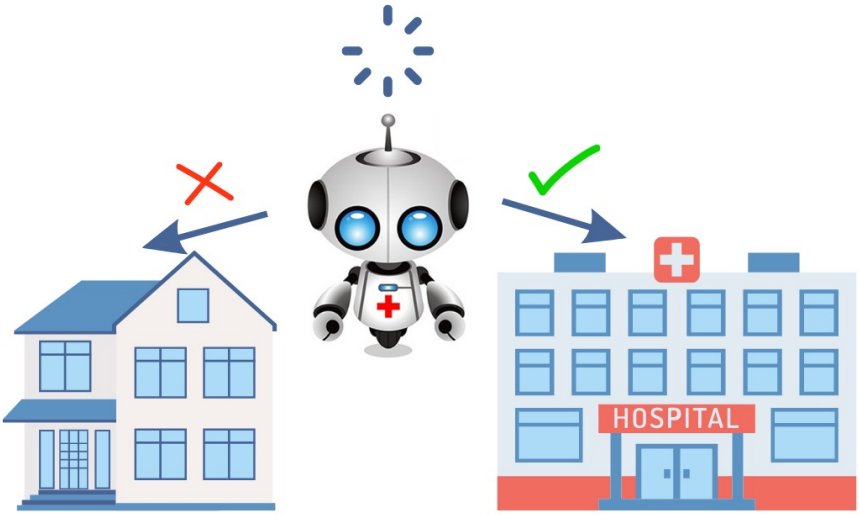


Applying Machine Learning to Intelligent Chatbot for Preventive Care

Sofia Malpique
Mestrado em Ciência de Computadores
Departamento de Ciência de Computadores
2023

Orientador
Prof. Investigadora Sénior Eva Maia, Instituto Superior de Engenharia do Porto

Coorientador
Prof. Auxiliar Rita Ribeiro, Faculdade de Ciências da Universidade do Porto

UNIVERSIDADE DO PORTO

MASTERS THESIS

Applying Machine Learning to Intelligent Chatbot for Preventive Care

Author:
Sofia MALPIQUE

Supervisor:
Eva MAIA

Co-supervisor:
Rita RIBEIRO

A thesis submitted in fulfilment of the requirements
for the degree of MSc. Computer Science

at the

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores
GECAD - Grupo de Investigação em Engenharia e Computação Inteligente para a Inovação e o Desenvolvimento

November 27, 2023

"The pursuit of knowledge is an ongoing process of doubt and inquiry. By subjecting our beliefs to rigorous scrutiny, we can uncover errors, refine our understanding, and approach closer to the truth. Doubt is not an obstacle but a means to attain certainty."

René Descartes

Acknowledgements

I want to express my gratitude to the project iCare4NextG for providing me with this incredible opportunity. Along this journey, I have been fortunate to have the guidance and support of two remarkable women, Researcher Eva Maia and Professor Rita Ribeiro. Their presence has been invaluable at every step of the way. I extend my heartfelt appreciation to Professor Isabel Praça from GECAD, whose passion has inspired my tech-savvy journey.

To Andreia, Nuno, Tiago, and Vito, your off-record insights and brainstorming sessions have significantly shaped my research. I am grateful for your engagement. My mother's motivation and uplifting talks have been a constant driving force, and my father's introduction to the world of computers ignited my curiosity.

Diogo Capela, your extensive knowledge has made complex concepts understandable and your passion for computers has been trully inspiring. Thank you for being such a source of strength. To everyone mentioned and others who contributed in many ways, my deepest gratitude for believing in me. Your support has made this journey possible. Thank you.

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores
GECAD - Grupo de Investigação em Engenharia e Computação Inteligente para a Inovação e o Desenvolvimento

MSc. Computer Science

Applying Machine Learning to Intelligent Chatbot for Preventive Care

by Sofia MALPIQUE

This dissertation aims to enhance preventive care with two primary contributions. Firstly, it constructs a machine learning model that, when integrated into a medical chatbot, can predict the likelihood of hospitalization for COVID-19 patients in home quarantine, thereby facilitating early identification and improving patient care.

The research focuses on utilizing straightforward patient data, such as age, sex, symptoms, and underlying medical conditions, to construct a robust classification model. It begins with the careful selection of an appropriate COVID-19 patient dataset, taking into consideration factors like dataset size and data completeness.

An exploratory data analysis (EDA) aids in understanding the dataset, including examining relationships between variables and hospitalization. Cluster analysis reveals distinctive patterns in symptoms and comorbidities.

To enhance model performance, feature selection techniques are employed, showcasing the complexity of the dataset. Given a class imbalance with only 3.9% positive cases, techniques like random under-sampling, SMOTE, and SMOTEENN are applied to boost model effectiveness.

Hyperparameter tuning via grid search optimizes the selected models, and their performance is assessed using metrics such as G-means, F1-Score, and ROC-AUC on an independent test dataset.

The results underscore the substantial impact of sampling techniques, with Gradient Boosting demonstrating exceptional performance. The combined effect of addressing

class imbalance and employing feature selection techniques markedly improves model efficacy.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores
GECAD - Grupo de Investigação em Engenharia e Computação Inteligente para a Inovação e o Desenvolvimento

Mestrado em Ciência de Computadores

Aplicação de Machine Learning a um Chatbot Inteligente para Cuidados Preventivos

por Sofia MALPIQUE

Esta dissertação contribui para o campo de *preventive care* através do desenvolvimento de um modelo de *Machine Learning (ML)* para um chatbot, que prevê a necessidade de hospitalização de doentes COVID-19 durante a quarentena. Esta funcionalidade contribui para a melhoria do atendimento de doentes COVID-19 e para a alocação de diferentes recursos.

O modelo desenvolvido utiliza técnicas de inteligência artificial (IA) para analisar variáveis de dados *rudimentares* - idade, sexo, sintomas, comorbidades - a fim de estabelecer um sistema de classificação robusto.

A pesquisa inicial envolveu a identificação de *datasets* adequados de doentes COVID-19, considerando certas características como tamanho e quantidade de valores nulos. O *dataset* escolhido impulsionou a análise subsequente e o desenvolvimento do modelo.

O estudo começou com uma análise exploratória de dados (AED) para compreender o *dataset*. As variáveis foram analisadas individualmente, juntamente com sua relação com a variável *target*. Foi também realizada uma análise com uso de *clusters* para estudar o agrupamento de doentes COVID-19, que revelou padrões de sintomas e comorbidades.

Foram empregues técnicas de *feature selection* através de modelos baseados em árvores, filtrando assim inicialmente o *dataset*. Para lidar com o desbalanceamento de classes - a classe positiva era apenas 3,9% - foram aplicadas técnicas de re-amostragem como *random undersampling*, SMOTE e SMOTEENN, que contribuiram para um *dataset* mais equilibrado, melhorando, desta forma, a robustez do modelo.

A *grid-search* facilitou o *tunning* de hiperparâmetros para os modelos considerados. Os models, depois de optimizados, foram avaliados usando métricas como G-means, F1-Score e AUC-ROC num *testset* independente.

Os resultados destacaram o impacto das técnicas de re-amostragem. RUS e SMOTE-ENN tiveram um desempenho consideravelemnte melhor. Destaca-se o modelo *Gradient Boosting (GB)* nas diferentes métricas e versões do *dataset*. As descobertas realçam a importância de abordar o desbalanceamento de classes e de empregar técnicas de *feature selection* para uma melhor eficácia do modelo.

Contents

Acknowledgements v

Abstract vii

Resumo ix

Contents xi

List of Figures xv

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Context . . . 2
- 1.3 Objectives . . . 3
- 1.4 Outline of the Dissertation . . . 4

2 State of the Art 5
- 2.1 Concepts & Definitions . . . 5
- 2.2 Remote Patient Monitoring Systems . . . 14
- 2.3 Health Chatbots . . . 15
  - 2.3.1 COVID-19 Health Chatbots . . . 16
  - 2.3.2 Challenges of Health Chatbots . . . 17
- 2.4 Applications of AI and ML for COVID-19 . . . 19
  - 2.4.1 Early detection, diagnosis, and prediction of the disease . . . 19
  - 2.4.2 Individuals' contact tracing . . . 20
  - 2.4.3 Mortality and number of cases projection . . . 21
  - 2.4.4 Reducing the workload of healthcare workers . . . 22
  - 2.4.5 Prevention . . . 22
  - 2.4.6 Monitoring the treatment . . . 23
  - 2.4.7 Conclusion . . . 24
- 2.5 Tools . . . 24

3 An overview of COVID-19 public data 27
- 3.1 COVID-19 Datasets . . . 27
- 3.2 Analysis and Pre-Processing . . . 30
  - 3.2.1 nCov2019 . . . 30

3.2.2 DS4C . . . 31
3.2.3 TriCovB . . . 32
3.3 Comparison of Datasets . . . 33
3.4 Conclusions . . . 35

4 Case Study on Brazil COVID-19 Hospitalization **37**
4.1 Dataset Description . . . 37
4.2 Univariate Analysis . . . 39
4.2.1 Relevant Dates . . . 39
4.2.2 Quarantine Duration . . . 41
4.2.3 Patient Outcome . . . 42
4.2.4 Patient Hospitalization . . . 44
4.2.5 Demographics . . . 45
4.2.6 Ethnicity . . . 46
4.2.7 Symptoms . . . 48
4.2.8 Comorbidities . . . 50
4.2.9 Extra Patient Information . . . 51
4.3 Multivariate Analysis . . . 52
4.3.1 Age and Sex . . . 53
4.3.2 Quarantine Duration . . . 56
4.3.3 Ethnicity . . . 57
4.3.4 Symptoms . . . 58
4.3.5 Comorbidities . . . 59
4.3.6 Extra Patient Information . . . 60
4.4 Clustering Analysis . . . 63
4.4.1 Symptoms . . . 64
4.4.2 Comorbidities . . . 66

5 Hospitalization Prediction Task **71**
5.1 Feature Selection . . . 71
5.2 Tackling Imbalance . . . 75
5.3 Hyperparameter Tuning . . . 78
5.3.1 Original Dataset . . . 80
5.3.2 RUS Dataset . . . 82
5.3.3 SMOTE Dataset . . . 82
5.3.4 SMOTEENN Dataset . . . 83
5.4 Performance Estimation . . . 84
5.4.1 Original Dataset . . . 84
5.4.2 RUS Dataset . . . 85
5.4.3 SMOTE Dataset . . . 86
5.4.4 SMOTEENN Dataset . . . 86
5.5 Discussion . . . 87

6 Conclusions **91**
6.1 Contributions . . . 91
6.2 Future Work . . . 93

Bibliography **95**

List of Figures

3.1 Percentage of Missing Values on each Dataset . . . 34

4.1 Monthly COVID-19 Cases . . . 40
4.2 Seasonal COVID-19 Cases . . . 41
4.3 Erroneous Data Points . . . 41
4.4 COVID-19 cases distribution by Quarantine Duration . . . 42
4.5 Fatalities Distribution . . . 43
4.6 Hospitalization Distribution . . . 44
4.7 COVID-19 cases distribution by Age . . . 46
4.8 COVID-19 Cases Distribution by Ethnicity . . . 47
4.9 Percentage of patients with symptoms . . . 49
4.10 Percentage of patients with comorbidities . . . 51
4.11 Extra Patient Information . . . 52
4.12 Hospitalization Cases by Survivance . . . 53
4.13 COVID-19 cases distribution by Sex . . . 54
4.14 Age distribution by hospitalization outcome . . . 55
4.15 Sex distribution by hospitalization outcome . . . 55
4.16 Quarantine duration distribution by hospitalization outcome . . . 56
4.17 Ethnicity distribution by hospitalization outcome . . . 57
4.18 Symptoms in Hospitalized Patients VS Non-Hospitalized Patients . . . 58
4.19 Comorbidities in Hospitalized Patients VS Non-Hospitalized Patients . . . 59
4.20 Pregnancy distribution by hospitalization outcome . . . 60
4.21 Health Professionals distribution by hospitalization outcome . . . 62
4.22 Incapacity distribution by hospitalization outcome . . . 62
4.23 Percentage of patients per cluster of symptoms. . . . 64
4.24 Percentage of patients within each comorbidity cluster . . . 66

5.1 Heatmap of Feature Importance . . . 73
5.2 Random Undersampling Process . . . 76
5.3 SMOTE Process . . . 77
5.4 SMOTEENN Process . . . 77

Chapter 1

Introduction

This chapter lays the foundation for the research presented in this dissertation, which aims to contribute to the field of preventive care using artificial intelligence (AI) techniques. It begins by providing the motivation behind the study, highlighting the importance of preventive care and the potential of intelligent chatbots in healthcare. Subsequently, the context of the research is established. Next, the research objectives are outlined, which include contributing to the field with insights gained from the available data and also developing a classification model to predict the need for hospitalization of COVID-19 patients. Finally, the chapter concludes with a brief overview of the structure and outline of the subsequent chapters, providing the reader with a clear roadmap of the dissertation's content.

1.1 Motivation

The COVID-19 pandemic has not only worsened the existing challenges but also shed light on the lasting issue of poor resource management within hospitals. Healthcare systems worldwide have been quite burdened, struggling to meet the overwhelming demand for medical care as the cases continued to surge [1]. Despite these pre-existing challenges, healthcare systems persistently strive to adapt and respond to the crisis, working tirelessly to ensure patients receive the necessary care during these challenging times [2]. Furthermore, in response to the pressing need for improved resource management, the COVID-19 pandemic has reinforced the significance of implementing remote patient monitoring (RPM) systems [3].

Over the past few years, these systems have provided a new solution by enabling healthcare providers to remotely monitor and care for patients, helping to alleviate the strain on hospital resources, while offering patients the ability to safely transition to their homes. Thereby ensuring they receive appropriate support and providing a considerable degree of flexibility and convenience in managing their health. With RPM systems, individuals can experience a greater sense of comfort and autonomy while ensuring continuous care [4]. RPM systems leverage digital monitoring tools, and they can be further improved through the integration of AI techniques [5, 6]. These tools may include digital oximeters and thermometers, as illustrated in the iCare4NextG project's use case (explained in Section 1.2). AI integration with digital monitoring tools enables the early detection of potential health issues, offering significant benefits, such as the ability to predict hospitalization needs in COVID-19 patients before their conditions deteriorate [7]. This dissertation demonstrates the application of machine learning (ML) within RPM systems.

1.2 Context

The iCare4NextG project focuses on advancing healthcare through the integration of innovative technologies. It aims to improve RPM through the integration of digital solutions. The project aims to enable proactive healthcare management by employing digital monitoring tools and cutting-edge AI techniques. This approach is intended to create a more efficient and responsive healthcare system by anticipating health concerns, optimizing patient care, and contributing to improved patient outcomes and resource utilization [8].

As part of its broader objectives, the iCare4NextG project includes the development of an intelligent chatbot for preventive care. This chatbot is designed to assist in achieving the project's tasks and goals. With the incorporation of ML and other AI capabilities, the chatbot aims to provide insights, predictions, and support to both patients and healthcare providers, precisely when needed. The project encompasses various use cases, one of which involves the monitoring of COVID-19 patients. While COVID-19 monitoring is a specific application within the project, iCare4NextG's aspirations extend beyond this use case. The project's broader mission is to reshape remote patient management in healthcare through technological innovation, with the intelligent chatbot and RPM framework playing core roles in achieving this transformation [9].

Typically, RPM systems operate in settings where collecting more complex data isn't always feasible due to the absence of a hospital environment or specialized tools [10]. So

there is a preference for an RPM system that excels in its performance by relying only on *simple data*, especially in cases like COVID-19 patients who are sent home without access to lab results or specialized medical tools. *Simple data* includes basic information individuals can provide about themselves, for instance, age, sex, symptoms experienced, and comorbidities. When AI techniques are applied to this information, the derived insights can hold significant value for healthcare providers when making decisions about the patient's condition, even in home settings with limited medical resources [11]. Leveraging this information, RPM systems can proficiently monitor patient health, facilitate timely interventions, and elevate the overall quality of healthcare delivery [12].

1.3 Objectives

The objective of this dissertation is to contribute to the field of preventive care by developing a tool that can improve RPM systems. This broader objective can be sectioned into smaller goals, presented in the natural order of events.

The initial phase involves data exploration and comprehension. This deep dive into the data is essential to gain a thorough understanding of COVID-19 patients' characteristics and outcomes. It provides insights that will contribute to subsequent stages of the research. Building upon the insights obtained from the data exploration, the next step is the construction of a predictive classifier. This classifier's primary role is to effectively predict whether a COVID-19 patient, who is under home quarantine, requires hospitalization. It is supposed to leverage simple and readily available data, mirroring the information typically accessible to patients in their home environments. This construction phase aligns with the broader goal of enhancing the efficiency and reliability of an RPM system.

The predictive model aims to inspire confidence in patients undergoing home quarantine. It seeks to alleviate concerns and ensure patient comfort during this challenging period, by delivering robust predictive capabilities. Subsequently, such a model is intended to be integrated as a feature into an intelligent chatbot designed for preventive care. This integration represents a pivotal improvement to the chatbot's functionality, further aligning with the project's objectives. However, it's important to note that this thesis primarily focuses on constructing and refining the predictive model, while the integration process into the chatbot falls outside the current study's scope.

Additionally, this research commits itself to maximizing the application of information collected from the working dataset. This data-driven approach not only enriches the

understanding of patient conditions but also contributes to informed decision-making in the realm of preventive care, by reveling insights and patterns through the exploratory phase. Each of these steps in the research process plays an important role in achieving the broader objective of enhancing patient safety and well-being through an advanced RPM system. In this way, the research significantly contributes to the field of preventive care.

1.4 Outline of the Dissertation

1. **Chapter 1: Introduction** - Overview of the research work and its goals. Explanation of the motivation behind developing an ML model for predicting hospitalization. Specific objectives and significance of the study in preventive care.

2. **Chapter 2: State of the Art** - Extensive review of the definitions of the necessary concepts related to the present study. Examination of existing ML applications in healthcare. Overview of previous studies related to predicting hospitalization in COVID-19 patients. Identification of the types of publicly available datasets COVID-19 related.

3. **Chapter 3: Datasets** - Information regarding datasets pertaining to COVID-19 patients, including origins, attributes, and data quality. Explanation of data preprocessing procedures, containing the treatment of missing values and identification of outliers. A concise overview of the refined dataset tailored for the predictive model.

4. **Chapter 4: Exploratory Data Analysis (EDA)** - Comprehensive analysis of each variable in the dataset. Examination of relationships between variables and the target variable (hospitalization indication). Results of cluster analysis revealing grouping patterns of symptoms and comorbidities.

5. **Chapter 5: Hospitalization - A Classification Task** - Development of the predictive model for hospitalization as a classification task. Application of feature selection techniques to identify influential variables. Discussion of sampling techniques to address class imbalance. Summary of the selected predictive model architecture.

6. **Chapter 6: Conclusion** - Summary of key findings and implications of the research. Reiteration of significant contributions of the intelligent chatbot in preventive care. Discussion of limitations encountered during the research. Outline of future research directions and potential enhancements to the predictive model.

Chapter 2

State of the Art

In this chapter, the following sections will explain the key concepts necessary to understand the subsequent discussions. The groundwork will be laid for a better understanding of the state-of-the-art RPM technologies, health chatbots, ML techniques behind these types of technologies, and COVID-19-related datasets. Existing studies and research on these topics will be examined in each of the areas mentioned before. The advantages and disadvantages of various approaches will be examined, and their contribution to advancing the healthcare system will be evaluated.

2.1 Concepts & Definitions

AI is a field of computer science dedicated to creating intelligent systems. These systems are designed to perform tasks that usually require human intelligence, such as learning, reasoning, problem-solving, and decision-making. A significant component of AI is ML, centered around developing algorithms and models that empower computers to learn from data and make predictions or take actions without being explicitly programmed. This learning process involves identifying patterns and insights from data to enhance performance over time [13].

In the realm of ML, there are two main types: supervised learning and unsupervised learning. In supervised learning, the model is trained on labeled examples, learning to generalize patterns from the labeled data to make predictions or classifications on unseen data (explained in more detail in upcoming paragraphs). On the other hand, unsupervised learning involves the model learning patterns and relationships in the data without

explicit labels. This type of learning helps discover hidden structures or groupings within the data [14].

Table 2.1 compares supervised and unsupervised learning based on several key characteristics, including the type of input data, the type of output data, the overall goal, some examples of tasks, and some common algorithms used for each approach [15–17].

TABLE 2.1: Comparison of Supervised and Unsupervised Learning

	Supervised Learning	**Unsupervised Learning**
Input data	Labeled data	Unlabeled data
Output data	Predicted output	No output
Goal	Predictive modeling	Discovering patterns and relationships
Examples	Classification, regression	Clustering, anomaly detection
Common algorithms	Naive Bayes, SVM, Decision Trees	K-Means, PCA, DBSCAN

The focus of this dissertation will be on classification, regression, and clustering, even though various other ML tasks exist [18].

Classification, as a supervised learning task, entails the ML model learning to categorize input data into predefined classes or categories. It establishes a mapping between the data and corresponding class labels. It's important to note that classification can be categorized into two main types: binary classification, where data is sorted into two distinct classes, and multi-class classification, which expands to multiple categories. For instance, binary classification can be exemplified by the task of distinguishing between spam and non-spam emails, while multi-class classification involves categorizing images into classes like cats, dogs, and birds [15].

On the other hand, regression represents another dimension of supervised learning, with a focus on predicting continuous numerical values. In regression, the model learns to establish relationships between input features and output values, often used for tasks like predicting house prices based on attributes such as square footage, bedroom count, and location [17].

Unlike classification, clustering does not rely on predefined classes but rather aims to reveal underlying patterns within the data. This is why Clustering is considered unsupervised learning. This technique involves grouping similar data points together based on their inherent characteristics. This proves particularly valuable for discovering specific segments or identifying similarities among data points [19].

In essence, these three tasks form the foundation of ML empowering diverse applications. Classification assigns data to classes, regression predicts continuous values, and clustering uncovers hidden relationships.

Now, transitioning from discussing the fundamental aspects of ML tasks to the importance of data, it is necessary to understand how these ML tasks rely on data for their functionality. It all starts with a dataset (or multiple), which is a collection of organized information, either structured or unstructured, used for training and testing ML models. It consists of input details along with corresponding output labels or target values [18].

As data is essential for training an ML model and assessing its performance, typically, a division of 70:30 or 80:20 is made, where the larger portion, referred to as training data, is used to teach ML models, while the smaller part goes for testing. During training, models learn from input features and their known corresponding output labels to make accurate predictions or classifications. The smaller segment, known as test data, remains unused during training. Instead, it's used to assess the trained model's accuracy and its ability to perform well on new, unseen data. This process ensures that the model doesn't just memorize the training data but generalizes its learnings effectively [20].

Assessing how well a trained model performs on unseen data is an indispensable step in the ML process, known as model evaluation. The choice of evaluation metrics depends on the type of ML task at hand. In classification tasks, commonly used metrics include accuracy, precision, sensitivity, specificity, F1-Score, G-Mean, and area under the ROC curve (AUC-ROC). These metrics help people understand how well the model can distinguish between different categories while managing errors. We will provide detailed definitions and explanations of each metric, starting with individual metrics and then moving on to composite metrics. These definitions are drawn from the book 'Machine Learning: A Probabilistic Perspective' by Kevin P. Murphy [21].

The basic concepts are:

- True Positives (TP) represent the correctly predicted positive instances.
- True Negatives (TN) represent the correctly predicted negative instances.
- False Positives (FP) represent the instances that were predicted as positive but were actually negative.
- False Negatives (FN) represent the instances that were predicted as negative but were actually positive.

Accuracy measures the proportion of correct predictions made by the model out of all the predictions it made. It provides an overall view of how well the model is performing in terms of both true positives and true negatives. The accuracy formula is given as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

Precision measures the proportion of positive predictions made by the model that are actually correct. It provides insights into the model's ability to avoid false positives, which can be crucial in scenarios where false alarms have significant consequences (Equation 2).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Sensitivity measures the proportion of actual positive cases that are correctly identified by the model. It helps assess the model's effectiveness in capturing all positive instances and minimizing false negatives, which is important when missing positive cases can lead to severe consequences (Equation 3).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

Specificity measures the proportion of actual negative cases that are correctly identified by the model. It's especially valuable when correctly identifying negative cases is essential, such as in medical diagnostics (Equation 4).

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

The F1-Score combines Precision and Sensitivity into a single value. It offers a balanced assessment of the model's ability to minimize both false positives and false negatives. The harmonic mean takes into account both precision and recall, making it useful when there's a need to balance these aspects (Equation 5).

$$\text{F1-Score} = \frac{2 \times (\text{Precision} \times \text{Sensitivity})}{(\text{Precision} + \text{Sensitivity})} \tag{5}$$

G-Mean is a performance metric that considers both sensitivity and specificity. It provides a balanced evaluation of classification performance across multiple classes, making it suitable for scenarios with imbalanced datasets (Equation 6).

$$\text{G-Mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{6}$$

AUC-ROC is a metric used to evaluate the model's capability to distinguish between the positive and negative classes across varying classification thresholds. The ROC curve illustrates the true positive rate (TPR) - a.k.a. sensitivity - against the false positive rate (FPR) - i.e. 1 - specificity - at different thresholds. The AUC-ROC quantifies the overall performance of the model in discerning between the two classes, providing a comprehensive view of its discriminatory power (Equation 7).

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(FPR)\text{d}FPR \tag{7}$$

Similarly, in regression tasks, standard metrics encompass mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared. These metrics provide insights into the model's precision when predicting numerical values and its overall performance [22]. However, since they are not pertinent to this dissertation, detailed explanations will not be provided.

In tasks involving classification, two or more classes are presented, and when there's a significant imbalance in the distribution of these classes, the data becomes imbalanced. In such situations, careful consideration is needed regarding the metrics employed for evaluating the model's performance. Some metrics, like accuracy, may not be suitable for such scenarios. A model trained on highly imbalanced data tends to classify most instances as the majority class, resulting in a high accuracy rate. However, if accurately predicting the minority class is crucial, other metrics are more appropriate for evaluating the model's effectiveness [23].

Indeed, the choice of suitable metrics plays an important role in improving model performance. However, it is equally imperative to acknowledge that achieving superior results relies on meticulous data preparation [17]. In this context, this dissertation will delve into some of the common steps in the subsequent list:

- **Data cleaning:** This fundamental step involves identifying and rectifying errors or inconsistencies in the dataset, such as duplicate records, inaccurate values, or outliers. It ensures dataset accuracy and reliability, forming a solid foundation for analysis [15].

- **Handling missing values:** Addressing missing data points is crucial to avoid biased results and incomplete analyses. Various techniques, like imputation or data removal, can be applied based on the dataset's characteristics [16].

- **Feature engineering:** This process focuses on creating new features or modifying existing ones to enhance ML model performance. It aims to capture essential data patterns and relationships, improving predictive power [17].

- **Feature selection:** Feature selection narrows down attributes to those with the most significant impact on model performance, reducing complexity, computational load, and overfitting [16].

- **Encoding categorical variables:** Transforming categorical data into numerical form bridges different data types and enhances the model's capacity to derive insights from diverse attributes [15].

- **Data reduction:** Techniques like Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE) simplify dimensionality, preserving vital data information and facilitating more efficient pattern extraction [17].

- **Outlier detection:** This step ensures data integrity by identifying and addressing anomalous data points, preventing skewed model interpretations and enhancing accuracy [16].

- **Data standardization:** Normalizing data to a consistent range prevents certain attributes from overshadowing others due to their magnitudes, ensuring model fairness [17].

- **Dealing with class imbalance:** Addressing scenarios where class distribution is skewed ensures each class is fairly represented, preventing biased model predictions and promoting accuracy and fairness [15].

With a foundation in data pre-processing established, the focus now shifts to exploring various ML algorithms, such as Decision Tree (DT), K-Nearest Neighbours (KNN), Logistic Regression (LR), AdaBoost, Bagging, Gradient Boosting (GB), Random Forrest (RF), and eXtreme Gradient Boosting (XGBoost). The definitions follow the idea presented in the book *Introduction to Machine Learning* by the author Ethem Alpaydin [16].

Before delving into the definitions of each model, it is necessary to understand the concept of ensemble methods in ML. Ensemble methods, also known as ensemble techniques, represent a powerful approach that amalgamates predictions or decisions from multiple individual algorithms, often referred to as "base models" or "weak learners." The goal of ensemble methods is to generate a prediction or classification that is not only more accurate but also more robust. The underlying principle of ensemble methods revolves around leveraging the diversity and collective intelligence of these multiple models to enhance predictive performance.

The core idea driving ensemble methods is the acknowledgment that combining predictions from various models can frequently yield superior results compared to relying solely on a single model. Each individual model within the ensemble may possess its unique strengths and weaknesses. Ensemble methods effectively mitigate weaknesses and amplify strengths by aggregating their predictions, ultimately yielding predictions that are not only more precise but also more stable and reliable.

- **Decision Tree:** A decision tree [24] is a graphical representation of possible decisions based on certain conditions. It starts with a root node representing the entire dataset and branches out to internal nodes that correspond to specific features. Each internal node makes a decision, and the leaf nodes provide the final classification or prediction. Decision trees are easy to understand and visualize.

- **K-Nearest Neighbors (KNN):** K-Nearest Neighbors [25] is a simple classification and regression algorithm that relies on the similarity between data points. Given a new data point, KNN identifies the K nearest neighbors based on a chosen distance metric. The algorithm then predicts the class label (classification) or value (regression) based on the majority class or average of the neighbors' values. KNN is intuitive and effective for smaller datasets.

- **Logistic Regression:** Despite its name, logistic regression [26] is a classification algorithm. It models the probability of a binary outcome using a linear combination

of input features, transformed by the logistic function. The output represents the likelihood of belonging to a certain class. Logistic regression is interpretable, and its coefficients provide insights into feature importance.

- **AdaBoost:** AdaBoost [27] (Adaptive Boosting) is an ensemble learning technique that combines multiple weak learners to create a strong classifier. It assigns higher weights to misclassified data points, allowing subsequent weak learners to focus on correcting these mistakes. AdaBoost's final prediction is based on the weighted decisions of all weak learners. It's an effective method for improving classification performance by addressing complex datasets.

- **Bagging:** Bagging [28] (Bootstrap Aggregating) is an ensemble method that reduces prediction variance by combining multiple models trained on bootstrapped datasets. Each model learns from a subset of the data, and the final prediction is obtained by aggregating individual model outputs. Bagging is particularly useful for models with high variance, enhancing stability and overall accuracy.

- **Gradient Boosting:** Gradient Boosting [29] is an advanced ensemble technique that builds a strong model by iteratively improving upon the mistakes of previous models. Each new model focuses on the residuals of the previous model's predictions. Gradient Boosting combines these weak models to create a powerful predictor capable of capturing complex relationships in the data.

- **Random Forest:** Random Forest [30] is an ensemble technique that constructs multiple decision trees and aggregates their predictions to make a final decision. Each tree is trained on a subset of the data, reducing overfitting and increasing generalization. Random Forest is versatile, effective, and handles high-dimensional datasets well.

- **XGBoost:** XGBoost [31] (eXtreme Gradient Boosting) is an optimized version of gradient boosting that incorporates regularization, parallel processing, and advanced optimization techniques. XGBoost provides superior predictive power and efficiency.

We selected these algorithms based on strong support from existing literature. For instance, in a study conducted by S. S. Aljameel et al. [32], LR, XGBoost, and RF were used as algorithms, and RF outshone the others with an impressive accuracy of 0.95 and an exceptional area under the curve (AUC) of 0.99. Similarly, K. Moulaei [33] evaluated

seven ML algorithms, including J48 decision tree (J48), RF, KNN, multi-layer perceptron (MLP), Naïve Bayes (NB), XGBoost, and LR, and found that RF consistently delivered superior performance. RF achieved remarkable accuracy, sensitivity, precision, specificity, and receiver operating characteristic (ROC) values, standing at 95.03%, 90.70%, 94.23%, 95.10%, and 99.02%, respectively.

Furthermore, S. S. Zakariaee et al. [34] delved into the evaluation of eight ML algorithms, encompassing J48, support vector machine (SVM), MLP, KNN, NB, LR, RF, and XGBoost. In this comprehensive analysis, the RF algorithm once again demonstrated outstanding performance, boasting an accuracy of 97.2%, sensitivity of 100%, precision of 94.8%, specificity of 94.5%, F1-Score of 97.3%, and an AUC of 99.9%. These consistent findings across multiple studies strongly suggest that the RF model is poised to deliver top-tier performance, potentially even securing the highest scores in our research.

As we conclude our examination of ML models and their explanations, it's important to note that while they perform well in many situations, they can struggle with complex tasks like image recognition or speech understanding. The introduction of Deep Learning (DL) is a must as it signifies a significant AI evolution. DL, a subset of ML, emphasizes multi-layered Neural Networks (NNs), inspired by human brain neurons, serving as building blocks for both DL and ML algorithms [35].

NNs consist of interconnected neurons organized into layers. These neurons process data using weighted connections and activation functions. Training involves adjusting weights to minimize prediction errors, with NNs excelling in image recognition and natural language processing [36].

ML and DL differ in data representation, training, model complexity, interpretability, and performance. ML relies on manual feature engineering, while DL learns features directly from raw data. DL requires more computational power for extensive datasets, resulting in complex models. ML is more interpretable, whereas DL can be challenging to interpret. DL outperforms in handling unstructured data and complex tasks [35, 36].

Regarding algorithms, ML uses RF, NB, and KNN, while DL employs specialized algorithms like Convolutional Neural Networks (CNNs) for image analysis and Long short-term memory (LSTM) for sequential data tasks. These NNs have applications in language processing, speech recognition, and more [37].

DL's adaptability extends to Natural Language Processing (NLP), enabling tasks like sentiment analysis, language translation, and chatbot creation [38].

This technology enhances patient care by providing real-time insights into health status and potential concerns, aligning technology with healthcare for improved monitoring and personalized interventions [39].

2.2 Remote Patient Monitoring Systems

The concept of RPM and the technologies beyond RPM systems are introduced in Chapter 1. One of the significant benefits of RPM is its ability to facilitate timely interventions if necessary [40]. Healthcare providers can detect and address problems before they escalate, by continuously monitoring patients remotely, improving patient outcomes, and reducing the risk of adverse events [41].

RPM systems, as highlighted in the study by Kaur et al. [42], can collect data from diverse sources, including wearable devices, sensors, and medical equipment. These systems possess the unique capability to discern intricate patterns and detect anomalies that may elude human observation, by subjecting this data to analysis through ML techniques. Consequently, this analytical achievement empowers these systems to offer invaluable insights into a patient's health status, as demonstrated in the research by Oliver et al. [43]. Furthermore, this higher analytical robustness translates into improved healthcare delivery, ultimately benefiting patient care.

Data from various sources, such as wearable devices, sensors, and medical equipment, can be gathered by RPM systems [42]. RPM systems can detect patterns and anomalies that may not be apparent to the human eye, by analyzing this data and applying ML techniques, providing, this way, valuable insights into a patient's health status [43]. Also, it can improve their robustness, resulting in better healthcare delivery.

With the COVID-19 pandemic, RPM systems became increasingly important for remotely monitoring and managing COVID-19 patients. G. Saranya et al. [44] present an IoT and cloud-assisted health monitoring system designed for RPM. The proposed system employs multiple sensors to detect and monitor the severity of COVID-19 in patients. It collects disease-specific parameters such as heart rate, temperature, oxygen level, and pulse rate. The collected data is processed on a cloud server, and CNN models are applied to identify the severity of the disease. The system also generates an alert for healthcare providers if any abnormalities are detected during the computation of sensed data on the CNN. This model can be used as a prediction and forecasting technique to determine the severity of the patient based on their health data.

Some other examples of RPM systems specifically designed for COVID-19 patients are the Vivify Health RPM system [45] and the Philips eCareCoordinator [46]. Vivify Health operates as both a company and a remote patient monitoring platform, providing inventive solutions to elevate patient engagement and the management of care. With a focus on enabling healthcare providers, the platform facilitates the remote monitoring of patients' health, the gathering of pertinent data, and the empowerment of patients in managing their care through personalized interventions and communication. The underlying technology from Vivify Health strives to enhance patient outcomes, diminish instances of hospital readmissions, and elevate the overall healthcare journey by harnessing the potential of remote monitoring and patient engagement tactics [45].

The Philips eCareCoordinator represents a comprehensive and advanced solution for remote patient monitoring, crafted by Philips, a prominent healthcare technology company. This innovative platform is meticulously designed to equip healthcare providers with the necessary tools and capabilities to proficiently oversee and track patients' well-being from a distance. Through the eCareCoordinator, healthcare professionals can conveniently and remotely monitor patients' crucial health metrics, pertinent health data, and ongoing progress in real-time. This real-time insight enables timely interventions and adaptable adjustments to treatment strategies as required. Additionally, the platform fosters seamless communication channels connecting patients and their care teams, enhancing patient engagement and facilitating personalized care delivery. The Philips eCareCoordinator harnesses advanced technology and sophisticated data analytics to pursue the goal of enhancing patient outcomes, streamlining healthcare delivery, and elevating the overall patient experience [46].

2.3 Health Chatbots

A chatbot serves as a concrete example of an AI system and stands as one of the most essential illustrations of intelligent Human-Computer Interaction (HCI) [47]. Functioning as a computer program, it emulates an intelligent entity by proficiently engaging in text or voice-based conversations and understanding one or more human languages through the application of NLP techniques [48]. In linguistic terms, a chatbot is defined as "A computer program designed to simulate conversation with human users" [49]. The chatbots are also recognized under various names, including intelligent bots, interactive agents, digital assistants, or artificial conversational entities.

There are COVID-19 health chatbots for different purposes and to perform different tasks. These tasks include, for example, answering questions [50, 51], asking questions [52], creating health records and history of use [53, 54], filling forms and generating reports [55].

2.3.1 COVID-19 Health Chatbots

The versatile applications of COVID-19 health chatbots have been pivotal in tackling the multifaceted challenges posed by the pandemic. These chatbots have harnessed their capabilities to offer indispensable support and services, addressing various aspects of the crisis.

One of the foremost roles undertaken by health chatbots was the dissemination of crucial health information and knowledge. They became a reliable source for resources related to COVID-19 symptoms, medication, and precautionary measures. These resources were made available through different forms and formats, including textual content, medical catalogs, audio clips, animated videos, and maps, catering to a range of preferences and needs [50–54].

Furthermore, health chatbots have played an important role in enabling self-triage and personalized risk assessment during the pandemic [51–53]. Operating based on guidelines from reputable organizations like the WHO and local health authorities, these chatbots facilitated self-screening to determine the need for inpatient care. Some were even employed for employee self-assessment before entering workplaces [55], while others provided information about nearby medical services and emergency hotlines for individualized risk evaluation [56].

Another significant application of these chatbots has been the monitoring of potential exposure to the virus and the provision of timely notifications. For instance, the COOPERA system in Japan has been instrumental in assessing the epidemiological situation, monitoring high-risk groups, and extending support where necessary [50].

Additionally, health chatbots have been adept at tracking health symptoms and mental well-being associated with the pandemic [50, 54, 57]. Users have been able to record various factors, such as nutrition and physical activity during self-isolation periods, alongside monitoring mood status to address psychological effects like anxiety and depression [58].

Notably, these chatbots have also been at the forefront of the fight against COVID-19 misinformation and fake news [57]. Distinguished examples include the WHO's chatbot, which disseminates reliable information and best medical practices, and the "COVID-19 Preventable" chatbot in Thailand, dedicated to sharing accurate information and raising awareness [52].

Table 2.2 provides an overview of various applications and roles that health chatbots have played during the COVID-19 pandemic. It encapsulates the roles and functions of these chatbots within the framework of pandemic management and response, categorizing them based on their respective applications and chatbot roles.

TABLE 2.2: Applications of COVID-19 Health Chatbots

Application	**Role of Chatbot**	**Reference**
Dissemination of Information	Providing resources on COVID-19 symptoms, medication, and precautionary measures through various formats.	[50–54]
Self-Triage and Risk Assessment	Enabling self-screening based on guidelines, determining the need for inpatient care.	[51–53]
Monitoring Exposure and Notifications	Assessing the epidemiological situation, monitoring high-risk groups, and providing support.	[50]
Tracking Health Symptoms and Mental Well-Being	Recording factors like nutrition, physical activity during self-isolation, and monitoring mood status.	[50, 54, 57, 58]
Combating Misinformation	Dispelling COVID-19 misinformation and fake news, providing reliable information and best practices.	[57]

2.3.2 Challenges of Health Chatbots

Naturally, there are challenges associated with the utilization of these technologies, both on social and technical system levels. At the social system level, some health chatbots experienced limited engagement from the community, resulting in virtual inactivity [57]. Moreover, a disparity existed between users' perceptions of these technologies and the capabilities provided by health chatbots, influencing acceptance [58]. Negative user perceptions regarding chatbot integrity, benevolence, the accuracy of the information, and privacy preservation hindered their willingness to use health chatbots. Additionally, individuals without access to technology or the Internet could not benefit from these tools,

resulting in information gaps within the population and decreased accuracy in identifying and predicting cases of infection [54].

At the technical system level, fact-checking information in real-time posed a significant challenge for chatbots, as a vast amount of data is updated daily. Ensuring the processing and delivery of accurate and up-to-date information often necessitated the intervention of multidisciplinary teams [57]. Integrating information from multiple sources sometimes resulted in incoherence, leaving users needing clarification and searching for reliable answers. Furthermore, current chatbot capabilities may not be sufficiently developed to address sensitive topics like mental health. Empathetic Natural Language Generation, for example, is not yet considered sophisticated, limiting the chatbots' suitability in assisting individuals experiencing nervous breakdowns or suicidal thoughts – critical issues prevalent during pandemics [53]. Other concerns encompass the need for accurate medical translation from professional jargon to day-to-day terms to prevent misunderstandings and misguided actions. Given the novelty of the pandemic, different terminologies used to describe the same condition can lead to confusion and inappropriate user actions [55].

Among the mentioned challenges, the integration of AI techniques offers promising solutions. AI can play an important role in improving the functionality and effectiveness of health chatbots. Advanced NLP algorithms can improve the accuracy and coherence of information delivery, addressing concerns related to fact-checking and data integration. AI-powered sentiment analysis can assist in identifying users' emotions and mental states, helping prepare responses and interventions more effectively, especially in cases of sensitive topics like mental health [59]. ML models can be trained to identify and rectify misinformation, reducing the spread of inaccurate data. Moreover, the deployment of AI technologies can facilitate seamless language translation, bridging the gap between medical terminology and everyday language, and ensuring accurate communication even in diverse linguistic contexts. As these technologies continue to evolve, the integration of AI in health chatbots holds the potential to mitigate existing challenges and improve the overall user experience, contributing to more effective pandemic management and response strategies [60].

2.4 Applications of AI and ML for COVID-19

Among the challenges posed by the COVID-19 pandemic, a surge of innovative solutions has emerged to address its multiple complexities [61]. This section delves into the diverse applications of AI and ML technologies in the battle against COVID-19 [62]. These growing technologies have been applied in numerous medical studies, resulting in improved scalability, timely and reliable outcomes, and increased efficiency [63]. In some healthcare tasks, it has even surpassed human performance [64]. Tools based on AI are employed for the identification, classification, and diagnosis of medical images to manage disease spread [65, 66]. Recent advancements in AI research have significantly enhanced COVID-19 screening, diagnostics, and prediction [67].

ML-based techniques have been very helpful in identifying patterns and forecasting epidemics. In the context of the COVID-19 pandemic, many researchers have applied these techniques to facilitate early detection and diagnosis of the infection, contact tracing of individuals, projection of cases and mortality rates, development of drugs and vaccines, reducing the workload of healthcare workers, and monitoring the treatment of patients. Various studies have documented the successful application of ML-based techniques in these areas [68–70].

These algorithms have been applied to COVID-19 detection, diagnosis, classification, screening, drug repurposing, prediction, and forecasting [71, 72]. The following subsections cover each use-case/application.

2.4.1 Early detection, diagnosis, and prediction of the disease

AI can quickly analyze irregular symptoms and identify potential red flags, providing timely alerts to both patients and healthcare professionals. This facilitates decision-making and the implementation of cost-effective solutions. AI helps develop new diagnosis and management systems for COVID-19 cases using useful algorithms. In particular, sophisticated DL algorithms, such as CNNs, have a significant impact on extracting critical features, especially in the realm of medical imaging [69, 73]. These algorithms exhibit efficacy in diagnosing infections, as their performance is enhanced when coupled with medical imaging technologies, such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI) scans, or X-ray images of various body parts. A new diagnostic framework emerges, by synergizing AI-driven analyses with advanced imaging, allowing for a more detailed assessment of COVID-19 cases [72, 74].

The significance of ML and DL-based techniques is highlighted in several studies. For instance, M. Otoom et al. [75] utilized a dataset from the CORD-19 repository to evaluate the effectiveness of eight ML algorithms in identifying COVID-19 cases. CORD-19 is a COVID-19 research initiative that offers an extensive dataset containing scientific articles and publications related to COVID-19 and has been a valuable resource for researchers and healthcare professionals studying the pandemic [76]. Despite some challenges with the data, five of these algorithms achieved over 90% accuracy. L.H. Nguyen et al. [77] proposed a novel DL method that uses cough sounds for early detection of COVID-19. This method involves converting audio signals into Log-Mel spectrograms, which are then processed by a two-stage deep neural network.

Further, a model using X-ray images and support vector machines was suggested for early COVID-19 detection and diagnosis [78], demonstrating high accuracy. L. Wang et al. [76] proposed a method for early detection of suspected COVID-19 cases without the need for CT scans, using Lasso regression for feature selection and also as the base model, having a 100% recall score. Additionally, a model for diagnosing COVID-19 was developed using multivariate logistic regression based on a dataset of 620 laboratory samples [79], showing satisfactory performance with high predictive values. Besides these techniques, other ML-based techniques, including NN, k-means, XGBoost, gaussian process regression, and multilayer perceptron, have also been employed for COVID-19 screening, detection, prediction, and forecasting [80–83]. CNNs have also become a popular tool in the fight against the COVID-19 pandemic. Its applications range from COVID-19 screening, diagnostics, classification, and prediction, to forecasting, as demonstrated in various studies [84–86].

2.4.2 Individuals' contact tracing

AI has the potential to analyze the level of infection caused by COVID-19 by identifying clusters and hotspots and can also aid in contact tracing and monitoring of individuals. According to various studies, such technology can predict the future course of the disease and its likelihood of reappearance [87]. One major strategy for preventing the spread of the virus is tracing confirmed cases of COVID-19, given the potential for transmission through the air via coughing, sneezing, or talking [88]. It was recommended that not only those who have tested positive for COVID-19 but also those who have been in close

contact with confirmed cases be quarantined for 14 days. Contact tracing applications have been implemented worldwide using various methods [89].

Contact tracing begins after a case has been diagnosed since the individual needs to be tracked [90]. The data collected by contact tracing apps are then analyzed using AI techniques to determine the extent of the disease's spread [91]. Although these apps are useful during the pandemic, privacy concerns have arisen due to the large amount of data collected, and governments may surveil individuals [92]. This is where digital footprint data provided by these apps and ML technology can be utilized to detect infected patients and enforce social distancing measures [93]. One example of this application is the SQREEM platform, originally developed in Singapore to track individuals who may have contracted COVID-19, which has been utilized in South Africa for real-time contact tracing with AI technology [94].

2.4.3 Mortality and number of cases projection

ML has the capability to forecast the nature of the COVID-19 virus using available data, social media, and media platforms to identify the risks of infection and its spread [95]. It can also predict the number of positive cases and deaths in any region and identify the most vulnerable regions, people, and countries. With this information, appropriate measures can be taken to mitigate the spread of the virus. These capabilities have been demonstrated in several studies, as reported in various publications [96].

Multiple studies have compared several ML models. For instance, M. Pourhomayoun et al. compared the performance of DT, RF, KNN, SVM, LR, and ANN, for predicting the mortality rate in patients with COVID-19. The dataset used in this study included 117,000 cases of COVID-19 infection of both genders. The model achieved an accuracy of 93% for predicting the mortality rate. The DT method, when used with 10-fold cross-validation, achieved an accuracy of 90.63% on its own [97]. S.S. Zakariaee et al. [34] compared the performance of eight ML algorithms for predicting the mortality of COVID-19 patients. The RF algorithm stood out with an accuracy of 97.2%, sensitivity of 100%, precision of 94.8%, specificity of 94.5%, F1-Score of 97.3%, and an exceptional Area Under the ROC Curve of 99.9%. Other algorithms also demonstrated good prediction abilities, achieving AUC values ranging from 81.2% to 93.9%. The proposed model, especially when utilizing a dataset including chest CT severity score (CT-SS), proved effective in promptly assessing

COVID-19 patient risk, optimizing hospital resources, and improving patient survival probabilities.

2.4.4 Reducing the workload of healthcare workers

The COVID-19 pandemic has resulted in an abrupt rise in patient numbers, leading to an unprecedented workload for healthcare professionals, as already introduced in Chapter 1. AI can provide training to medical students and doctors to better understand the disease. The use of AI can have a significant impact on future patient care and tackle potential challenges, thereby reducing the workload of doctors [37, 69, 73, 74, 98].

N. Galo et al. [99] discusses a decision-making process for triaging suspected COVID-19 patients in Brazil. The paper suggests the use of computational techniques based on fuzzy inference systems, arguing that fuzzy set theory [100] is suitable for this problem since it allows natural language to describe the patient's symptoms, making it easier for healthcare professionals. The fuzzy system is modeled with symptoms that health professionals currently use to analyze COVID-19 cases, and a pilot test was conducted. The results suggest that the model aligns with the sample data and has the potential to support triage for classifying the severity of COVID-19 cases.

With the surge in the number of COVID-19 cases, manual severity assessment has become a challenging and time-consuming task. The authors Tang et al. [101] proposed an ML-based model that can automatically identify the severity level of COVID-19 patients. The RF model is trained using CT images of 176 COVID-19-positive patients for severity assessment. The results of this study are promising, with 87.5% accuracy using 3-fold cross-validation. The authors also identified various quantitative features that have the potential to assess the severity of COVID-19 cases.

2.4.5 Prevention

AI, combined with data analysis, can provide valuable and up-to-date information for the prevention of COVID-19 [102]. It can predict the areas with the highest likelihood of infection, the extent of the virus spread, the need for hospital beds, and the demand for healthcare professionals during this crisis [103]. Moreover, AI can help prevent future viruses and diseases by analyzing previous data and identifying trends, causes, and reasons for the spread of diseases. It can offer preventive measures and assist in combating other diseases [104].

One other facet of AI that has shown remarkable promise during the COVID-19 pandemic is the utilization of health chatbots, as mentioned earlier. These chatbots, available in multiple languages, have proven instrumental in aiding patients during the initial phases of the illness [105]. An illustrative example of such implementation is Aapka Chkitsak, an AI-driven chatbot developed in India by U. Bharti et al. [51].

2.4.6 Monitoring the treatment

It has been suggested the development of an intelligent platform for the automatic monitoring of COVID-19 patients [106]. These platforms, leveraging AI and data analytics, could provide daily updates on patient conditions, offer solutions to combat the pandemic, and significantly aid healthcare professionals in managing patients and controlling the virus's spread. Such platforms could provide insights for decision-making, and facilitate remote consultations [107].

H. Yu et al. focused on model-based decision trees to detect the severity of COVID-19 in pediatric cases [108] involved the collection of clinical laboratory and epidemiological data from 105 infected children. The outcomes of the study were encouraging, with the proposed model showcasing promising results. Impressively, it achieved a flawless F1 score of 100, underscoring the predictive potential of ML in gauging disease severity, a crucial aspect of the monitoring process.

B. S. Yelure et al. [109] utilized IoT and AI in remote monitoring by analyzing cough sounds. The detection of coughs was performed using Mel-frequency cepstral coefficients (MFCC) features and deep neural networks (DNN), as well as CNNs. This research highlights the potential of AI in non-invasive monitoring techniques.

A significant advancement in monitoring COVID-19 is the study by Kim et al. [110], which used an automated ML technique to develop prediction models using easily obtainable characteristics—baseline demographics, comorbidities, and symptoms. The primary outcome was the need for intensive care, and the model used was XGBoost. This study shows the potential of ML in predicting intensive care needs based on non-invasive parameters. The proposed model in this research not only utilizes non-invasive parameters but also leverages a wider time window, which provides a comprehensive view of the pandemic's progression and the associated changes in patient characteristics and outcomes. This approach facilitates a deeper understanding of the factors influencing hospitalization among COVID-19 patients, revealing distinct patient categories that could

inform personalized treatment plans.

2.4.7 Conclusion

In conclusion, the integration of AI has ushered in a new era of agile responses to the COVID-19 pandemic. Beyond its role in patient treatment, AI is very helpful to monitor and manage the health of infected individuals. This technology can operate on multiple scales, spanning from molecular insights to epidemiological trends, providing healthcare professionals with the tools to make informed decisions. Medical practitioners can develop personalized treatment regimens and preventive strategies tailored to each patient's unique profile, by using AI's predictive capabilities, ultimately improving patient outcomes.

Furthermore, AI-driven data analysis accelerates the extraction of meaningful insights from vast datasets, allowing researchers to identify patterns, potential treatment targets, and novel therapeutic interventions. The development of intelligent platforms and the adoption of non-invasive monitoring techniques represent innovative pathways in pandemic management. These approaches can enable real-time monitoring and assessment, ensuring timely interventions and resource allocation. As the global community continues to combat COVID-19 and prepares for future health challenges, the partnership between AI and healthcare exemplifies humanity's resilience and adaptability in using technology to solve these challenges.

Overall, AI has demonstrated its transformative power in the fight against the pandemic, offering a varied arsenal of tools to aid healthcare professionals and researchers. As this symbiotic relationship continues to evolve, the potential for AI-driven solutions to shape more effective and responsive healthcare strategies becomes increasingly evident.

2.5 Tools

The study employed diverse computational tools, all accessed and executed within the cloud-based Python development environment, Google Colaboratory [111], generously offered by Google. Additionally, we used GitHub [112] as the repository for hosting all the project notebooks.

- **Python:** The primary programming language used in this research was Python. The specific version used should be confirmed in the Google Colaboratory environment.

- **Pandas:** Pandas (1.5.3.) used for data manipulation and analysis.
- **NumPy:** NumPy (1.22.4) used for numerical computations.
- **Matplotlib:** Matplotlib (3.7.1.) used for data visualization.
- **Seaborn:** Seaborn (0.12.2) is a statistical data visualization library based on Matplotlib, used for creating more informative and attractive statistical graphics.
- **Scikit-learn (sklearn):** Scikit-learn (1.2.2.) used to perform various ML tasks.
- **XGBoost:** XGBoost (1.7.5.), another gradient boosting framework, was used for building predictive models.
- **Imbalanced-learn (imblearn):** Imbalanced-learn (0.10.1) used to tackle the issue of imbalanced datasets.

Chapter 3

An overview of COVID-19 public data

This chapter is about the process of dataset selection and analysis. We begin by outlining the key criteria and considerations that guided our selection process, highlighting the great impact of data completeness, structure, and relevance in our decision-making. Furthermore, we provide a comparative examination of three publicly available COVID-19 datasets, presenting their strengths and limitations. The chosen dataset becomes the base of our research, as it will establish the subsequent chapters' analysis and model development. In addition, we shed light on the challenges associated with missing data and emphasize the significance of data quality in ensuring the integrity of our study.

3.1 COVID-19 Datasets

The pandemic has resulted in the gathering of data, encompassing publicly accessible general information as well as sensitive patient-specific data. While general information is readily available online, the healthcare sector generates a trove of confidential data, including patient records, subject to stringent access controls. These same access restrictions apply to COVID-19 data. This section aims to present the challenges associated with COVID-19 datasets, offering insights into various examples and providing succinct explanations to illuminate these points of view.

The development of accurate and effective AI systems leans on the structure and quality of the dataset. The method of data collection, storage, and the inclusion of specific variables within the dataset play a vital role in this process. Concerning the context of

COVID-19, various types of datasets exist, ranging from publicly available ones, such as state-level statistics, to private datasets that require special access permissions. Some datasets are created by researchers for specific purposes, as exemplified in the work by Yelure et al. [109], where sounds are recorded and categorized for building a ML model. However, datasets of this kind are often kept private and not widely shared within the research community. This underscores the significance of data quality and access in developing AI systems, especially in the context of healthcare and pandemic-related research.

Regarding publicly available datasets, several public COVID-19 datasets contain unstructured data. The CORD-19 dataset [76] contains scientific papers on COVID-19, and the COVID-19 Image Data Collection contains chest X-rays and CT scans of patients [113] are two examples of datasets containing unstructured data.

A considerable amount of work on diagnostics, screening, classification, disease prediction, and medication development has already been done using not only tabular data but mostly CT-Scans, X-rays, and MRI images [114]. Nevertheless, other topics, such as contact tracing [115], projecting mortality and case numbers [97], reducing the workload of healthcare professionals [99], preventing disease [51], and monitoring treatment [109] have received less attention [72].

Concerning tabular data, some examples related to COVID-19 patients are datasets containing information on patient demographics, laboratory test results, and medical history. Such organized data helps in keeping track of how patients are doing, spotting symptoms, and understanding how the disease is developing [116]. The structured format of this data makes it easier to study, which can potentially lead to better patient outcomes [117]. There are various COVID-19 patient datasets that follow this structured pattern, like the COVID-19 Open Data Portal [118] and the dataset provided by the Johns Hopkins Coronavirus Resource Center [119]. Researchers have used these datasets to explore things like what factors make someone more likely to get sick, how severe the illness can become, and the rates of people dying from the disease [120].

Using this form of data (tabular data) allows for the monitoring of patients' conditions, symptom identification, and the tracking of disease progression, as highlighted in Gordon et al. (2020) [116]. Structured data also facilitates a more straightforward analysis of patient information, potentially leading to enhanced patient outcomes, as emphasized in Wallace et al. [117]. Various COVID-19 patient datasets with structured data are available, including the COVID-19 Open Data Portal dataset [118], encompassing data from

multiple countries, and the Johns Hopkins Coronavirus Resource Center dataset [119]. These datasets have been instrumental in analyzing COVID-19 risk factors, disease severity, and mortality rates, as demonstrated in research such as that conducted by Pericles et al. [120].

J. Zhang et al. emphasize that laboratory test results contain pertinent information and serve as effective markers of disease severity [86]. Nevertheless, these test outcomes were deemed less relevant due to the infrequent nature of post-discharge blood tests for patients, especially during home quarantine periods. Demographic details like age and gender help in analyzing how the virus affects different groups. Information about symptoms helps in recognizing common signs of the disease. Patient outcomes indicate the effectiveness of treatments and interventions. Comorbidities reveal underlying health conditions that could worsen COVID-19 outcomes. Dates provide a timeline for tracking the progression of the disease. Collecting and analyzing such data is crucial for developing strategies, making informed decisions, and improving healthcare responses during the pandemic [121]. Consequently, this dissertation focused only on the previously mentioned data types.

We identified three datasets that hold information on COVID-19 patients: two of them containing covid patients' characteristics, such as the ones mentioned before - demographics, relevant dates, symptoms, comorbidities, and patient outcome - and one dataset containing the same features except for the symptoms and the comorbidities.

In chronological order, starting from the earliest inception to the most recent data within the initial time frame, let's introduce these datasets. The first dataset, known as Novel COVID-19 (nCov2019) [122], contains comprehensive information on COVID-19 cases reported in Hubei and various other Chinese provinces. Following that, the second dataset, Data Science for COVID-19 (DS4C) [123], was collaboratively developed by the Korea Centers for Disease Control and Prevention (KCDC) and Seoul National University Bundang Hospital in South Korea. Lastly, we have the third dataset, named TriCovB [99], which was specifically designed for COVID-19 triage purposes within hospitals in Brazil.

Table 3.1 summarizes the ML studies performed on these identified datasets. These studies predicted COVID-19 outcomes, classified patients based on disease severity, and identified potential risk factors.

TABLE 3.1: COVID-19 Patient Datasets and Predictive Tasks

Dataset	Predictive Tasks
nCov2019 [122]	Prediction of mortality rate [97] Assess risk factors of mortality [124]
DS4C [123]	Prediction of isolation, released, and deceased states [125] Prediction of the number of recovered and deceased cases [126] Prediction of the number of days to recover [127]
TriCovB [99]	Severity Assessment [99]

3.2 Analysis and Pre-Processing

This section begins by describing the datasets and then focuses on the pre-processing. The following subsections will present each dataset and the steps that were taken to obtain the final result. Our goal is to ensure that the data is properly prepared and meets the quality criteria for further analysis and ML tasks. Initially, the initial size of the datasets will be considered, followed by the removal of lines containing missing values in any of the relevant features (*simple data*). Subsequently, their sizes will be reassessed once the datasets have been tidied to determine if they possess a feasible size for the continuation of the analysis and the execution of the remaining ML tasks.

3.2.1 nCov2019

The nCov2019 dataset [122] encompasses a comprehensive collection of patients information predominantly gathered from China for three months (December 2019 - February 2020). The dataset comprises 31 distinct features and consists of 18,527 entries. The extensive features include patient ID, age, sex, location, relevant dates, symptoms, comorbidities, travel history, administrative units, source of patient information, and patient outcomes.

In this study, the approach adopted was to consolidate the analysis by removing unnecessary features. Attributes like patient ID, location-related information, travel data, and data source details were deliberately omitted. This consolidation resulted in a refined dataset containing nine essential attributes: age, gender, confirmation date, date of death, release date, symptoms, comorbidities, and patient outcomes.

During the preliminary analysis, the percentage of missing values was computed for each category within the dataset. The dataset exhibited an alarmingly high proportion of missing values. Consequently, all entries containing missing values were excluded from

further analysis. Therefore, the final dataset consisted of 39 entries, rendering further analysis unfeasible.

In addition to the limited size of the dataset, several significant issues necessitate acknowledgment. Foremost among these concerns is the datasets reference period, which predates the WHOs declaration of the COVID-19 outbreak as a pandemic. Furthermore, notable challenges arise regarding the datasets demographic feature, namely age. Although the dataset contains many missing age values, instances where age values are present often exhibit wide age intervals, such as 15-88, or provide precise age values, such as 42. Moreover, the symptom and comorbidity features are presented in a textual format lacking a standardized protocol. Consequently, the processing of this information is complicated by multiple entries with different names but representing the same symptom or disease.

3.2.2 DS4C

The DS4C dataset [123], referred to as the Data Science for COVID-19 dataset, encompasses a period of six months (January 2020 - June 2020). It comprises multiple tables containing detailed information on the number of COVID-19 cases, patient-related data, time series data illustrating the progression of case numbers, and supplementary information such as weather data. However, for this study, the primary focus was on the patient information table, as it was the only table containing essential data regarding demographics, significant dates, and patient outcomes.

Initially, the patient information table encompassed 14 different features for 5,165 patients. However, only seven pertinent features were retained through a careful selection process. These included age, sex, confirmation date, release date, deceased date, patient state, and a newly derived feature called ”quarantine duration”. The sex of the patients was represented as binary values, while age was categorized into 10-year intervals. The relevant dates considered for analysis were the confirmation date and either the deceased or release dates. The patient state variable encompassed three distinct values: ”deceased,” ”isolated,” or ”released.” It is noteworthy to mention that this dataset did not contain information regarding patient symptoms or comorbidities.

Following an initial phase of data preparation, entries with the patient state labeled as ”isolated” were removed from consideration. This selection focused solely on cases that had already reached a definitive status of recovery or death. Entries with negative values

in this new column were subsequently eliminated from the dataset. The final version of the dataset comprised the following seven characteristics: age, sex, confirmation date, release date, deceased date, quarantine duration, and patient state.

3.2.3 TriCovB

The TriCovB dataset, as documented in Galo et al. (2022) [99], offers an extensive and comprehensive collection of data spanning over a year, from January 2020 to July 2021. This dataset holds an impressive array of 45 features and a substantial 1,679,329 entries. It includes detailed patient information, containing both individuals who are positive for COVID-19 and those not affected. This data compilation includes all the aforementioned features, ranging from patient demographic information and relevant dates to symptoms, comorbidities, and patient outcomes. Notably, the dataset organizes gender as binary values and presents age through two distinct features, one indicating precise age and the other denoting age intervals. Furthermore, the symptoms and comorbidities in the dataset were encoded using the one-hot encoding technique.

Multiple dates are included within the relevant dates category; however, for this study, only diagnostic, release, and decease dates were considered. Furthermore, patient outcomes encompass not only ”discharged” or ”deceased” but also additional values such as ”deceased, but not from COVID-19” and ”not infected.”

To prepare the dataset for analysis, a series of filtering operations were performed to eliminate irrelevant information or features. The initial step involved filtering entries specifically related to confirmed cases of COVID-19, as this study focuses on monitoring COVID-19 patients who were sent home. Features such as location-related attributes, test and lab results, and social attributes were removed as they are not pertinent to the scope of this study.

Additionally, travel information columns were eliminated, as they are irrelevant to the study’s objectives. Relevant date features were further filtered, and only confirmation, decease, and release dates were retained. These features were utilized to calculate the duration of quarantine (in days), which was subsequently used to create a new feature. The ”Ethnicity” column was retained for further analysis to explore potential correlations with other attributes, as it is easily accessible when a patient is at home.

To ensure data quality, entries with negative values in the quarantine duration column were excluded, and any missing information within the symptoms, ethnicity, comorbidities, or relevant dates columns was also eliminated. After implementing these filters, the dataset was refined, resulting in a final version comprising 188,383 entries and encompassing 22 features. These features can be categorized into various aspects, including relevant dates, quarantine duration, patient outcomes, hospitalization status, demographics, ethnicity, symptoms, comorbidities, and additional patient details. The organization entails seven symptom variables encoded as one-hot vectors, six comorbidity variables encoded similarly, age represented both as an exact value and within 10-year intervals, sex, ethnicity, three relevant dates, and patient outcomes. This meticulously curated dataset subsequently served as the foundation for subsequent analyses.

3.3 Comparison of Datasets

A summary of the three COVID-19 datasets under investigation, namely DS4C, nCov2019, and TriCovB, is presented in Table 3.2. Each dataset was analyzed based on important features like their sizes, starting and ending dates, leading to the time they cover, whether they include age and sex information, whether they provide details about symptoms and other health conditions, and whether relevant dates are present.

TABLE 3.2: Summary description of the three public COVID-19 datasets.

	nCov2019 [122]	**DS4C** [123]	**TriCovB** [99]
Original Size	18,527 × 32	5,165 × 14	1,679,329 × 45
Final Size	39 × 9	1,633 × 6	189,625 × 24
Period	Dec. 2019 - Feb. 2020	Jan. 2020 - Jun. 2020	Jan. 2020 - Jul. 2021
Age	one variable: 10-year intervals and precise age	one variable: 10-year intervals (e.g. 20s, 30s,)	multiple variables: one per each 10-year interval and one with precise age
Sex	M/F	M/F	M/F
Symptoms	format: text	NA	format: one-hot encoding
Comorbidities	format: text	NA	format: one-hot encoding
Relevant Dates	symptom onset, hospitalization, confirmation, release/death	symptom onset, confirmation, release, death	notification, testing, diagnose, register, release, death
Outcome	dead, alive	dead, alive, isolated	dead, released, not infected

The nCov2019 dataset contained information on symptoms and comorbidities in text format instead of readily usable one-hot encoding variables. The occurrence of missing values was evident, and this can be better understood through the analysis presented in Figure 3.1. Subsequently, following the dataset's pre-processing, its size was reduced, resulting in a final count of 39 records with 9 features. This considerable reduction in size notably constrains its potential for facilitating a thorough data analysis as needed for our specific task.

The DS4C dataset was reduced after pre-processing, indicating that a large portion of the data was either irrelevant or incomplete for the purposes of this dissertation. Furthermore, it lacked crucial information on symptoms and comorbidities, which are essential variables in COVID-19 studies.

On the other hand, the TriCovB dataset, with its massive original size, still retained a substantial amount of data after pre-processing. It also provided a more detailed representation of age and included one-hot encoding for symptoms and comorbidities, which is a more suitable format for ML tasks. Moreover, it covered a longer period and included more relevant dates, which could provide more comprehensive insights into the progression of the disease.

Figure 3.1 presents a series of bar plots that visually represent the percentage of missing values for each type of feature in each dataset. The bar plots offer a clear and concise way to understand the completeness of the data in each category, highlighting areas where data may be sparse or missing. The presence of missing data is a crucial factor in data analysis, as it can substantially influence the outcomes and the derived conclusions.

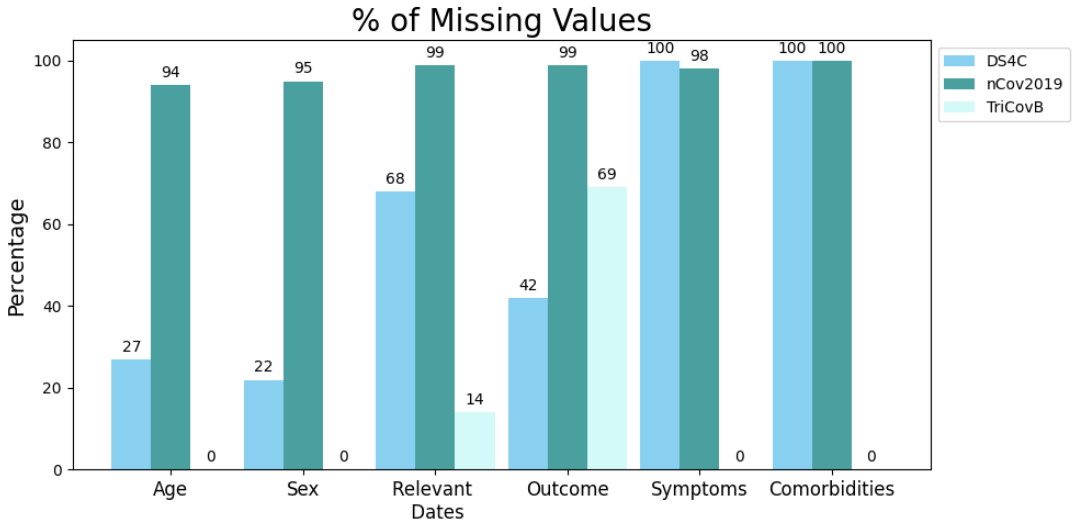


FIGURE 3.1: Percentage of Missing Values on each Dataset

In the case of the DS4C dataset, there is a considerable percentage of missing data across all variables. Specifically, the Relevant Dates and Outcome variables exhibit more than 40% missing data. This high percentage of missing values could potentially undermine the efficacy of any analysis conducted using this dataset, given that a significant portion of the data is absent.

The situation is even more pronounced with the nCov2019 dataset, which displays an exceedingly high percentage of missing values. Specifically, the Age, Sex, Relevant Dates, and Outcome variables all have over 90% missing data. This severe deficiency implies that the vast majority of the data for these variables is absent, which could result in biased or unreliable outcomes. The near-total absence of Symptoms and Comorbidities variables further restricts the utility of this dataset for COVID-19 studies.

In stark contrast, the TriCovB dataset exhibits a significantly lower percentage of missing values for the majority of its variables. Both the Age and Sex variables are completely filled, and the Relevant Dates variable has a mere 14% missing data. However, the Outcome variable does present a relatively high percentage of missing values at 69%, which could pose a potential issue.

3.4 Conclusions

The process of dataset selection for this study involved a careful evaluation of the DS4C, nCov2019, and TriCovB datasets. Each dataset was assessed based on its completeness, structure, size, and relevance to the research objectives.

The DS4C dataset, despite its substantial size, was not selected for this study. The primary reason for this decision was the dataset's lack of information on patient symptoms and comorbidities. These variables are crucial for our study, as they provide essential insights into the patients' conditions. Without this information, the DS4C dataset does not fully meet the requirements of our research.

Similarly, the nCov2019 dataset was also not chosen for this study. The reported period of this dataset predates the declaration of the pandemic by the WHO, which limits its relevance to our research objectives. Furthermore, the size of the dataset, after the cleaning process, was not sufficient to support robust analysis. These factors collectively rendered the nCov2019 dataset less suitable for our study.

The TriCovB dataset was selected for further analysis. This decision was based on the several key strengths of the dataset. Firstly, the TriCovB dataset has a minimal percentage of missing values, which sets it apart from the other datasets considered. This completeness provides a robust foundation for our research. Secondly, the TriCovB dataset is characterized by its tidy structure and substantial size, with nearly 200,000 entries. The dataset encompasses a comprehensive range of information, including detailed demographics, pertinent dates, symptoms, comorbidities, and patient outcomes. This wealth of data enables us to gain a thorough understanding of the patients and their respective conditions, which is indispensable for conducting effective analysis.

In summary, the selection of the TriCovB dataset for this study was guided by its completeness, structure, size, and relevance to the research objectives. Despite the high percentage of missing values in the Outcome variable, the dataset still offers the most comprehensive and reliable data among the three datasets. However, its important to note that the high percentage of missing values in the Outcome variable should be acknowledged in the analysis, as it could potentially introduce bias or uncertainty in the results.

Chapter 4

Case Study on Brazil COVID-19 Hospitalization

Studying the TriCovB dataset through exploratory data analysis (EDA) can help uncover insights and spot possible reasons for negative outcomes. This involves looking at details like patient characteristics, symptoms, and results, as well as other factors. Performing an EDA allows a closer look at the dataset to understand its main parts, such as distribution, variance, and relationships. Using tools like pictures and simple summaries, an EDA shows patterns and possible connections in the data.

This chapter begins by providing a more comprehensive introduction to the TriCovB dataset (previously mentioned in Chapter 3), then proceeds to focus on Univariate Analysis, followed by Multivariate Analysis, and ultimately grouping similar elements together.

4.1 Dataset Description

In this section, our aim is to provide an in-depth exploration of each feature category, as introduced in Chapter 3. This endeavor is designed to empower the reader with a comprehensive grasp of the dataset's composition, prioritizing a nuanced comprehension of the dataset's nature over a mere focus on feature types:

- **Relevant Dates:** This feature group contains dates in a specific format, which may represent significant events or milestones related to each patient. These dates can provide crucial temporal context during the analysis.

- **Quarantine Duration:** This numerical feature denotes the duration of the patient's quarantine period, measured in days. It reflects the amount of time the patient spent in isolation or under observation.
- **Patient Outcome:** This binary feature indicates the patient's ultimate outcome. A value of '0' signifies that the patient has recovered from COVID-19, while a value of '1' indicates that the patient passed away.
- **Patient Hospitalization:** This binary feature captures whether or not a patient requires hospitalization. A value of '0' denotes no hospitalization, while a value of '1' indicates that the patient was admitted to a hospital.
- **Demographics:** This feature group encompasses demographic information about the patients, including sex and age. The 'sex' feature is represented as a binary variable, with '0' denoting male and '1' denoting female. The 'age' feature is a numerical variable measured in years, indicating the patient's age.
- **Ethnicity:** This categorical feature represents the ethnicity of each patient. It consists of five distinct classes, allowing for an analysis of the potential impact of ethnicity on hospitalization.
- **Symptoms:** This feature group consists of seven one-hot encoded variables, each representing a specific symptom exhibited by the patient. The possible symptoms are fever, cough, headache, difficulty breathing, runny nose, sore throat, and diarrhea. By examining these symptoms, it is possible to explore their individual and collective relationship with hospitalization.
- **Comorbidities:** Similar to the symptoms group, the comorbidities feature group comprises six one-hot encoded variables, reflecting the presence or absence of specific comorbidities in each patient. The comorbidities covered in the dataset are obesity, diabetes, smoking, cardiac, lung, and renal problems. Analyzing these comorbidities can provide insights into their association with the likelihood of hospitalization.
- **Extra Patient Information:** This group includes additional patient-related information, such as whether the patient is pregnant, has any incapability, or works as a healthcare professional. These variables offer supplementary context that may influence the patient's hospitalization status.

4.2 Univariate Analysis

Univariate analysis is a fundamental component of EDA that focuses on examining individual variables in isolation. In this section, we delve into the univariate analysis of the TriCovB dataset to gain a deeper understanding of each feature's characteristics and its relationship with the target variable. The primary objective of the univariate analysis is to explore the distribution, central tendency, and variability of each feature independently.

Numerical attributes, like age and quarantine duration, are examined to discern their central tendencies and dispersion. Categorical and binary features also undergo scrutiny, revealing frequency distributions that shed light on the prevalence of each category or value. Complementing these statistics, the deployment of visualizations – encompassing histograms, box plots, bar charts, and pie charts – aids in intuitively conveying feature distributions and proportions, facilitating the identification of noteworthy patterns or discrepancies.

4.2.1 Relevant Dates

The temporal constraints of the available data necessitated a meticulous examination of pertinent dates. The dataset spanned a period of eighteen months, leading to an unequal representation of the months within a year. The first half of the year was twice as represented as the second half. To rectify this imbalance, a specific time frame was defined for our analysis: from July 2020 to June 2021.

This period represents a critical phase in the global trajectory of the COVID-19 pandemic, marked by significant developments and challenges. It is the most recent full-year window available in our dataset, making it particularly relevant for informing current and future public health strategies. During this period, the world witnessed the rise and fall of multiple waves of infections, the emergence of new variants of the virus, and the initiation of global vaccination campaigns [128]. The start of this window, July 2020, saw many countries grappling with the aftermath of the first wave and preparing for potential subsequent waves [129]. By the end of this window, in June 2021, many countries had launched vaccination campaigns and were starting to see the impact of these efforts on case numbers and hospitalizations [130].

Upon resolving the temporal limitations, the investigation turned toward the exploration and analysis of the temporal distribution of COVID-19 cases. The diagnosis date served as the foundation for this analysis. Two methodologies were employed: monthly

segmentation and seasonal segmentation. These techniques facilitated a comprehensive understanding of potential temporal patterns or trends in the incidence of COVID-19. Figures 4.2 and 4.1 visually compare these two methodologies, illustrating the monthly and seasonal distributions.

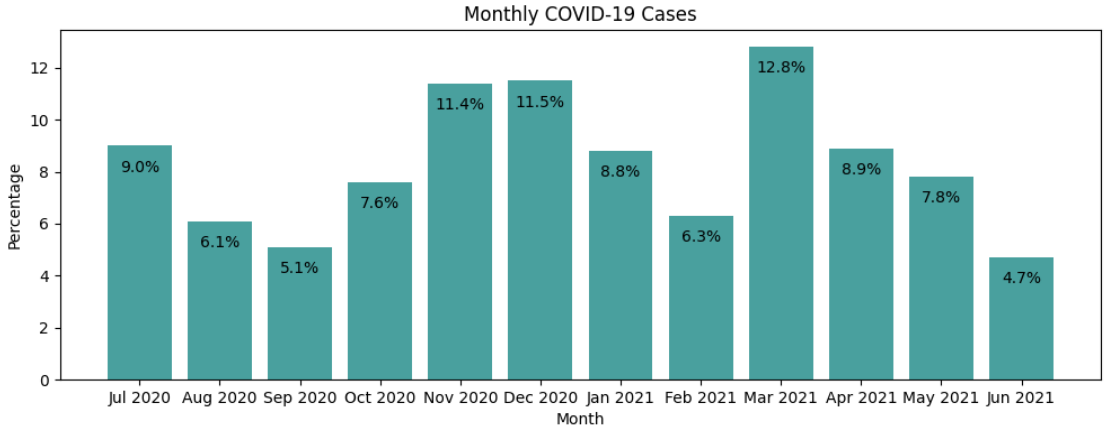


FIGURE 4.1: Monthly COVID-19 Cases

The monthly distribution of cases within this time frame is represented in Figure 4.1. March 2021 accounted for the highest number of cases, approximately 12.8% of the total, followed by November 2020 and December 2020 with 11.4% and 11.5% respectively. June 2021 recorded the least number of cases, contributing only 4.7% to the total.

A parallel analysis was conducted for the seasonal distribution of cases, as depicted in Figure 4.2. Spring had the highest case count, constituting 34.2% of the total cases, followed by Winter and Autumn with 27.1% and 20.8% respectively. Summer recorded the least cases, contributing approximately 17.9% to the total.

In terms of seasonal distribution, the highest number of cases were reported in Spring, followed by Winter, Autumn, and Summer. This could be due to the fact that respiratory viruses, including the one that causes COVID-19, often show seasonal variation with higher transmission rates in colder months [131]. However, the impact of seasonality on COVID-19 is still not fully understood and is likely to be influenced by a combination of factors including human behavior, host immunity, and environmental conditions [132].

These findings are in line with some of the existing literature on COVID-19. Belay et al. [133] highlighted the trends in the geographic and temporal distribution of COVID-19 cases among children in the US. Furthermore, a perspective by Dhanasekaran et al. [134] discussed the potential short and long-term evolutionary dynamics of seasonal influenza and the potential consequences as global travel gradually returns to pre-pandemic levels.

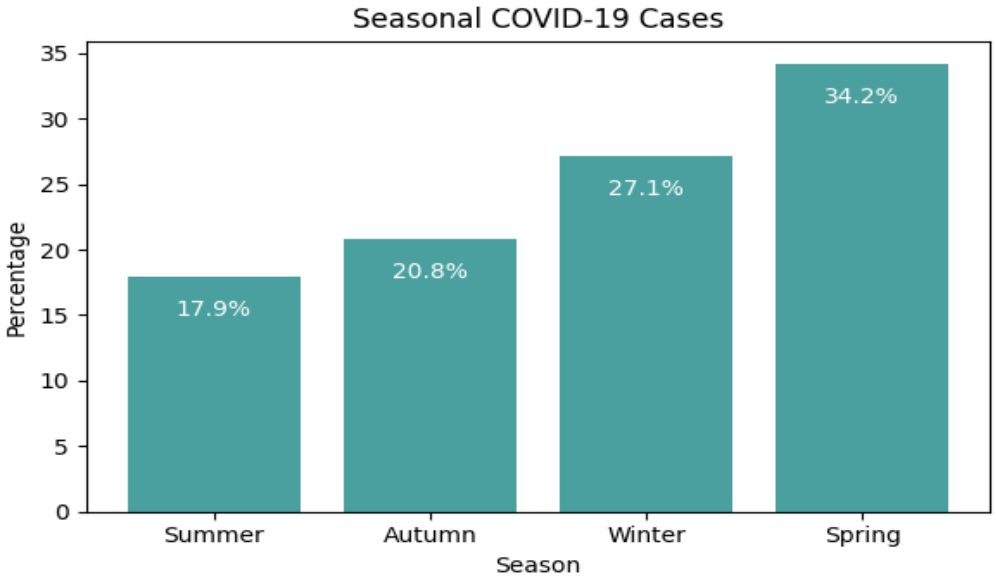


FIGURE 4.2: Seasonal COVID-19 Cases

4.2.2 Quarantine Duration

The quarantine duration is a numerical feature that denotes the length of time each patient spent in isolation or under observation. Measured in days, this feature provides an indication of the duration of the quarantine period. The quarantine duration feature was derived from the relevant dates in the dataset, specifically the difference between the end date and the diagnosis date.

The Quarantine Duration feature initially presented a maximum value of 7,485 days in the TriCovB dataset. This value was deemed anomalous, especially considering the timeframe under study, which spanned from January 2020 to July 2021, a total of 546 days. As depicted in Figure 4.3, a box plot visualization was instrumental in identifying two data points with quarantine durations exceeding 6,000 days. These were considered erroneous and subsequently removed from the dataset.

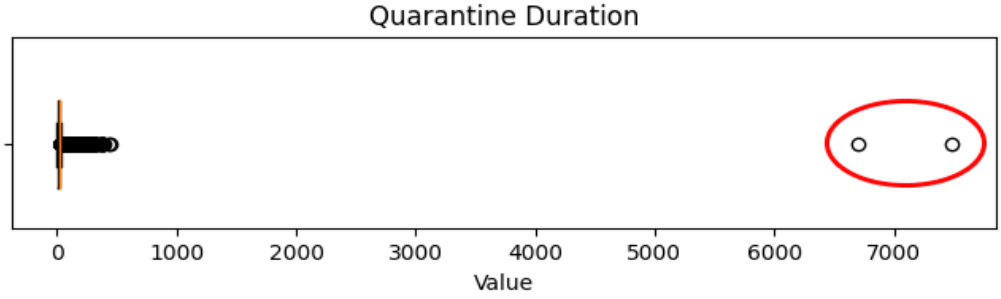
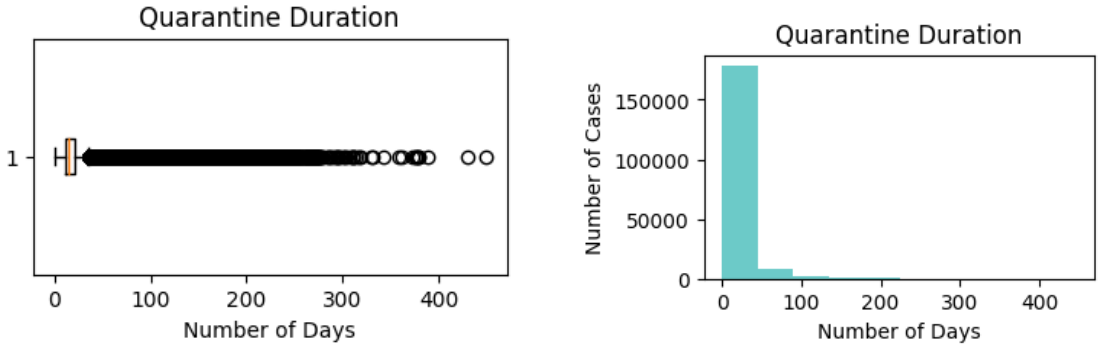


FIGURE 4.3: Erroneous Data Points

Post removal of these two data points, the revised dataset comprised 189,625 cases. The cleaned dataset presents a mean quarantine duration of approximately 20 days, with a standard deviation of around 20 days, indicating a substantial variation in quarantine durations among individuals. The minimum duration is 0 days, while the maximum is significantly reduced to 448 days. The interquartile range, extending from the 25th percentile (11 days) to the 75th percentile (21 days), encapsulates the middle 50% of the quarantine durations, with the median duration being 15 days.

The box plot in Figure 4.4a illustrates the distribution of Quarantine Duration after this data cleaning process. Notably, a significant number of data points lie beyond the box plot's right whisker, indicating numerous outliers with relatively long quarantine durations. These outliers represent individuals with exceptionally lengthy periods of quarantine compared to most of the population in the dataset.

The histogram in Figure 4.4b provides another perspective on the distribution of Quarantine Duration. The concentration of data towards the left of the histogram indicates a right-skewed or positively skewed distribution, suggesting that a larger proportion of individuals in the dataset had relatively short quarantine durations. However, the distribution's skewness is likely influenced by the presence of outliers, as indicated in the box plot.



(A) Box-plot of Quarantine Duration (B) Histogram of Quarantine Duration

FIGURE 4.4: COVID-19 cases distribution by Quarantine Duration

4.2.3 Patient Outcome

The patient outcome is a binary feature that indicates the ultimate outcome for each individual in the dataset. A value of '0' signifies that the patient was cured, while a value of '1' indicates that the patient passed away. Analyzing the patient outcome is essential in

understanding the overall prognosis and mortality rates within the dataset. By examining the distribution and proportions of these outcomes, we can gain insights into the severity of the disease and its impact on patient health. This information is crucial for assessing the effectiveness of healthcare interventions and identifying potential risk factors associated with poor outcomes.

Figure 4.5 presents the distribution of patient outcomes, specifically focusing on fatalities within the dataset. It provides a visual representation of the proportions of patients who unfortunately succumbed to the disease.

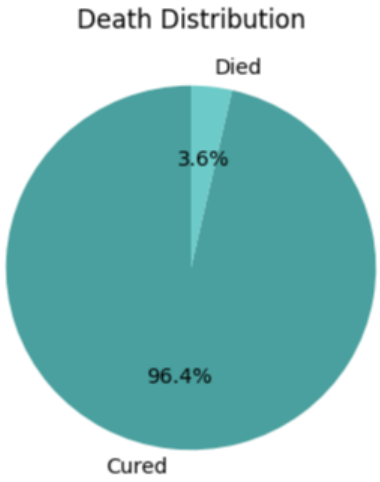


FIGURE 4.5: Fatalities Distribution

The analysis of COVID-19 cases within the dataset revealed a case fatality rate (CFR) of 3.6%. The CFR is a critical measure in epidemiology as it provides insights into the severity of a disease and the effectiveness of healthcare systems in managing it.

The observed CFR of 3.6% falls within the range reported in the literature for COVID-19, albeit on the lower end. For instance, a study conducted in Italy during the early stages of the pandemic reported a CFR of 7.2% [135]. A systematic review and meta-analysis of multiple studies found a pooled CFR estimate of 3.38% [136]. These figures, however, should be interpreted with caution as the CFR can be influenced by several factors. These include the demographic characteristics of the population, the capacity and quality of healthcare systems, and the strategies used to test, report, and manage cases.

In relation to our research, the relatively lower CFR could be influenced by a range of factors. These factors might encompass the specific time frame covered by the dataset, extensive testing and accurate case identification, a population composition that skews

towards younger age groups, as discussed in Section 4.2.5, or efficient healthcare and treatment approaches for individuals with COVID-19. It is also possible that a significant number of mild or asymptomatic cases were detected and reported, which would lower the observed CFR. In conclusion, the CFR of 3.6% observed in this study is consistent with the existing literature on COVID-19, although it is on the lower end of reported rates [137].

4.2.4 Patient Hospitalization

The patient hospitalization feature is a binary variable designed to capture the necessity of hospital care for each individual. A value of '0' signifies that hospitalization was not required, indicating that the patient did not undergo admission to a medical facility. Conversely, a value of '1' indicates that the patient was admitted to a hospital. The analysis of the patient hospitalization feature allows for a comprehensive understanding of the proportion of individuals who necessitated medical care beyond the standard quarantine period.

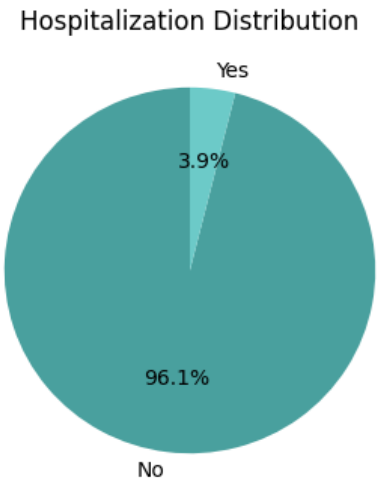


FIGURE 4.6: Hospitalization Distribution

A visual representation of the distribution of COVID-19 cases based on hospitalization status is provided in Figure 4.6. In the dataset under study, a hospitalization rate of 3.9% was observed. The hospitalization rate is a crucial indicator of the severity of the disease in the population and the potential strain on healthcare resources.

The value observed in this study is within the range reported in the literature. For instance, a study conducted in New York City during the early stages of the pandemic

reported a hospitalization rate of 21% among confirmed cases [138]. In Denmark, a nationwide cohort study reported a hospitalization rate of 17.2% among confirmed COVID-19 cases [139]. This rate is significantly higher than the one observed in our study, which could be due to differences in the demographic characteristics of the populations or the strategies for managing COVID-19 patients.

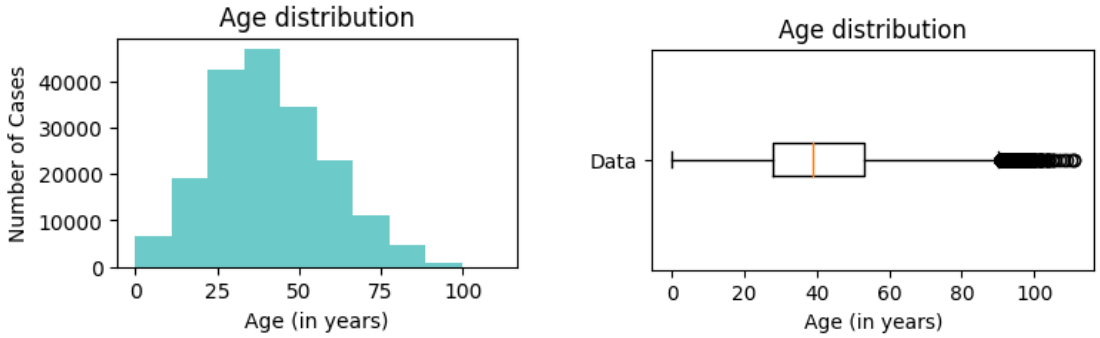
In the context of Brazil, a study developed a risk prediction algorithm for hospital admission due to COVID-19 and found that factors such as age, sex, ethnicity, and comorbidities significantly influenced the risk of hospitalization [140]. These factors could also explain the relatively low hospitalization rate observed in our study.

In conclusion, the hospitalization rate of 3.9% observed in this study is consistent with the existing literature, although it is on the lower end of reported rates. This finding underscores the importance of context-specific factors in influencing the outcomes of the COVID-19 pandemic. As the situation continues to evolve, ongoing research is needed to monitor these trends and inform public health strategies.

4.2.5 Demographics

The demographics feature group encompasses crucial patient information, including sex and age, which contribute to a comprehensive understanding of the dataset. The 'sex' feature is represented as a binary variable, with '0' denoting male and '1' denoting female. This enables an analysis of potential gender-based differences in hospitalization rates and outcomes. The 'age' feature, measured in years, provides valuable insights into the age-dependent vulnerability and risk factors associated with the disease.

The dataset exhibits a broad range of ages, spanning from newborns (minimum age of 0 years) to elderly individuals (maximum age of 111 years). The mean age of the dataset is approximately 41 years, with a standard deviation of around 18 years, indicating a notable dispersion in age among the individuals. It is worth noting that the mean age is slightly higher than the reported average age of 40 in the city of São Paulo, as documented in a previous study [141]. However, this disparity may be attributed to the inclusion of patients from diverse regions in Brazil, limiting the dataset's representativeness for the population in São Paulo. Further investigation is required to determine the statistical significance of this difference and identify potential contributing factors.



(A) Histogram of Age Distribution

(B) Box-plot of Age Distribution

FIGURE 4.7: COVID-19 cases distribution by Age

The interquartile range, ranging from the 25th percentile (28 years) to the 75th percentile (53 years), encapsulates the middle 50% of the ages, with a median age of 39 years. The distribution of age is graphically depicted in Figure 4.7.

The histogram in Figure 4.7a reveals a left-skewed or negatively skewed distribution, with a higher frequency of younger ages. This indicates a larger proportion of relatively young individuals in the dataset.

The box plot in Figure 4.7b further illustrates this skewness and identifies several outliers on the higher end of the age spectrum. These outliers, represented as points beyond the right whisker of the box plot, indicate individuals significantly older than the majority of the population in the dataset.

Regarding the gender distribution, a minor disparity is observed within the dataset. Females account for 54.7% of the cases, while males comprise 45.3%. Interestingly, this distribution contradicts existing literature, which often reports a higher incidence of COVID-19 cases in males compared to females [142].

4.2.6 Ethnicity

The ethnicity feature serves as a categorical variable that characterizes the ethnicity of each patient in the dataset. It encompasses five distinct classes, allowing for an analysis of the potential influence of ethnicity on hospitalization rates.

From the outset of our analysis, we recognized ethnicity as a potentially significant categorical feature, considering that certain diseases exhibit varying prevalence across different ethnic groups [143]. For instance, skin cancer is more commonly observed in individuals with lighter skin, including those of Caucasian descent [144]. Thus, we closely

monitored this feature and explored its potential associations with other variables to gain valuable insights.

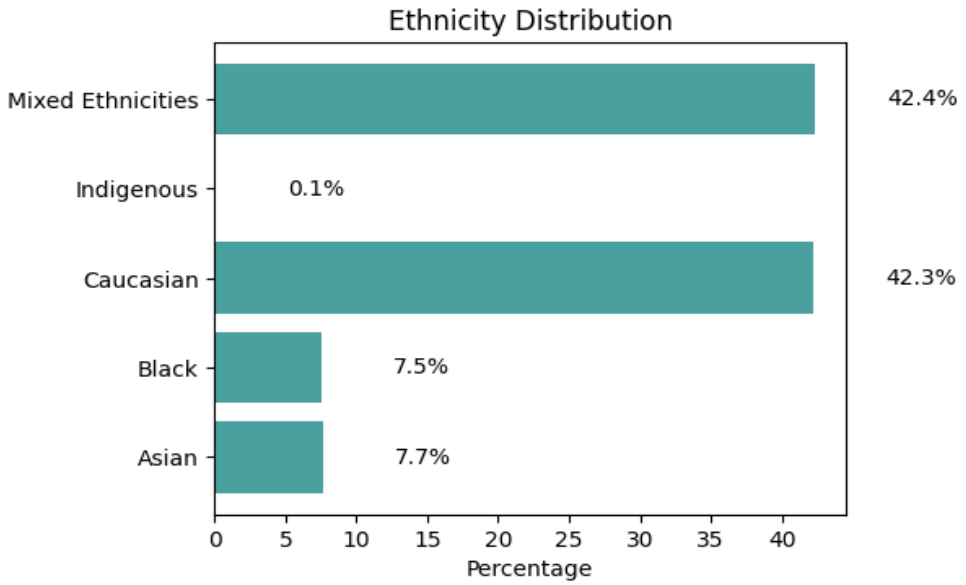


FIGURE 4.8: COVID-19 Cases Distribution by Ethnicity

The distribution of COVID-19 cases across different ethnicities in the dataset under study reveals certain disparities. Mixed ethnicities and white individuals each account for 42.3% of the cases, followed by Asian individuals at 7.7% and Black individuals at 7.5%. Indigenous people constitute the smallest group, with only 0.1% of the cases.

These disparities could be attributed to a combination of social, economic, and health-related factors. Socioeconomic factors could play a significant role in the observed ethnic disparities. For instance, individuals from certain ethnic groups might be more likely to live in crowded housing conditions, work in jobs with a higher risk of exposure to the virus, or have limited access to healthcare, all of which could increase their risk of COVID-19 [145].

Pre-existing health disparities could also contribute to the observed trends. Certain ethnic groups might have a higher prevalence of underlying health conditions that increase the risk of severe COVID-19, such as diabetes or cardiovascular disease. These health disparities could lead to a higher incidence of cases among these groups.

Differences in access to testing could also influence the observed distribution of cases. If testing is more accessible to certain ethnic groups, it could lead to a higher detection rate among those groups. This could result in an over-representation of these groups in the dataset.

Cultural factors, such as language barriers or mistrust in healthcare systems, could also influence the likelihood of seeking testing or healthcare, thereby affecting the observed distribution of cases. These factors could lead to an under-representation of certain ethnic groups in the dataset.

4.2.7 Symptoms

The symptoms feature group consists of seven one-hot encoded variables, each representing a specific symptom exhibited by the patient. By examining these symptoms, we can explore their individual and collective relationship with hospitalization. Analyzing the prevalence of symptoms allows us to identify symptom patterns that may be indicative of severe illness and the need for hospitalization.

Initially, we analyzed the symptomatology of the patients in our dataset. We found that the top-3 most common symptoms reported were cough, headache, and fever. Interestingly, our findings differ slightly from those reported in the literature [146], which indicate that the top-3 symptoms during the same period were fever, cough, and fatigue. However, since fatigue was not included as a symptom in our dataset, the next most prevalent symptom after cough and fever was dyspnea, which can be interpreted as difficulty breathing. Notably, difficulty breathing was the third least common symptom in our dataset.

The remaining symptoms and their corresponding percentages are presented in Figure 4.9, where they are displayed in descending order of prevalence.

The distribution of symptoms among COVID-19 patients in TriCovB reveals a diverse range of manifestations. The most prevalent symptom is a cough, reported in 57.03% of patients. This is closely followed by a headache, experienced by 55.44% of patients. Fever is the third most common symptom, reported in 45.77% of cases. Other symptoms include a runny nose, reported in 37.49% of patients, a sore throat, experienced by 32.17% of individuals, difficulty breathing, reported by 16.97% of patients, and diarrhea, reported by 16.06% of patients. Interestingly, 12.08% of patients report no symptoms, highlighting the potential for asymptomatic transmission of the virus.

These results provide valuable insights into the symptomatology of COVID-19 patients in TriCovB, which can inform future research and clinical practices. The symptoms of COVID-19 are a result of the body's immune response to the virus and the damage caused by the virus to different organs. The respiratory system is often the most affected,

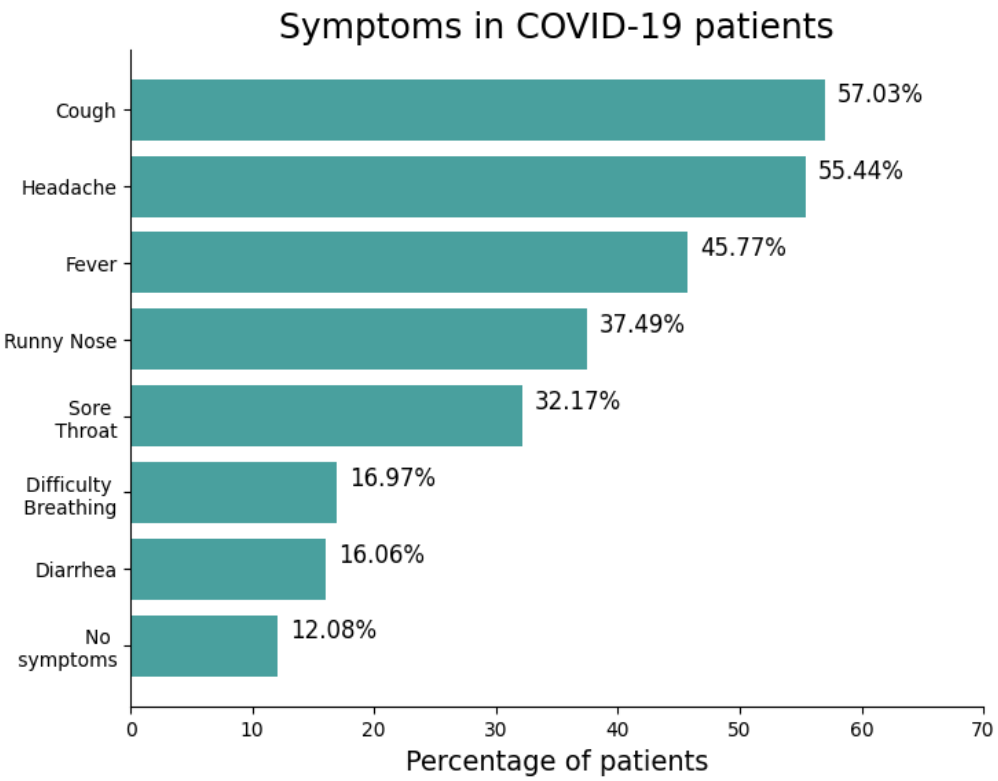


FIGURE 4.9: Percentage of patients with symptoms

which explains the high prevalence of cough and difficulty breathing. The virus can also affect other systems, leading to symptoms such as headache, fever, and diarrhea.

The range and severity of symptoms can vary widely among individuals, depending on factors such as age, sex, underlying health conditions, and genetic factors. This could explain the diverse range of symptoms observed in the dataset.

The existence of asymptomatic cases, constituting 12.08% of the patients in the dataset, is a recognized characteristic of COVID-19 [147]. However, this proportion is notably lower compared to findings reported in existing literature [148]. These individuals test positive for the virus but do not exhibit any symptoms. Asymptomatic cases pose a significant challenge in controlling the spread of the virus, as these individuals might unknowingly transmit the virus to others.

It's worth emphasizing that these are plausible explanations, and the precise causes might differ based on particular contexts and conditions.

4.2.8 Comorbidities

Similar to the symptoms group, the comorbidities feature group comprises six one-hot encoded variables. These variables reflect the presence or absence of specific comorbidities in each patient. This analysis aids in understanding the interplay between COVID-19 and pre-existing health conditions.

In this study, the three most prevalent comorbidities among patients in the dataset were cardiac conditions, diabetes, and obesity, in that order. Interestingly, these results differ from the literature [149], which suggests that the most common comorbidities in COVID-19 patients are hypertension, cardiovascular diseases, and diabetes. However, hypertension was not included in the dataset. Hence, the researchers compared the top three comorbidities common to both the dataset and the literature: cardiovascular diseases, diabetes, and chronic kidney disease. It is worth noting that renal conditions were found to be the least common comorbidity in the dataset.

The prevalence of the remaining comorbidities in the dataset and their corresponding percentages can be found in Figure 4.10, where they are displayed in descending order. These findings suggest that the distribution of comorbidities in COVID-19 patients may vary depending on the population studied, which highlights the importance of understanding the specific characteristics of each population to provide appropriate medical care.

The comorbidity profile among COVID-19 patients in the dataset under study is largely characterized by the absence of any reported comorbidities, with 74.76% of patients falling into this category. However, among those with comorbidities, cardiac issues are the most prevalent, affecting 18.11% of patients. Diabetes is the next most common comorbidity, reported in 6.83% of cases. Obesity is present in 3.38% of patients, while lung-related issues are reported in 3% of cases. Smoking, a risk factor for many health conditions, is reported in 2.12% of patients. Renal issues are the least common comorbidity, affecting just 0.6% of patients.

Older individuals are more likely to have comorbidities such as cardiac issues and diabetes [150]. Individuals with comorbidities are at a higher risk of severe COVID-19. Therefore, they might be more likely to get tested and be represented in the dataset.

Lifestyle factors such as diet, physical activity, and smoking can contribute to the development of comorbidities such as obesity, cardiac issues, and lung-related issues. These factors could influence the observed distribution of comorbidities.

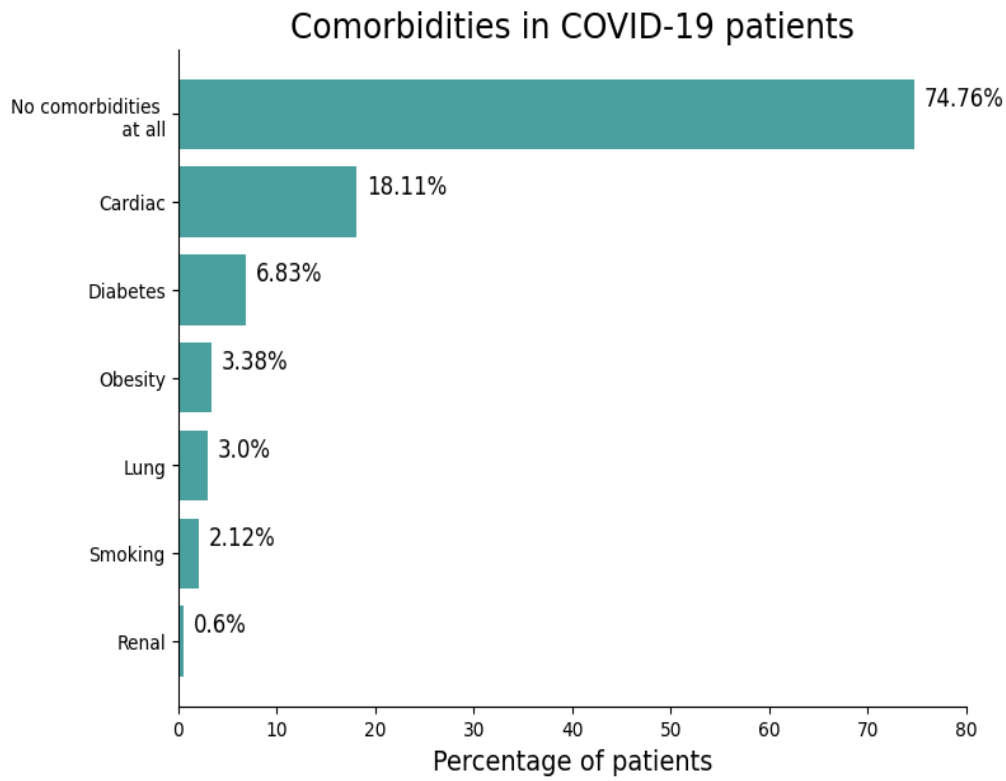


FIGURE 4.10: Percentage of patients with comorbidities

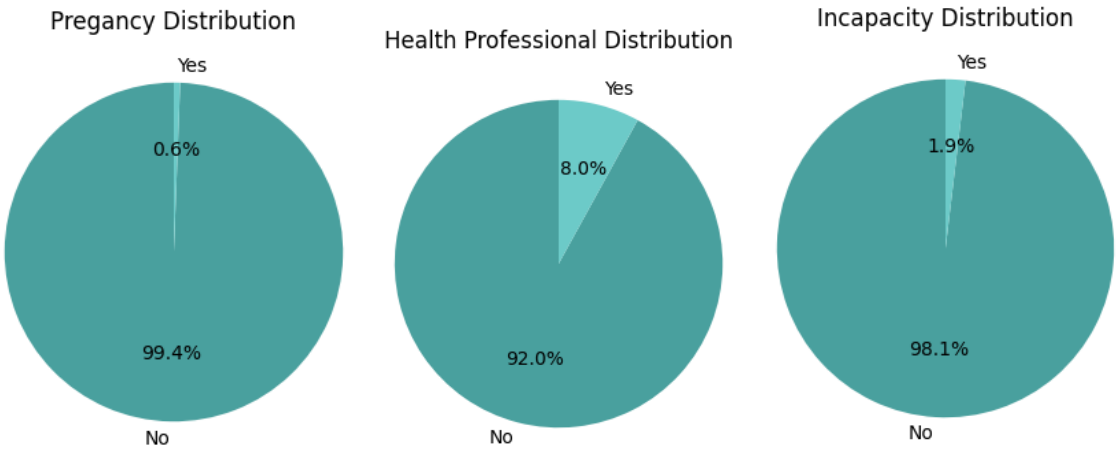
It's important to highlight that the high proportion of patients without any reported comorbidities could reflect the demographic profile of the dataset (e.g., younger age, healthier population) or testing strategies (e.g., widespread testing, including asymptomatic individuals).

4.2.9 Extra Patient Information

The extra patient information group includes additional patient-related variables, such as pregnancy status, incapability, and healthcare professional status. This information provides insights into the unique characteristics of certain patient subgroups and their potential vulnerability to severe COVID-19 outcomes.

The dataset reveals that 0.6% of the patients are pregnant, as depicted in Figure 4.11a. This finding suggests that pregnant individuals constitute a small proportion of the overall COVID-19 cases in the dataset.

Furthermore, the dataset indicates that 8% of the patients are health professionals, as shown in Figure 4.11b. This finding is noteworthy as it highlights the occupational exposure and potential vulnerability of healthcare workers to COVID-19. Analyzing the



(A) Pregnancy Distribution

(B) Healthcare Professionals Distribution

(C) Incapacity Cases Distribution

FIGURE 4.11: Extra Patient Information

hospitalization rates and patient outcomes specifically among healthcare professionals can provide insights into the effectiveness of infection control measures within healthcare settings and help identify any additional support or interventions required to safeguard the health and well-being of frontline workers.

In addition, the dataset reveals that 1.9% of the patients have some form of incapacity or deficiency, as presented in Figure 4.11c. This finding underscores the importance of understanding the impact of COVID-19 on individuals with pre-existing conditions or disabilities. This information can contribute to the development of targeted interventions and support systems to ensure equitable care for all individuals, regardless of their physical or cognitive abilities.

4.3 Multivariate Analysis

In the next phase, a multivariate analysis was conducted by examining the hospitalization feature paired with other variables. The dataset will be partitioned into two subsets: cases requiring hospitalization and cases not requiring hospitalization. This analysis aims to aid in the identification of factors that display significant variation between the two groups.

In this part, the hospitalization rates and patient outcomes were explored and analyzed. The analysis of TriCovB uncovered that severe outcomes were experienced by a relatively minor segment of the population within our dataset, as shown before in Figures 4.6 and 4.5. The distinction was made between survivors and non-survivors, aiming to

determine the proportion of individuals from each group that were hospitalized. This approach was taken with the intention of investigating whether hospitalization could serve as an indicator of unfavorable outcomes. Figure 4.12 illustrates these distributions.

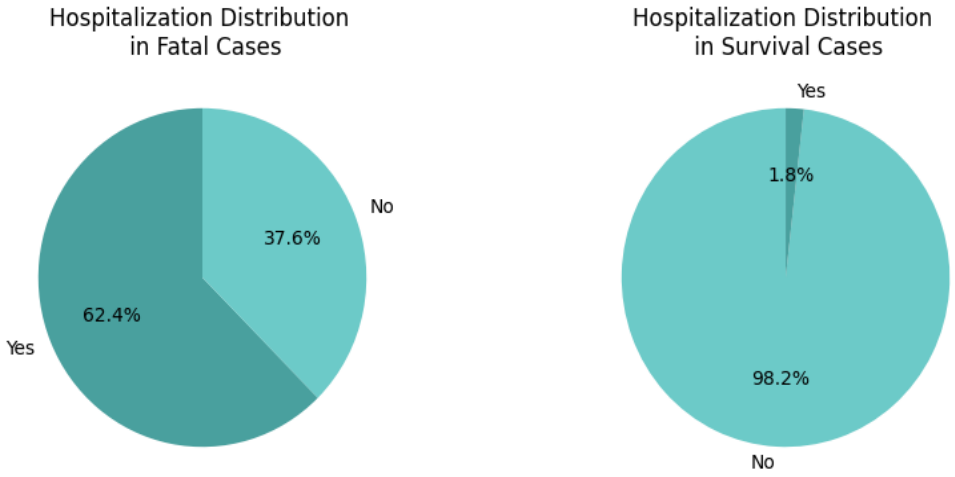


FIGURE 4.12: Hospitalization Cases by Survivance

It is noteworthy that 62.4% of the patients who unfortunately did not survive were initially hospitalized, while 37.6% were not. In contrast, a mere 1.8% of the survivors required hospitalization. This observation suggests that hospitalization is indeed a relevant indicator of unfavorable outcomes.

With our focus directed toward predicting COVID-19 patient hospitalization, subsequent sub-sections will present an overview and conduct a comparative analysis between the hospitalized and non-hospitalized groups, aiming to enhance comprehension of factors associated with hospitalization and, ultimately, mitigate unfavorable outcomes.

4.3.1 Age and Sex

In order to conduct a combined analysis of the age and sex features, we explore potential differences in age distributions between the sexes. The density graph in Figure 4.13 visually represents these distributions, offering insights into variations in the age distribution between males and females.

The multivariate analysis of the dataset reveals specific trends in the prevalence of COVID-19 cases, particularly an increased prevalence in older men compared to older women. Among patients aged 60 and above, the count of men surpasses that of women, indicating a higher incidence of confirmed cases in older men. This pattern is also observable in the age brackets of 10, 30s, and 50s. Conversely, women in their 20s, 40s, and

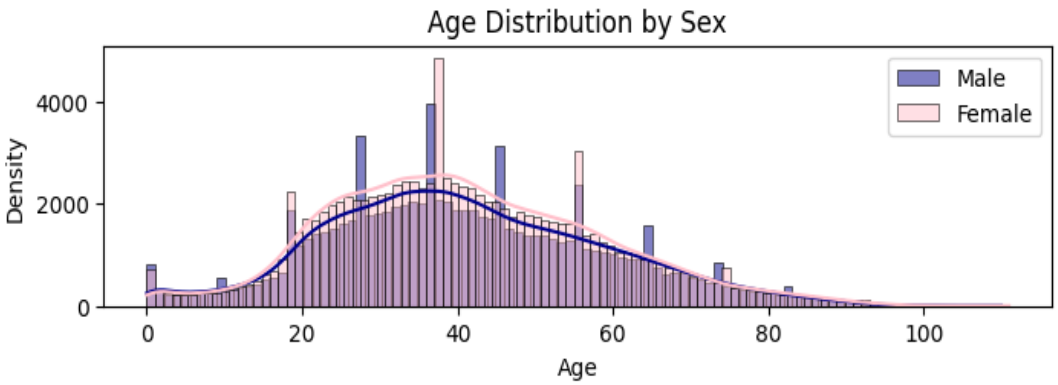


FIGURE 4.13: COVID-19 cases distribution by Sex

around the age of 55 exhibit a higher incidence rate. The most significant disparity in the number of cases between the genders is observed at around 40.

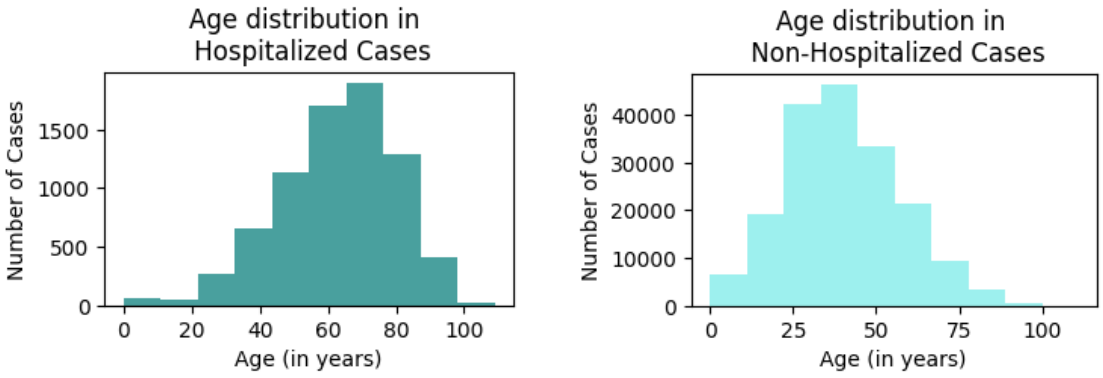
These observed gender and age-specific trends could be attributed to a combination of biological, behavioral, and social factors. Some studies suggest that biological differences between men and women could influence their susceptibility to infections, including COVID-19 [142]. For instance, sex hormones and the X chromosome in females have been associated with a stronger immune response, which could potentially explain the lower incidence of cases in women in certain age groups [151].

The observed patterns could also be influenced by lifestyle and behavioral elements. For instance, habits like smoking and alcohol consumption, which tend to be more common among men, could elevate the susceptibility to severe COVID-19 outcomes [152]. Moreover, men might display a greater tendency to seek medical attention and undergo testing, potentially resulting in a higher detection rate within specific age brackets [153].

To examine potential variations in age distribution, distinct age histograms were generated for hospitalized COVID-19 patients and non-hospitalized patients, as illustrated in Figure 4.14a and Figure 4.14b, respectively.

Variations in age distribution emerged between hospitalized and non-hospitalized COVID-19 cases. Hospitalized patients exhibited an average age of 63, somewhat lower than certain literature references suggest [154, 155]. Regarding non-hospitalized patients, the average age of 40 years old is slightly below the figure of 45 years reported in the literature [156]. This finding supports the conclusion that older individuals tend to face an elevated probability of requiring hospitalization.

Quartile analysis further reinforces this observation, demonstrating a larger proportion of elderly individuals among hospitalized cases compared to non-hospitalized cases.



(A) Hospitalized cases (B) Non-hospitalized cases

FIGURE 4.14: Age distribution by hospitalization outcome

Despite both groups sharing a similar maximum age (111 for non-hospitalized and 109 for hospitalized), the age range was more extensive among hospitalized cases, spanning from infants to older individuals.

Figure 4.15 provides a visual representation of the gender distribution, clearly delineating the comparative proportions of males and females within the two distinct categories of COVID-19 patients: those who required hospitalization and those who did not.

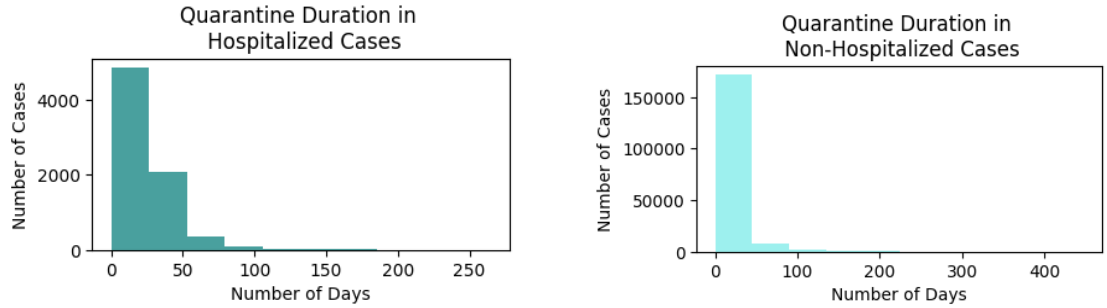
FIGURE 4.15: Sex distribution by hospitalization outcome

A slight imbalance in hospitalization rates based on gender was observed. Hospitalized patients were mainly male, comprising 53.4% of the cases, while females constituted the majority at 55% among non-hospitalized patients. These findings align with existing literature [157]. These results suggest a potential gender-based difference in COVID-19 outcomes, indicating that males might be more susceptible to a severe disease requiring hospitalization [158].

The results indicate a significant influence of age on the probability of COVID-19 patients requiring hospitalization, aligning with the existing literature's expectations. Advanced age is linked to a heightened vulnerability to severe conditions necessitating hospital admission.

4.3.2 Quarantine Duration

To investigate potential differences in quarantine duration distribution , we generated separate histograms. The histograms are presented in Figure 4.16a and Figure 4.16b, respectively.



(A) Hospitalized Cases

(B) Non-Hospitalized Cases

FIGURE 4.16: Quarantine duration distribution by hospitalization outcome

When comparing the quarantine duration between hospitalized and non-hospitalized cases, several notable differences emerge. For non-hospitalized cases, the mean quarantine duration is approximately 20 days, with a standard deviation of around 19 days. The minimum duration is 0 days, indicating individuals who may not have undergone a quarantine period.

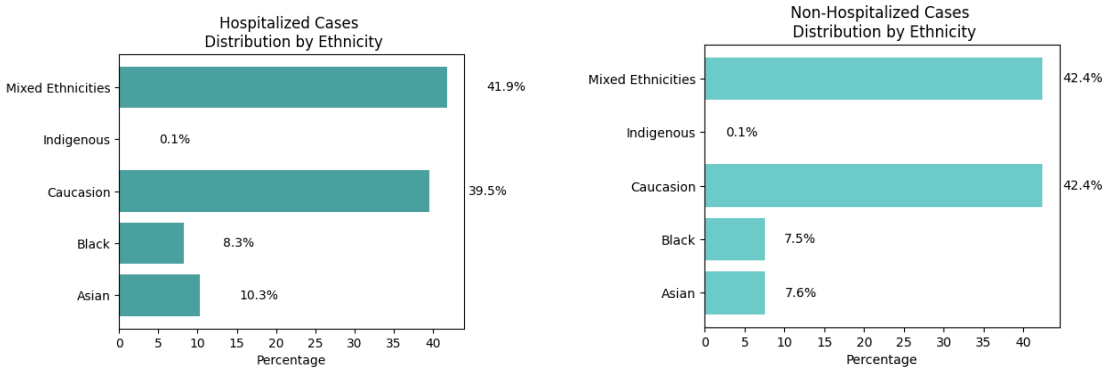
The quartile analysis reveals that 25% of non-hospitalized cases have a quarantine duration of 11 days or less, while 50% have a duration of 15 days or less. The 75th percentile indicates that 75% of non-hospitalized cases have a duration of 21 days or less. The maximum duration observed among non-hospitalized cases is 448 days.

On the other hand, for hospitalized cases, the mean quarantine duration is approximately 63 days, with a standard deviation of around 17 days. Similar to non-hospitalized cases, the minimum duration is 0 days. The quartile analysis shows that 25% of hospitalized cases have a duration of 52.5 days or less, while 50% have a duration of 65 days or less. The 75th percentile indicates that 75% of hospitalized cases have a duration of 75 days or less. The maximum duration observed among hospitalized cases is 109 days.

These findings indicate a substantial disparity in quarantine duration between the two groups. Non-hospitalized cases tend to have shorter quarantine periods, with the majority falling within the range of 15 to 21 days. In contrast, hospitalized cases have significantly longer quarantine durations, with the majority falling within the range of 65 to 75 days. It is important to note that the comparison of quarantine duration between the two groups should be interpreted with caution.

4.3.3 Ethnicity

This dissertation explored how different ethnic groups relate to the chances of getting hospitalized due to COVID-19. We can see if there are any differences, by comparing the numbers of hospitalized and non-hospitalized cases in each ethnic group. This helps us understand if some ethnic groups are more likely to end up in the hospital or not. The outcomes are visualized in Figure 4.17.



(A) Hospitalized Cases (B) Non-Hospitalized Cases

FIGURE 4.17: Ethnicity distribution by hospitalization outcome

Figure 4.17a illustrates the ethnic distribution among hospitalized COVID-19 cases, showcasing that Caucasian and Mixed Ethnicities are prevalent, comprising 39.50% and 41.87% of cases, respectively. Asian and Black ethnic groups account for 10.29% and 8.27%, while the Indigenous group has the lowest representation at 0.07%. Comparatively, Figure 4.17b reveals a balanced distribution among ethnic groups for both hospitalized and non-hospitalized cases, with Caucasian and Mixed Ethnicities prevailing at 42.39% and 42.41%, respectively, followed by Asian (7.58%) and Black (7.50%) groups.

4.3.4 Symptoms

Figure 4.18 shows a comparison of the prevalence of various symptoms among hospitalized and non-hospitalized COVID-19 patients.

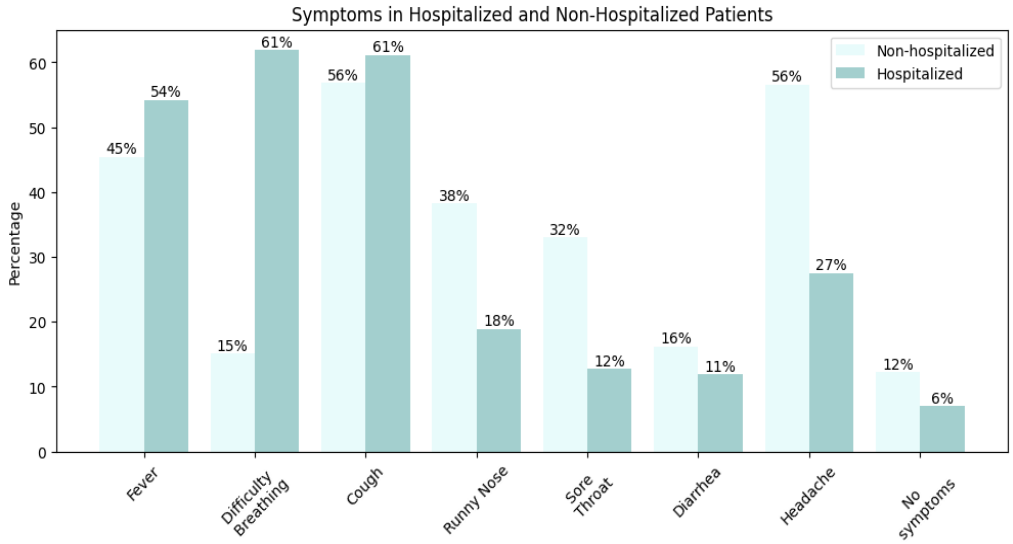


FIGURE 4.18: Symptoms in Hospitalized Patients VS Non-Hospitalized Patients

For non-hospitalized patients, the most frequently reported symptoms are cough and headache, present in 56.86% and 56.58% of patients, respectively. Fever is also common, reported by 45.43% of patients. Other symptoms such as runny nose and sore throat are present in 38.25% and 32.97% of patients, respectively. Diarrhea is less common, reported by 16.23% of patients. Notably, a significant proportion of non-hospitalized patients, 12.29%, reported no symptoms.

In contrast, the distribution of symptoms among hospitalized patients is markedly different. The most common symptom in this group is difficulty breathing, reported by 61.93% of patients. This is followed closely by cough, present in 61.15% of patients, and fever, present in 54.21% of patients. Other symptoms such as runny nose, sore throat, and diarrhea are less common, reported by 18.92%, 12.74%, and 11.86% of patients, respectively. Headache is present in 27.55% of patients. Notably, only 6.96% of hospitalized patients reported no symptoms.

These findings suggest that symptoms such as difficulty breathing, cough, and fever are more common among hospitalized COVID-19 patients compared to those who are not hospitalized. This could indicate that these symptoms are associated with more severe COVID-19 outcomes, leading to hospitalization.

It's also important to note that a significant proportion of patients in both groups reported no symptoms. This underscores the fact that COVID-19 can affect individuals of all health statuses and that even those without symptoms can spread the virus.

4.3.5 Comorbidities

In order to understand the role of underlying health conditions in COVID-19 outcomes, we examined the prevalence of various comorbidities among patients. Figure 4.19 provides a comparative analysis of these comorbidities, offering insights into potential differences in health profiles between these two groups

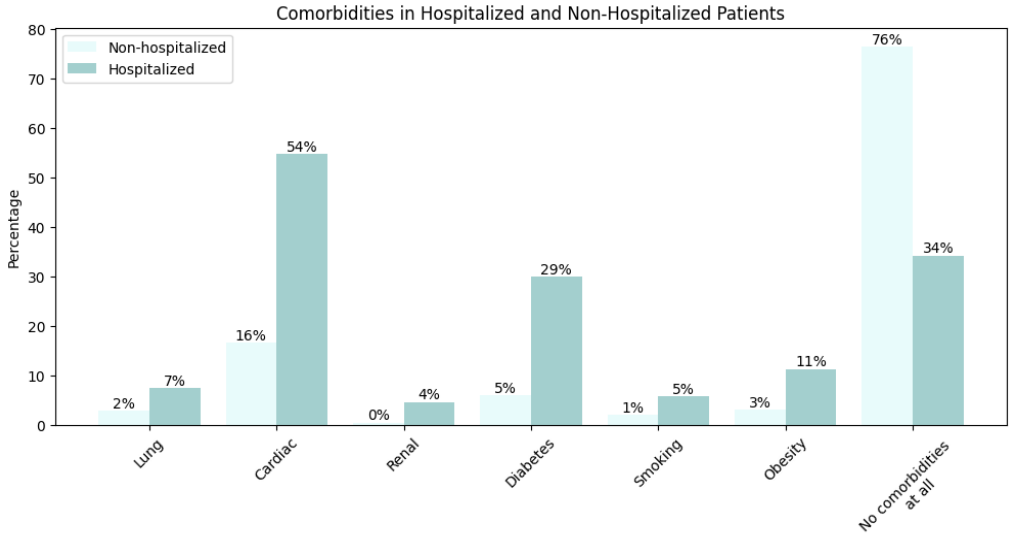


FIGURE 4.19: Comorbidities in Hospitalized Patients VS Non-Hospitalized Patients

For non-hospitalized patients, the most common comorbidity is cardiovascular disease, present in 16.61% of patients. This is followed by diabetes, which is seen in 5.88% of patients. Pulmonary disease and obesity are present in 2.82% and 3.06% of patients, respectively. Renal disease and smoking (tabacism) are less common, with prevalence rates of 0.43% and 1.97%, respectively. Notably, a significant majority of non-hospitalized patients, 76.42%, reported no comorbidities.

In contrast, the distribution of comorbidities among hospitalized patients is markedly different. Over half of these patients, 54.68%, have cardiovascular disease. Diabetes is also significantly more common in this group, present in 29.93% of patients. Pulmonary disease and obesity are seen in 7.43% and 11.24% of patients, respectively. Renal disease and smoking are present in 4.59% and 5.79% of patients, respectively. Notably, only 34.20% of hospitalized patients reported no comorbidities.

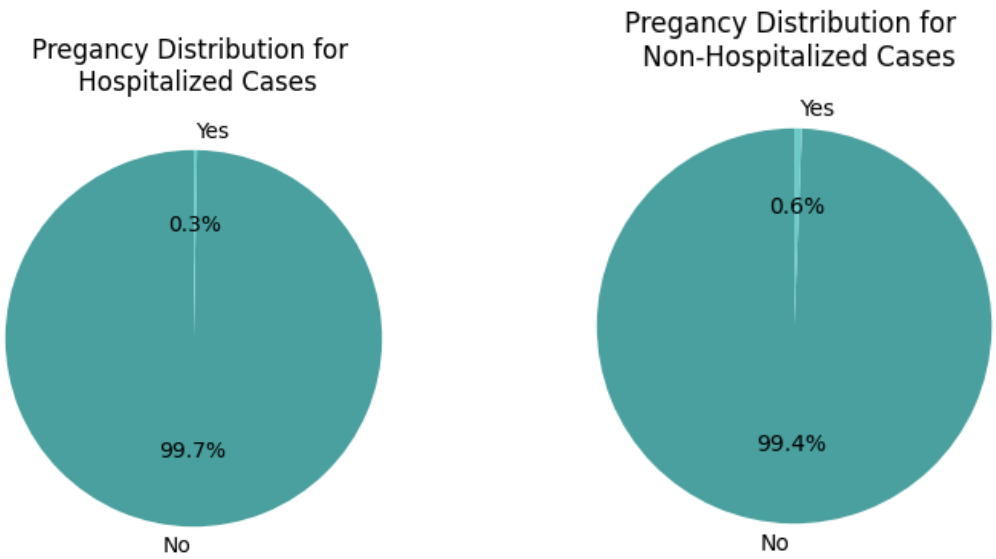
These findings suggest that comorbidities are more common among hospitalized COVID-19 patients compared to those who are not hospitalized. Specifically, cardiovascular disease and diabetes are significantly more prevalent in the hospitalized group, as seen in figure 4.19.

It's also important to note that a significant proportion of patients in both groups reported no comorbidities. This complements the fact that COVID-19 can affect individuals of all health statuses and that even those without underlying health conditions can experience severe outcomes.

4.3.6 Extra Patient Information

In light of the limited representation of pregnancy in our dataset, as indicated in 4.2.9, and our aim to discern features significantly associated with hospitalization, we chose to examine the percentage of pregnant individuals among both hospitalized and non-hospitalized cases. This approach facilitated an investigation into the potential influence of pregnancy on hospitalization rates amid the COVID-19 pandemic.

As illustrated in Figure 4.20, the differences between the two groups in terms of pregnancy are minimal, further underscoring the limited representation of this demographic in our dataset.



(A) Hospitalized Cases (B) Non-Hospitalized Cases

FIGURE 4.20: Pregnancy distribution by hospitalization outcome

Our analysis revealed that pregnant individuals constituted a small proportion of both hospitalized and non-hospitalized cases, at 0.3% and 0.6% respectively. This low representation of pregnancy in the dataset suggests that it may not significantly influence the overall patterns and trends in our analysis.

Interestingly, there appears to be a slightly higher percentage of pregnant individuals among non-hospitalized cases. However, given the minimal overall representation, this difference is unlikely to have a significant impact on the broader trends and patterns in the data.

Given these findings, we decided to exclude the pregnancy column from the dataset. This decision was made to streamline our analysis and focus on features with more substantial representation and potential impact on hospitalization rates.

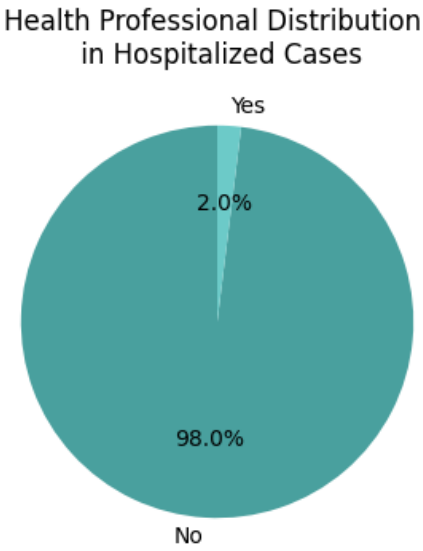
It is important to note, however, that this does not diminish the relevance of studying the impact of pregnancy on COVID-19 outcomes in a dataset with a more substantial representation of this demographic. Future research with a larger sample of pregnant individuals could provide valuable insights into the effects of pregnancy on COVID-19 severity and outcomes.

Before we delve into specific disparities observed in our dataset, it's important to take a closer look at the representation of health professionals within our study's population. As illustrated in Figure 4.21, there are noticeable but modest disparities between the two groups in terms of the presence of health professionals. This observation serves to accentuate the dataset's continued underrepresentation of this specific feature.

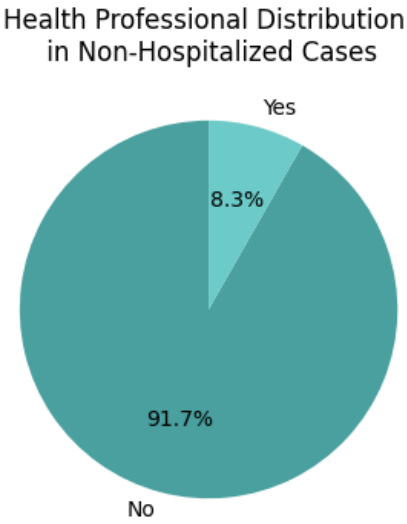
Our analysis reveals a distinct difference in the representation of health professionals among hospitalized and non-hospitalized cases. Specifically, only 2% of hospitalized patients are health professionals, compared to a significantly higher proportion of 8.3% in non-hospitalized cases.

This disparity could suggest that health professionals, despite their increased exposure to the virus, are less likely to require hospitalization when infected with COVID-19. This could be attributed to a variety of factors, including potentially higher rates of vaccination, better access to personal protective equipment, and early access to treatment among health professionals.

Turning the attention to some other two distinct groups in our dataset, it's important to consider the representation of incapacitated individuals. As illustrated in Figure 4.22,



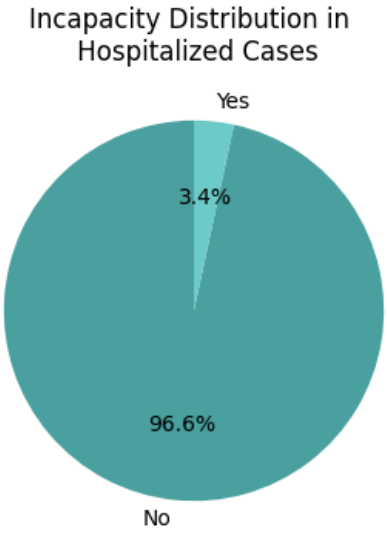
(A) Hospitalized Cases



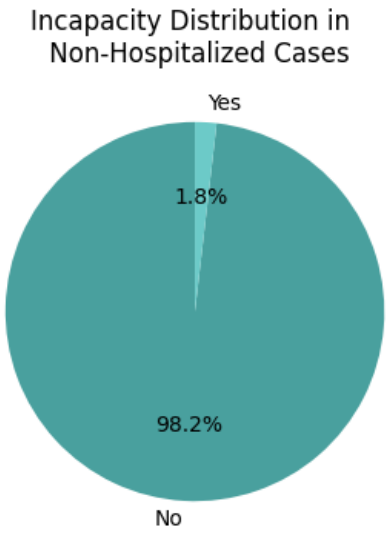
(B) Non-Hospitalized Cases

FIGURE 4.21: Health Professionals distribution by hospitalization outcome

the differences between those two groups are minimal, highlighting the limited representation of this demographic in our dataset.



(A) Hospitalized Cases



(B) Non-Hospitalized Cases

FIGURE 4.22: Incapacity distribution by hospitalization outcome

Our analysis reveals a notable difference in the proportion of incapacitated individuals among hospitalized and non-hospitalized cases. Specifically, incapacitated individuals constitute 3.4% of hospitalized cases, which is almost double the proportion observed in non-hospitalized cases at 1.8%.

This disparity suggests that incapacitated individuals are more likely to require hospitalization when infected with COVID-19. Incapacitated individuals may have underlying health conditions or compromised immune systems that make them more susceptible to severe outcomes from the virus. Additionally, they may face challenges in accessing timely and appropriate care, which could exacerbate the severity of their illness.

4.4 Clustering Analysis

The process of clustering analysis, which groups similar data points based on their attributes, has been employed in this study to discern patterns in patient characteristics, such as symptoms and comorbidities.

The K-means algorithm [159] was chosen for this task due to its computational efficiency and its capacity to manage large datasets with high dimensionality. K-means is a clustering technique in data analysis used to group similar data points together into clusters. It operates by iteratively assigning data points to the nearest cluster center and then recalculating the cluster centers based on the newly assigned data points. This process continues until the cluster centers stabilize. K-means requires the pre-specification of the number of clusters (k) and aims to minimize the distance between data points and the cluster centers they belong to while maximizing the distance between different clusters. To ascertain the optimal value of k, the number of clusters, the elbow method [160] was employed, utilizing Within-Cluster-Sum-Squared (WCSS) [161] errors as the criterion. The Jaccard index [162] was chosen as the similarity measure for the application of K-means to our dataset.

The decision to segment the dataset into symptoms and comorbidities was driven by the desire to gain distinct insights from these two different aspects of a patient's health. Symptoms, as manifestations of the disease (in this case, COVID-19), and comorbidities, as pre-existing health conditions, can provide unique perspectives when analyzed separately. Clustering based on symptoms can reveal patterns related to the disease's manifestation and progression, thereby potentially contributing to a better understanding of the disease and aiding in the formulation of targeted treatment plans.

On the other hand, clustering based on comorbidities can shed light on how pre-existing conditions might influence the disease's severity and patient outcomes. This

could provide valuable insights for developing preventative measures and risk stratification strategies, particularly given that patients with comorbidities are often at higher risk for severe outcomes from COVID-19.

The decision to exclude patients without any comorbidities from the comorbidity analysis was made to address the issue of data sparsity. With a significant proportion of the patients in the dataset (74.82%) not having any comorbidities, including these patients in the comorbidity-based clustering could lead to many empty or zero values in the data matrix. This could negatively impact the clustering results, potentially leading to less reliable and less meaningful insights.

4.4.1 Symptoms

The clustering analysis performed on the symptom sub-dataset resulted in the identification of nine unique clusters. To help better understand the patient distribution across these clusters, a corresponding visualization was constructed, as illustrated in Fig. 4.23.

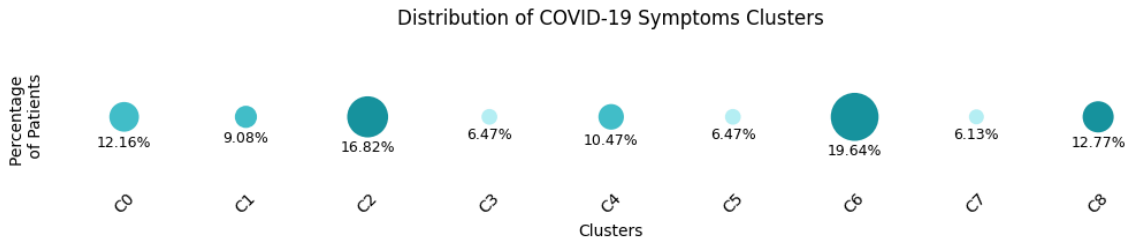


FIGURE 4.23: Percentage of patients per cluster of symptoms.

The three clusters that contain the greatest number of patients are indicated by the darkest shade of blue, specifically clusters C6, C2, and C8. In contrast, the three clusters with the smallest number of patients are denoted by the lightest shade of blue, namely clusters C3, C5, and C7. This graphical representation serves to elucidate the distribution of patients across the symptom clusters, thereby offering valuable insights into the prevalence of symptom patterns among the patient population.

Table 4.1 provides an overview of the symptom profiles associated with each cluster. The first and fourth columns identify the clusters under consideration, representing the top 3 and bottom 3 clusters, respectively. The second and fifth columns outline the symptom(s) essential for a patient's categorization within the respective cluster, while the third and sixth columns specify the symptom(s) that should not be present for the patient to be associated with that specific cluster.

TABLE 4.1: Top 3 and Bottom 3 symptoms clusters characterization

Top 3	Present	Absent	Bottom 3	Present	Absent
C6	headaches	runny nose, sore throat	C3	runny nose, sore throat	fever
C2	-	runny nose, sore throat	C5	cough	runny nose, fever, headaches
C8	runny nose	sore throat	C7	runny nose difficulty breathing	-

Given that we are discussing COVID-19, the distribution of patients across the clusters and the associated symptom profiles could be influenced by several factors related to the nature of this specific disease.

For the top three clusters (C6, C2, and C8), the common absence of symptoms such as a runny nose and sore throat might suggest that these symptoms are less prevalent or less specific to COVID-19. The presence of headaches in cluster C6 could indicate a common symptom among a significant subset of COVID-19 patients. The absence of any specific symptoms in cluster C2 might suggest a group of asymptomatic or mildly symptomatic COVID-19 patients. The presence of a runny nose in cluster C8 could indicate a subset of COVID-19 patients who experience this symptom without the accompanying sore throat, which could be related to individual variations in disease presentation.

Conversely, the bottom three clusters (C3, C5, and C7) highlight symptom profiles that are less commonly observed among COVID-19 patients. The coexistence of runny nose and sore throat in cluster C3 possibly characterizes a subset of patients with these concurrent symptoms. Within cluster C5, a unique presence of cough without accompanying symptoms like runny nose, fever, or headaches indicates a specific manifestation. Furthermore, the simultaneous occurrence of runny nose and breathing difficulties in cluster C7 might indicate a more severe symptom pattern associated with critical COVID-19 cases.

This analysis revealed intriguing patterns regarding the symptomatology of COVID-19. Cough, a symptom often mentioned by patients, emerged as a distinctive trait solely within cluster C5 (alongside the absence of runny nose, fever, and headaches). Notably, this cluster is among the three groups with the smallest number of patients. This observation suggests that despite the high prevalence of cough, it does not appear to be a distinguishing characteristic of the clusters that encompass the majority of patients in our dataset.

There are various factors that could contribute to this. Cough, although commonly associated with the illness, is not exclusive to COVID-19 and can be indicative of various

respiratory conditions. As a result, the presence of a cough alone might not be adequate to definitively assign a patient to a particular COVID-19 symptom cluster. Additionally, individual patient attributes like age, overall health condition, and the existence of underlying health issues are recognized as influencers of COVID-19 severity and progression. This implies that the observed symptom clusters may better mirror these fundamental factors rather than being solely dictated by the presence of an isolated symptom like a cough.

In conclusion, the analysis of symptom clusters among COVID-19 patients has unveiled intriguing insights. While some clusters align with common expectations, such as the presence of headaches in one group or the absence of specific symptoms in another, there are notable deviations that challenge conventional assumptions. For instance, the presence of cough, a symptom frequently associated with COVID-19, is not a defining feature of the larger patient clusters, suggesting its presence may not be as distinctive in the context of the disease. These findings emphasize the varied nature of COVID-19 symptoms, which can overlap with various other respiratory conditions. Additionally, individual patient characteristics and underlying health factors play a crucial role in shaping the manifestation and severity of the disease.

4.4.2 Comorbidities

The clustering analysis conducted on the comorbidities sub-dataset resulted in the identification of nine unique clusters, as depicted in Figure 4.24.

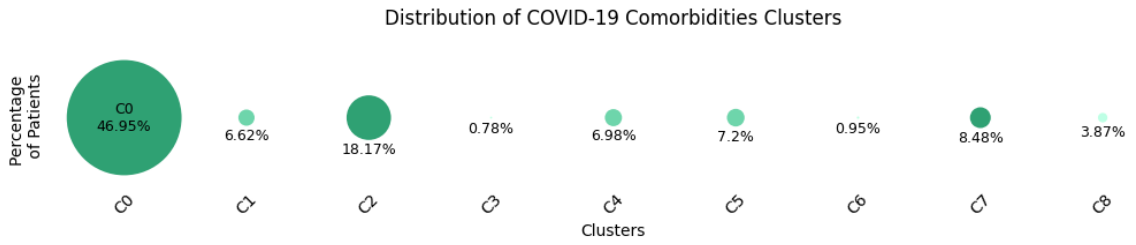


FIGURE 4.24: Percentage of patients within each comorbidity cluster

Clusters containing the largest patient populations are displayed in the deepest shade of green (namely, clusters C0, C2, and C7), while those with the lowest patient numbers are portrayed in the palest shade of green (clusters C8, C6, and C3).

Table 4.2 offers an overview of the comorbidity profiles associated with each cluster. The first and fourth columns identify the clusters under consideration, representing the

top 3 and bottom 3 clusters, respectively. The second and fifth columns outline the comorbidity(s) essential for a patient's categorization within the respective cluster, while the third and sixth columns specify the comorbidity(s) that should not be present for the patient to be associated with that specific cluster.

TABLE 4.2: Top 3 and Bottom 3 comorbidities clusters characterization

Top 3	**Present**	**Absent**	**Bottom 3**	**Present**	**Absent**
C0	diabetes	cardiovascular, lung, obesity	C8	lung	diabetes
C2	cardiovascular	renal, obesity	C6	renal, diabetes	cardiovascular, lung
C7	obesity	renal	C3	lung, obesity	-

Similarly to the preceding analysis of symptoms, this examination unveiled distinct patterns in the comorbidity profiles of COVID-19 patients. These clusters exhibit specific combinations of comorbidity presence and absence. The largest cluster, C0, is characterized by the presence of diabetes and the absence of cardiovascular, lung, and obesity issues. In other words, all individuals in cluster C0 have diabetes, while none of them have any issues related to cardiovascular health, lung function, or obesity. In cluster C2, the common factor is cardiovascular problems, with no occurrences of renal or obesity issues. Lastly, cluster C7 is marked by the presence of obesity among all its members, while no one in this cluster has renal problems.

On the other hand, the three clusters at the bottom (C8, C6, and C3) exhibit less frequent comorbidity patterns among COVID-19 patients. Cluster C8 is characterized by the presence of lung disease alongside the absence of diabetes. For cluster C6, all patients have diabetes and renal problems, yet none of them exhibit any cardiovascular or lung issues. In the case of cluster C3, the least populous cluster, its defining feature is the coexistence of lung and obesity problems.

The coexistence of diabetes while simultaneously lacking cardiovascular, lung, and obesity issues in cluster C0 raises questions about the relationship between diabetes and other comorbidities in COVID-19 patients. This observation implies that even though diabetes is widely recognized as a risk factor for severe outcomes, the lack of other prevalent comorbidities may point toward distinct pathways or factors that influence the severity of COVID-19 in individuals with diabetes [163, 164]. Moreover, the absence of cardiovascular disease within the same cluster, which encompasses a significant portion of our dataset, is noteworthy. This observation highlights that despite the widespread occurrence of cardiovascular disease among patients, it does not seem to be a defining feature

of the largest cluster in our dataset. This logic aligns with the notion that if a particular ailment is prevalent among nearly all or most individuals, it loses its discriminative value. On the contrary, its absence becomes more telling and indicative.

On the other hand, the coexistence of cardiovascular problems within cluster C2, coupled with the absence of renal and obesity issues, invites a deeper investigation into the complex interplay between cardiovascular health and other comorbidities within the realm of COVID-19. This discovery is rather unexpected, given the high prevalence of cardiovascular problems among COVID-19 patients. Typically, cardiovascular problems tend to be paired with other comorbidities like chronic kidney disease and other illnesses [165, 166]. However, in this specific case, where cluster C2 ranks as the second-largest cluster, its distinctive feature is the absence of renal and kidney problems while concurrently presenting cardiovascular problems.

The presence of obesity while lacking renal problems in cluster C7 raises questions about the direct association between obesity and renal complications in COVID-19 patients. The existing literature has consistently highlighted a relationship between obesity and kidney disease [167] within the context of COVID-19, which stands in contrast to the outcome revealed by the cluster analysis. Paradoxically, prevailing research underscores that obesity significantly heightens the risk of kidney disease development [168]. This disparity between the observed cluster outcome and established knowledge in the field accentuates the complexity of comorbidity patterns in COVID-19 patients.

Cluster C8's characteristic combination of lung issues without concurrent diabetes draws attention to a potential differentiation between lung health and diabetes within the context of COVID-19. Although both conditions can contribute to disease severity, this cluster implies that their coexistence is not ubiquitous. Considering the higher prevalence of diabetes among individuals with chronic obstructive pulmonary disease (COPD) compared to the general population [169], this result is particularly intriguing. This discovery challenges the conventional understanding that patients with lung disease typically have diabetes, as suggested by the literature. Nonetheless, it's crucial to bear in mind that this particular cluster constitutes a relatively smaller subset of cases.

In cluster C6, where both renal problems and diabetes are requisites while cardiovascular and lung issues are excluded, the evidence suggests a convergence between renal health and diabetes concerning COVID-19, as supported by the literature [170]. This implies that renal and metabolic factors may jointly contribute to the disease's severity in

individuals with diabetes, warranting further investigation into their combined impact. Similarly, in cluster C3, the co-occurrence of lung and obesity issues underscores a potential connection between respiratory and metabolic well-being among COVID-19 patients, which is also substantiated by existing literature [171]. This cluster potentially represents a subgroup with distinct vulnerability resulting from the interplay between respiratory and metabolic factors. It's worth noting that both clusters, C6 and C3, encompass less than 1% of the patient population, indicating their relatively limited prevalence.

In conclusion, the analysis of comorbidity patterns among COVID-19 patients has revealed complex relationships and unexpected insights. While some clusters agree with established understandings, such as the association between diabetes and renal health, others challenge conventional knowledge, like the disassociation between cardiovascular problems and certain comorbidities within a significant portion of the dataset. These findings highlight the complexity of comorbidity patterns in COVID-19 patients and the need for nuanced approaches to understanding the interplay between different health conditions. This study sheds light on potential subgroups within the COVID-19 patient population, each with unique vulnerabilities emerging from the complex interactions between comorbidities. Further research in this direction could offer insights into customized interventions and treatments for specific patient profiles, ultimately improving patient care and outcomes.

Chapter 5

Hospitalization Prediction Task

The main goal of this chapter is to predict whether patient hospitalization is necessary. The prediction relies solely on the TriCovB dataset that now comprises 19 patient features, including one numerical variable and 18 binary variables. The binary target variable indicates hospitalization necessity, with *0* representing *No* and *1* indicating *Yes*. The dataset, encompassing around 190,000 entries, presents a challenge due to its highly imbalanced nature. The target variable distribution is skewed, with *No* instances making up about 96.1%, while *Yes* instances constitute only 3.9%. To address this, the dataset is split into training and test sets for model development and evaluation, considering the need to prioritize sensitivity (identifying *Yes* cases) while maintaining specificity to prevent unnecessary hospitalizations. This chapter will present all the tasks executed on the training dataset for constructing our ML model. Each task is associated with a specific section, including Feature Selection, Addressing Imbalance, Preliminary Model Selection, Performance Evaluation, and Discussion and Conclusion.

5.1 Feature Selection

In light of the foundational concepts discussed in Chapter 2, let's provide a concise summary. Feature Selection is an important step in ML. It involves choosing the most influential features from the larger pool of variables. The primary goal is to streamline the dataset, retaining only those attributes that truly impact the model's performance. This process offers several advantages, including reducing noise in the data, enhancing model efficiency, and improving predictive accuracy. One common method employed during feature selection is Feature Importance analysis. This technique evaluates each feature's

contribution to an ML model's predictions. Essentially, it ranks features based on their influence. Features with higher importance scores are considered more vital in determining the model's outcomes. This approach helps researchers and data scientists pinpoint the key factors driving their models' performance. Feature importance with tree-based models is the chosen approach to perform Feature Selection.

Feature importance allows us to identify the most influential features that contribute to the prediction outcome. We can extract feature importance scores by training tree-based models such as Random Forest and Extreme Gradient Boosting, and these scores indicate the relative importance of each feature in making accurate predictions. Features with higher importance scores are considered more relevant and informative for predicting hospitalization. Then the focus can be laid on a subset of highly important features that contribute the most to the model's predictive performance.

To guide the decision on which features from the training dataset to retain and which to discard, a feature importance analysis was conducted using tree-based models. Specifically, algorithms such as RF, GB, XGBoost, DT, AdaBoost, and Bagging were employed - their definitions can be revised in Chapter 2. The results of this analysis are summarized in Figure 5.1, providing insights into the relative importance of the various features in the predictive models. This figure shows the top ten features that have collected the highest importance scores across the models applied. Features are listed alphabetically, and the importance scores are rounded to four decimal places. A dash ("-") signifies that a particular feature did not rank among the top 10 most influential features for a given model.

The features that were considered important across all models are highlighted in bold in Figure 5.1, and these features are Age, Diabetes, Difficulty Breathing, Headache, Runny Nose, and Sore Throat. The prominence of these features across all models suggests their pivotal role in the decision-making process of these models, indicating their potential as robust predictors for the outcome of COVID-19.

The Age feature persists across all models, corroborating the clinical understanding of COVID-19. It is well-documented in the literature that older individuals are at a heightened risk of severe disease and mortality due to COVID-19, which could elucidate why Age is a highly rated feature in all models [172–174]. These findings align with the results highlighted in Chapter 4, where a clear differentiation in the age distribution between hospitalized and non-hospitalized patients was noted.

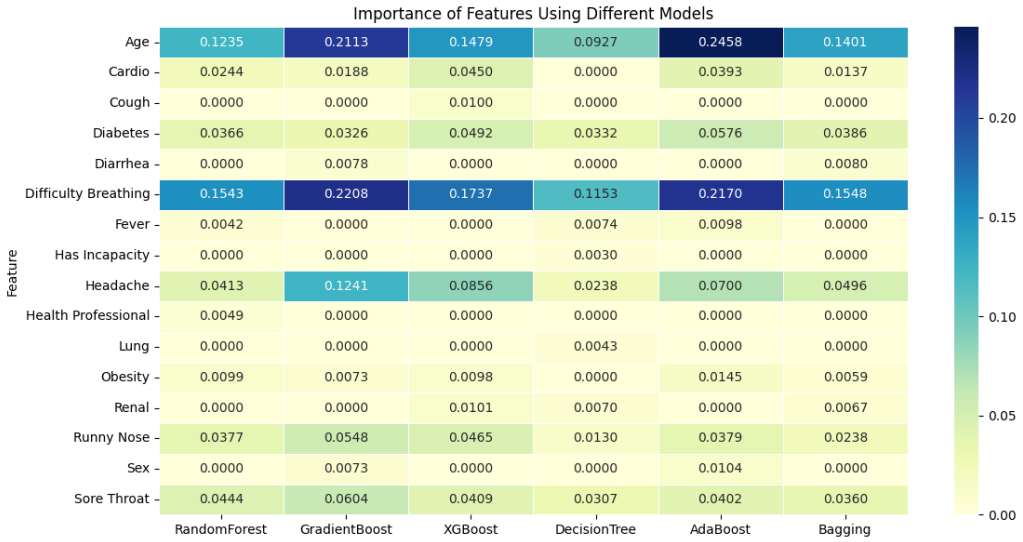


FIGURE 5.1: Heatmap of Feature Importance

The feature "Difficulty Breathing" consistently emerges as a predictive factor across all models. This consistency is further supported by the contrast observed in Chapter 4, where 61% of hospitalized patients exhibited this symptom, compared to only 15% of non-hospitalized patients. Notably, "Difficulty Breathing" is a recognized indicator of severe COVID-19, often heralding the onset of conditions such as pneumonia or acute respiratory distress syndrome (ARDS) [175–177]. Hence, it is not surprising that this feature emerges as a key determinant in the models' predictions.

Another feature that consistently emerges as significant across all models is Diabetes, a well-documented comorbidity known to exacerbate COVID-19 symptoms and contribute to more severe outcomes [164–166, 173]. It's worth noting that, while the disparities may not be as prominent as those observed for Difficulty Breathing, a considerable distinction persists. Specifically, as detailed in Chapter 4, 29% of hospitalized patients exhibited diabetes as a comorbidity, whereas only 5% of non-hospitalized patients had this characteristic. This observation reinforces its role as a predictive factor for hospitalization.

In essence, all the features that appear across all models also demonstrated a noteworthy contrast in Chapter 4 when comparing hospitalized and non-hospitalized patients. This occurrence serves to corroborate and validate the results obtained for these features in the current stage of analysis.

Let's delve into the attributes of "Headache," "Runny Nose," and "Sore Throat," which are frequently associated with various respiratory illnesses, including COVID-19 [178].

It's noteworthy that these symptoms have been identified as significant features in multiple models. This finding implies that, despite their non-specific nature, these symptoms could potentially serve as indicators of a COVID-19 infection.

In contrast, features such as Fever, Lung, and Renal are not deemed as important by any of the models which aligns with the results obtained in the Cluster Analysis in chapter 4, section 4.4. This could be interpreted as these features not contributing significantly to the models' predictive accuracy. Interestingly, Fever, a common symptom of COVID-19 [178], is not considered important. This could be attributed to the fact that fever is a common symptom for many illnesses and not specific to COVID-19, thereby reducing its predictive power.

The Cardio feature, which pertains to cardiovascular comorbidities, is considered significant by some models but not by others. This could be due to the inherent differences in the construction of decision trees across these models or due to the specific characteristics of the training data. The importance associated with this feature might be justified by the results outlined in Chapter 4, where it was noted that 54% of hospitalized patients had cardiovascular problems, while only 16% of non-hospitalized patients exhibited this comorbidity. Moreover, a noteworthy aspect to consider as well, as highlighted in Section 4.4 of the same chapter, is that the predominant cluster within our analysis demonstrated an absence of cardiac problems, while the second-largest cluster exclusively consisted of patients with cardiac issues. These findings suggest that the significance of the Cardio feature is context-dependent. While it may not hold universal importance across all models, its role becomes more pronounced when considering specific patient clusters or subsets.

For last, the Sex feature is only considered important by the GradientBoost and AdaBoost models. This could be reflecting the observed disparity in COVID-19 outcomes between genders, with men often experiencing more severe disease and higher mortality rates [95].

These findings have informed the creation of a refined dataset, solely comprising the features outlined in Figure 5.1. In the subsequent section, we will apply a combination of resampling techniques to this dataset, encompassing both oversampling and undersampling methods, to further refine our models.

5.2 Tackling Imbalance

The TriCovB dataset, as introduced before, has 96.1% of entries representing patients who were not hospitalized and only 3.9% representing hospitalized patients, which indicates a highly imbalanced data scenario. Given the imbalanced nature of the dataset, the positive class has been explicitly defined as being the value '1'. This explicit definition ensures that our models are more adept at detecting the minority class, which is often of greater significance in imbalanced datasets. This particular scenario of imbalanced data reflects the real-world nature of the problem[179].

In healthcare, hospitalizations are relatively less frequent compared to non-hospitalizations [180]. Accurately predicting the minority class (hospitalized patients) carries pivotal importance within this context. As detailed in chapter 4, particularly in section 4.3, it was observed that a considerable proportion of individuals who were admitted to the hospital ultimately succumbed to the condition. Identifying patients who require hospitalization is important for timely medical intervention, appropriate allocation of resources, and adequate healthcare management [181].

To tackle imbalanced data challenges, approaches like undersampling, oversampling, or a combination of both can be employed. Undersampling reduces instances in the majority class to balance it with the minority class, curbing bias and boosting overall model performance. Oversampling involves increasing minority class instances through synthetic data or duplications, achieving class distribution equilibrium, and potentially enhancing model performance [182].

Random Undersampling (RUS) is a method of undersampling [183]. It functions by randomly discarding instances from the majority class to achieve a balanced class distribution. This method proves particularly useful when dealing with large datasets or when data collection costs are high, as it minimizes the amount of data required for model training. However, it's important to apply RUS with caution, as it may result in the loss of significant information from the majority class, which could be vital in specific scenarios such as predicting rare diseases [184].

We have applied RUS to the original dataset to obtain a newer version of a dataset with a class distribution that is closer to a 50:50 balance between the minority class (the class of interest) and the majority class. Figure 5.2 illustrates the effect of applying the RUS technique on the original dataset. The plot shows a more balanced distribution of the two classes, achieved by reducing the instances of the majority class. However, it's important

to note that this method may result in the loss of potentially important information from the majority class.

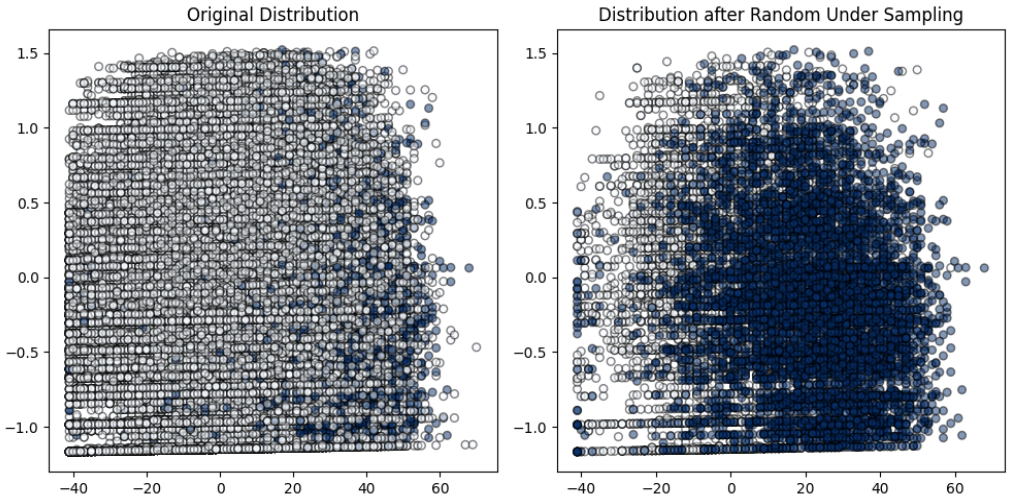


FIGURE 5.2: Random Undersampling Process

On the other hand, the Synthetic Minority Over-sampling Technique (SMOTE) is an oversampling method. SMOTE works by creating synthetic samples for the minority class through interpolation between neighboring instances. This method is especially beneficial when accurate prediction of the minority class is critical [185], such as in predicting hospital admissions.

As depicted in Figure 5.3, the SMOTE technique generates synthetic examples for the minority class, resulting in a more balanced dataset. The plot shows a denser cluster of minority class instances, indicating the creation of synthetic data points. However, this technique may lead to overfitting due to the synthetic nature of the new instances.

Moreover, combining oversampling and undersampling techniques, like in the case of SMOTEENN (SMOTE + Edited Nearest Neighbors), can boost performance by eliminating noisy samples and generating new instances. These resampling techniques offer an effective solution for enhancing the model's ability to understand the patterns and characteristics of the minority class, thereby improving predictive performance and mitigating the bias towards the majority class [186].

Figure 5.4 demonstrates the result of applying SMOTEENN to the dataset. This technique first over-samples the minority class using SMOTE method and then cleans the data using Edited Nearest Neighbors (ENN). The plot shows a more diverse and cleaner distribution of instances, combining the benefits of both over-sampling and under-sampling.

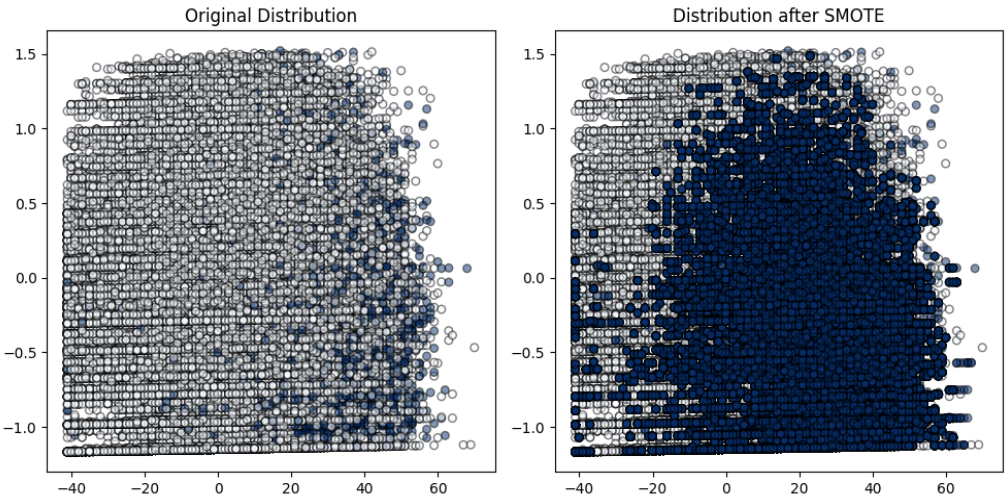


FIGURE 5.3: SMOTE Process

However, this method can be computationally expensive due to the complexity of the ENN algorithm.

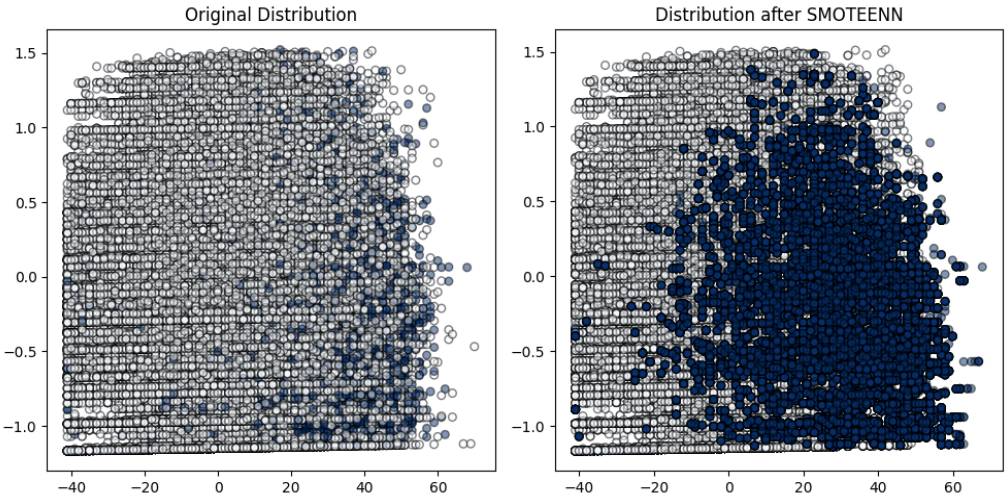


FIGURE 5.4: SMOTEENN Process

A comparison of these techniques, including an overview of their primary operational mechanisms, a discussion of their key advantages, and an examination of potential drawbacks, is presented in Table 5.1.

With all of this considered, the decision was made to create four versions of the TriCovB dataset, using each of the mentioned approaches: the **original** (imbalanced) dataset, the **RUS** dataset, the **SMOTE** dataset, and the **SMOTEENN** dataset. These dataset variations enable an examination of strategies to address the impact of imbalanced data on analyses.

TABLE 5.1: Comparison of Sampling Techniques

Technique	Description	Advantages	Disadvantages
RUS	Reduces majority class instances.	Simple, fast, improves computational efficiency.	Risk of losing information, potential underfitting.
SMOTE	Generates synthetic minority class instances.	Provides diverse training data.	Potential overfitting.
SMOTEENN	Applies SMOTE, then cleans data with ENN.	Results in cleaner, diverse dataset. Combines over-sampling and under-sampling.	Can be computationally expensive.

5.3 Hyperparameter Tuning

The initial stage of our model selection process involves conducting a grid search across different classifiers. Grid search is a technique used in ML to systematically explore and evaluate various combinations of hyperparameters for a given model, then train and evaluate the model using all the combinations set for the grid [187]. This helps to identify the best combination of hyperparameters that leads to the best performance of the model on a validation or test dataset. Hyperparameters, unlike other parameters, are not derived from the data during the learning process but are predetermined before the initiation of this process [188]. Essentially, it consolidates the search for the ideal hyperparameters for an ML algorithm, offering a time-efficient and less labor-intensive alternative to manual tuning [189].

To build robust ML models for predictive analysis, we have selected a diverse set of classifiers, each offering a unique approach to solving the given task. To get a better understanding of the inner workings of these classifiers and their associated hyperparameters, we present a concise overview.

- Decision Tree
  - **Maximum depth of the tree ('max_depth'):** This hyperparameter controls the maximum depth or levels of the decision tree. It determines how deep the tree can grow during training. A deeper tree can capture more complex patterns but is more prone to overfitting.
- K-Nearest Neighbors
  - **Number of neighbors ('n_neighbors'):** This hyperparameter defines the number of nearest data points (neighbors) to consider when making predictions for

a new data point. It influences the model's sensitivity to local patterns in the data.

- Logistic Regression
  - **Inverse of regularization strength ('C'):** This hyperparameter represents the inverse of the regularization strength. Smaller values of 'C' indicate stronger regularization, which can prevent overfitting by penalizing large coefficients in the logistic regression model.
- AdaBoost
  - **Maximum number of estimators ('n_estimators'):** This hyperparameter specifies the maximum number of weak learners (usually decision trees) that AdaBoost can use. Increasing this value can improve model performance but might lead to longer training times.
  - **Learning rate:** The learning rate controls the contribution of each weak learner to the final prediction. It is a small positive value (e.g., 0.1, 0.01) that scales the weight of each weak learner's prediction.
- Bagging
  - **Number of base estimators in the ensemble ('n_estimators'):** Bagging (Bootstrap Aggregating) creates an ensemble of base models. This hyperparameter determines how many base models are included in the ensemble.
- Gradient Boosting
  - **Number of boosting stages ('n_estimators'):** Gradient Boosting builds an ensemble of decision trees sequentially. This hyperparameter specifies the number of boosting stages or trees to be added to the ensemble.
  - **Learning rate:** Similar to AdaBoost, the learning rate controls the contribution of each tree added to the ensemble. It's a small positive value.
- Random Forest
  - **Maximum depth of the tree ('max_depth'):** Similar to the decision tree's 'max_depth,' this hyperparameter controls the maximum depth of individual trees in the random forest.

  - **Number of trees in the forest ('n_estimators'):** This hyperparameter sets the number of decision trees to include in the random forest ensemble.

- XGBoost

  - **Number of boosting stages ('n_estimators'):** XGBoost builds an ensemble of decision trees. This hyperparameter specifies the number of boosting stages or trees to be added.
  - **Learning rate:** As with other boosting methods, the learning rate governs the contribution of each tree to the ensemble's prediction.

We resort to the grid-search method for hyper-parameter tuning with 10-fold cross-validation process. Cross-validation is a resampling procedure utilized to evaluate ML models on a limited data sample [190]. The method of 10-fold cross-validation partitions the original sample into 10 subsamples, using nine for training purposes, and the remaining one for validation. This process is iteratively repeated 10 times, with each iteration using a different subsample as the validation set [191].

In the following part, we explore the grid search approaches, outlining the hyperparameter values explored for each model. Each ML model is associated with its distinct set of hyperparameters. Table 5.2 specifies the ranges of values examined for each hyperparameter.

5.3.1 Original Dataset

The grid search process was first conducted on the original dataset. This dataset, in its original form, presents the true distribution of the classes and serves as a baseline for comparison with the resampled datasets. The hyperparameters of various ML models were tuned on this original dataset, and the results are presented in Table 5.3.

For this preliminary stage, we have selected G-means as our evaluation metric for performing a grid search, as G-means is particularly suitable for imbalanced datasets as it provides a balanced measure of the model's performance on both the majority and minority classes 5.2. G-means balances sensitivity and specificity. This balance is crucial for our problem where both false positives and false negatives have significant implications. The aim is to minimize both types of errors by utilizing G-means, ensuring that neither class is disproportionately favored over the other during the hyperparameter tuning process.

TABLE 5.2: Search Space of Hyperparameters

Algorithm	**Hyperparameters**	**Tested Values**
Decision Tree	Maximum depth of the tree 'max_depth'	None, 5, 10
K-Nearest Neighbors	Number of neighbors 'n_neighbors'	3, 5, 7
Logistic Regression	Inverse of regularization strength 'C'	0.001, 0.01, 0.1, 1, 10, 100
AdaBoost	Maximum number of estimators 'n_estimators'	50, 100
	Learning rate	0.1, 0.01
Bagging	Number of base estimators in the ensemble 'n_estimators'	10, 20
Gradient Boosting	Number of boosting stages 'n_estimators'	100, 200
	Learning rate	0.1, 0.01
Random Forest	Maximum depth of the tree 'max_depth'	None, 5, 10
	Number of trees in the forest 'n_estimators'	100, 200
XGBoost	Number of boosting stages 'n_estimators'	100, 200
	Learning rate	0.1, 0.01

TABLE 5.3: Grid Search Results for the Original Dataset

	Decision Tree	**K-Nearest Neighbors**	**Logistic Regression**	**AdaBoost**	**Bagging**	**Gradient Boosting**	**Random Forest**	**XGBoost**
Best Hyperparameters	max depth = None	num neighbors = 3	C = 1	learning rate = 0.1, num estimators = 100	num estimators = 20	learning rate = 0.1, num estimators = 100	max depth = None, num estimators = 200	learning rate = 0.1, num estimators = 100
G-Mean Score	0.4607	0.4458	0.4379	0.3048	0.4743	0.4696	0.4678	0.4459

The grid search results on the original dataset indicate that the models are struggling to balance sensitivity and specificity, as evidenced by G-Mean scores below 0.5. This challenge may stem from the dataset's imbalance, with far fewer instances of patients requiring hospitalization compared to those who don't. The Bagging model with 20 estimators and the RF model with no maximum depth and 200 estimators achieved the highest G-Mean scores, suggesting they handle the class imbalance best. However, these scores are still insufficient for such a critical task. These findings emphasize the necessity of using resampling techniques like RUS, SMOTE, and SMOTEENN to address the dataset's class imbalance. These techniques could potentially enhance the models' performance.

In summary, while the initial results serve as a valuable benchmark, they also highlight the issues caused by the dataset's class imbalance. The next step involves applying

the resampling techniques presented before and repeating the grid search on the resampled datasets to find models that can more accurately and reliably predict COVID-19 patients' hospitalization needs.

5.3.2 RUS Dataset

After the grid search was conducted on the original dataset, the same procedure was applied to the RUS dataset. This adapted dataset presents a more balanced class representation to counteract the initial class imbalance. The hyperparameters of several ML models were fine-tuned on this RUS dataset. The outcomes, encompassing the optimal hyperparameters for each model along with their corresponding G-Mean scores, are detailed in Table 5.4.

TABLE 5.4: Grid Search Results for the RUS Dataset

	Decision Tree	**K-Nearest Neighbors**	**Logistic Regression**	**AdaBoost**	**Bagging**	**Gradient Boosting**	**Random Forest**	**XGBoost**
Best Hyperparameters	max depth = 5	num neighbors = 7	C = 0.1	learning rate = 0.1, num estimators = 100	num estimators = 20	learning rate = 0.1, num estimators = 100	max depth = 10, num estimators = 200	learning rate = 0.1, num estimators = 100
G-Mean Score	0.8109	0.7859	0.8256	0.8234	0.7827	0.8331	0.8226	0.8268

These scores are notably superior to those from the original dataset, suggesting improved model performance on a balanced dataset created via RUS. GB, with a learning rate of 0.1 and 100 estimators, tops the list with the highest G-Mean score. XGBoost and AdaBoost follow closely, demonstrating effective handling of the balanced dataset and a balanced sensitivity-specificity trade-off. These findings highlight the importance of RUS in enhancing ML model performance on imbalanced datasets.

5.3.3 SMOTE Dataset

Subsequent to the grid search conducted on the original and RUS datasets, the same methodology was employed on the SMOTE dataset. This version has been synthetically augmented to address the class imbalance, thereby providing a more equitable class representation. The results of the grid search, including the optimal hyperparameters for each model and their corresponding G-Mean scores, are displayed in Table 5.5.

The G-Mean scores, as shown in Table 5.5, are markedly superior to those from the original dataset, suggesting enhanced model performance on a balanced dataset created

TABLE 5.5: Grid Search Results for the SMOTE Dataset

	Decision Tree	K-Nearest Neighbors	Logistic Regression	AdaBoost	Bagging	Gradient Boosting	Random Forest	XGBoost
Best Hyperparameters	max depth = None	num neighbors = 7	C = 10	learning rate = 0.1, num estimators = 100	num estimators = 20	learning rate = 0.1, num estimators = 200	max depth = None, num estimators = 100	learning rate = 0.1, num estimators = 200
G-Mean Score	0.9011	0.7961	0.8332	0.8248	0.9030	0.8411	0.9029	0.8560

via SMOTE. Bagging and RF models stand out with the highest G-Mean scores, indicating their effectiveness in managing the balanced dataset and maintaining a good balance between sensitivity and specificity. The obtained results show the value of SMOTE in boosting the performance of ML models on imbalanced datasets.

5.3.4 SMOTEENN Dataset

After performing the grid search on the original, RUS, and SMOTE datasets, we extended the same procedure to the dataset modified using the SMOTEENN. This adjusted dataset has been synthetically expanded and refined to rectify the class imbalance, thus providing a more balanced class representation. The results, including the optimal hyperparameters for each model and their respective G-Mean scores, are presented in Table 5.6.

TABLE 5.6: Grid Search Results for the SMOTEENN Dataset

	Decision Tree	K-Nearest Neighbors	Logistic Regression	AdaBoost	Bagging	Gradient Boosting	Random Forest	XGBoost
Best Hyperparameters	max depth = None	num neighbors = 3	C = 0.01	learning rate = 0.1, num estimators = 100	num estimators = 20	learning rate = 0.1, num estimators = 200	max depth = None, num estimators = 200	learning rate = 0.1, num estimators = 200
G-Mean Score	0.9954	0.9974	0.9246	0.9133	0.9959	0.9328	0.9964	0.9599

These scores notably surpass those from the original, RUS, and SMOTE datasets, suggesting improved model performance on the balanced dataset created via the SMOTEENN. Bagging, DT, KNN, and RF models stand out with the highest G-Mean scores, indicating their effectiveness in managing the balanced dataset and maintaining a good balance between sensitivity and specificity. The values that were obtained from the grid search highlight the advantages of using SMOTEENN to boost the performance of ML models on imbalanced datasets.

5.4 Performance Estimation

To assess model performance, the hyperparameter configuration that yielded the highest G-Mean score will be utilized. Once these optimal hyperparameters are determined, the question of which dataset version should be employed for model training arises. In response, an evaluation of the models using the corresponding TriCovB dataset version will be presented in each subsection below. Model performance will be assessed on the test set using metrics such as G-Mean, F1-Score, and AUC-ROC, providing a comprehensive evaluation of their capabilities. This evaluation process will facilitate objective comparisons of model performance across different dataset versions, enhancing our understanding of their effectiveness in handling various data imbalance scenarios.

5.4.1 Original Dataset

Upon completing the training of models on the original dataset with the optimal parameters identified through grid search, the focus transitioned to the phase of model evaluation. This phase includes the analysis of the outcomes of this evaluation, along with the presentation of their corresponding scores, all of which can be observed in Table 5.7.

TABLE 5.7: Models Evaluation for Original Dataset

Metric	**Model**							
	AdaBoost	**Bagging**	**DT**	**GB**	**KNN**	**LR**	**RF**	**XGBoost**
G-Mean	0.2975	0.4647	0.4617	**0.4764**	0.4428	0.4499	0.4567	0.4442
F1-Score	0.1557	0.2801	0.2623	**0.3241**	0.2651	0.3006	0.2774	0.2950
AUC-ROC	0.9084	0.8139	0.6667	**0.9132**	0.7147	0.9087	0.8438	0.9129

The GB model emerges as the superior performer in terms of G-Mean. This metric value suggests an optimal balance between sensitivity and specificity. This implies that the GB model exhibits good performance in classifying the positive class. Regarding the F1-Score, the GB model also exhibits superior performance compared to the other models. However, it is important to highlight that even though the GB model achieves the best score, it remains below the level of randomness. This signifies a notable deficiency in the model's precision, suggesting that its ability to correctly classify positive instances is somewhat limited.

The XGBoost model outperforms other models in terms of the AUC-ROC score, indicating its superior ability to differentiate between classes. This means that the XGBoost

model exhibits a higher capability to distinguish between positive and negative classes, independently of the threshold selected for classification.

On the other hand, the AdaBoost model registers the lowest scores for both G-Mean and F1-Score, suggesting a struggle in achieving a balance between sensitivity and specificity, and between precision and sensitivity. The DT model, with the lowest AUC-ROC score, demonstrates the weakest performance in distinguishing between classes across different thresholds. These outcomes imply that the models encounter challenges in striking a harmonious equilibrium between sensitivity and specificity, as well as precision and sensitivity. This suggests that the AdaBoost and the DT models might not be the most suitable option for the final model selection.

5.4.2 RUS Dataset

Similarly to the preceding subsection, subsequent to finalizing the model training process using the optimal parameters acquired through grid search for the RUS dataset, the investigation has transitioned into the phase of model evaluation. In this phase, the results of this assessment, along with their respective scores, are presented, as illustrated in Table 5.8.

TABLE 5.8: Models Evaluation for RUS Dataset

Metric	**Model**							
	AdaBoost	**Bagging**	**DT**	**GB**	**KNN**	**LR**	**RF**	**XGBoost**
G-Mean	0.8349	0.7986	0.8191	0.8377	0.8113	0.8354	**0.8382**	0.8369
F1-Score	0.2951	0.2462	0.2853	0.2967	0.2661	0.2982	**0.2984**	0.2959
AUC-ROC	0.9080	0.8756	0.8957	**0.9132**	0.8784	0.9091	0.9121	0.9102

The RF model emerges as the superior performer in terms of G-Mean, indicating that it has the best balance between sensitivity and specificity among all models. This suggests that the RF model is particularly effective in classifying the positive class while maintaining a robust balance between false positives and false negatives. The F1-Score analysis reveals that the RF model also outperforms its counterparts. Yet, an intriguing observation is that the top-performing F1-Score still falls short of a random baseline, similar to the pattern observed in the original dataset.

The GB model takes the lead in terms of the AUC-ROC score, which measures the model's ability to distinguish between classes across different threshold settings. This suggests that the GB model is more capable of distinguishing between the positive and negative classes, regardless of the specific threshold chosen for classification.

Conversely, the Bagging model records the least favorable scores across all evaluated metrics, indicating challenges in striking a balance between sensitivity and specificity, as well as precision and sensitivity. Moreover, it exhibits the least effective performance in distinguishing between classes across varying threshold values. These combined results raise concerns about its suitability for eventual model selection.

5.4.3 SMOTE Dataset

Following the same pattern as with the original and RUS datasets, the models were trained on the SMOTE dataset using the optimal parameters found through grid search. The results are displayed in Table 5.9.

TABLE 5.9: Models Evaluation for SMOTE Dataset

Metric	**Model**							
	AdaBoost	**Bagging**	**DT**	**GB**	**KNN**	**LR**	**RF**	**XGBoost**
G-Mean	**0.8216**	0.6629	0.6587	0.8153	0.7777	0.8151	0.6637	0.7975
F1-Score	0.2848	0.2285	0.2252	**0.2955**	0.2595	0.2834	0.2323	0.2871
AUC-ROC	**0.8961**	0.7872	0.6830	0.8893	0.8406	0.8877	0.8139	0.8769

The AdaBoost model excels in achieving a high G-Mean score, highlighting its proficiency in accurately classifying the positive class while maintaining a balanced false positive to false negative ratio. This capability extends to the AUC-ROC score, where the model effectively distinguishes between positive and negative classes at different threshold settings, emphasizing its strength in class separation.

The GB model leads in terms of the F1-Score among the models. However, it's important to note that even the best F1-Score achieved is lower than what would be expected by chance, indicating limited precision in positive case classification.

Conversely, the DT model consistently shows less favorable results across all evaluated metrics, implying challenges in achieving a balance between sensitivity and specificity, as well as precision and sensitivity. Additionally, it exhibits a limited ability to differentiate between classes at various threshold levels. These findings raise questions about its suitability for model selection.

5.4.4 SMOTEENN Dataset

Following the same pattern as in the previous dataset versions, the models have been trained on the SMOTEENN dataset. The results of evaluating these models in the test set are shown in Table 5.10.

TABLE 5.10: Models Evaluation for SMOTEENN Dataset

Metric	Model							
	AdaBoost	**Bagging**	**DT**	**GB**	**KNN**	**LR**	**RF**	**XGBoost**
G-Mean	**0.8306**	0.7643	0.7582	0.8289	0.7172	0.8255	0.7727	0.8171
F1-Score	0.2807	0.2929	0.2902	0.2899	**0.3241**	0.2779	0.3066	0.2909
AUC-ROC	0.9017	0.8312	0.7663	**0.9035**	0.7522	0.9019	0.8579	0.8980

The highest G-Mean score is attained by the AdaBoost model, analogous to the pattern observed in the SMOTE dataset. Consequently, the significance of this prominent performance mirrors the same of the prior subsection.

The evaluation of the F1-Score highlights the KNN model's dominance over the other models. However, it's worth noting that even the highest F1-Score attained is not able to surpass the level of randomness. This observation brings to light a precision deficiency in the model, suggesting that its accuracy in identifying positive instances is relatively low.

The GB model takes the lead regarding the AUC-ROC score, which measures the model's ability to distinguish between classes across different threshold settings. This suggests that the GB model is more capable of distinguishing between the positive and negative classes, regardless of the specific threshold chosen for classification.

The KNN model exhibits the lowest scores for G-Mean and AUC-ROC, indicating difficulties in achieving a balanced combination of sensitivity and specificity, as well as challenges in differentiating between classes at varying threshold values. Similarly, LR holds the lowest score for F1-Score, suggesting struggles in maintaining a robust balance between precision and sensitivity. These findings collectively suggest that both the KNN and LR models might not be the most optimal choices for the final model selection, given their limitations in achieving balanced performance across the assessed metrics.

5.5 Discussion

The results underscore the importance of hyperparameter tuning in improving model performance. The optimal hyperparameters varied across the different models and datasets, highlighting the need for a thorough and systematic approach to hyperparameter tuning. Future studies could explore more advanced hyperparameter optimization techniques, such as Bayesian optimization or genetic algorithms, to further improve model performance but more of this will be discussed in the next chapter.

The results also show how different ML models performed on various dataset versions, the original version being the ground truth for comparison, and each of the other versions designed to tackle the issue of imbalanced data. Compared to the original imbalanced dataset, all the ML models had a better performance when trained on the resampled versions. These findings emphasize the importance of addressing class imbalance in training data, as it directly impacts the models' ability to make accurate predictions.

In the original dataset, the models struggled to achieve a balanced trade-off between sensitivity and specificity, evident from the low scores obtained across all models for the metrics G-Mean and F1-Score, even falling below random levels. However, through resampling techniques, we observed a substantial improvement in overall performance. Across all dataset versions, several common trends and distinctions have emerged, providing insights into the suitability and limitations of various ML models.

In terms of the G-Mean score, it is intriguing to observe consistent trends across different dataset versions. The GB, and the RF models got the highest G-Mean scores. The AdaBoost model was both among the models with the lowest and the highest G-Mean score. Bagging, DT, and KNN models also held the lowest G-Mean score. This observation shows the limitations of these models in achieving the desired equilibrium.

Regarding the F1-Score, a unique finding comes to the forefront. While different dataset versions display variations in performance, an intriguing trend prevails: even the top-performing F1-Scores across all versions remain below the random baseline. This suggests that the models, although showing improved sensitivity, struggle to maintain an adequate level of precision. This fact corroborates the difficulty of achieving high precision in predicting COVID-19 patient hospitalization, regardless of the technique employed.

In terms of the AUC-ROC score, a consistent trend is observed where the foremost positions are consistently held by the XGBoost, GB, and AdaBoost models, underscoring their proficiency in effectively discriminating between positive and negative cases. Conversely, lower scores are associated with the KNN, DT, and Bagging models, positioning them among the models with comparatively poorer performance in this regard.

To sum up the best scores and the corresponding models for each version of the dataset, the table 5.11 was built.

The results presented in this table allow some conclusions that go as follows:

- **G-Mean:** The optimal resampling technique in this context is Random Under-sampling. This approach is advantageous as it minimizes the dataset's size, thereby reducing

TABLE 5.11: Best Scores for Each Metric on Different Dataset Versions

Metric	Original	RUS	SMOTE	SMOTEENN
G-Mean	0.4764 (GB)	0.8382 (RF)	0.8216 (AdaBoost)	0.8306 (AdaBoost)
F1-Score	0.3241 (GB)	0.2984 (RF)	0.2955 (GB)	0.3241 (KNN)
AUC-ROC	0.9129 (XGBoost)	0.9132 (GB)	0.8961 (AdaBoost)	0.9035 (GB)

the computational resources required. The model of choice, Random Forest, brings several benefits. It is capable of managing high-dimensional spaces and numerous features effectively. It also exhibits robustness to outliers and possesses an inherent feature selection capability due to the random subspace method. Additionally, Random Forest models, through their ensemble nature, are less prone to overfitting compared to individual decision trees. This is because the ensemble approach averages out biases and diminishes variance.

- **F1-Score:** This metric, generally, offers a well-rounded evaluation of model performance, particularly in imbalanced dataset scenarios, nevertheless the values obtained are notably low. The models with the highest F1-Score are GB and KNN, but even so, it is worse than random. This usually happens when the model is not performing well in terms of precision, meaning that it is generating a significant number of false positives [192]. The low scores suggest that there are areas for further optimization. This could involve additional feature engineering, parameter tuning, or the exploration of other resampling or modeling techniques.

- **AUC-ROC:** High AUC-ROC values across all versions of the dataset suggests that the models are doing a good job of distinguishing between the two classes, regardless of the dataset version. This is a positive outcome, as it indicates that the models have learned meaningful patterns from the data that allow them to differentiate between the classes effectively. Noteworthy that this metric gets slightly better in the RUS version, although its value on the original dataset is the second best. This makes it a good metric for imbalanced datasets, as it's not sensitive to the class imbalance itself.

  However, a slight decrease in AUC-ROC for more complex resampling techniques like SMOTE and SMOTEENN was observed. This could be due to a variety of factors. One possibility is that these techniques, while they do a good job of balancing the classes, may also introduce some noise or artificial patterns into the data that

make the classification task slightly more challenging. For instance, SMOTE creates synthetic examples in the feature space, which can sometimes lead to over-generalization or the creation of instances that are harder to classify. Similarly, SMOTEENN combines over-sampling of the minority class (SMOTE) with under-sampling of the majority class (ENN), which can lead to a more complex decision boundary that might not necessarily result in better AUC-ROC.

Another possibility is that the models are overfitting to the resampled training data, which could result in lower performance on the test set, as reflected in the AUC-ROC. Overfitting is a common risk with more complex models or when using techniques that significantly alter the training data, as they can cause the model to learn patterns that are specific to the training data but do not generalize well to new data.

In conclusion, our initial findings highlight the challenges associated with class imbalance in the dataset and the complexities of predicting COVID-19 patient hospitalization. Our primary objective is to create a robust model that assists healthcare professionals in deciding when hospitalization is necessary for COVID-19 patients. To achieve this, we propose the use of RUS to tackle class imbalance. The choice lies between GB and RF as the model depends on the evaluation metric. If prioritizing AUC-ROC with a score of 0.91, go for GB; if focusing on G-Mean with a score of 0.82, opt for RF.

The resulting model can be seamlessly integrated into various health chatbots designed for RPM, enabling timely assessments of the need for hospitalization, which can substantially benefit healthcare decision-making.

Chapter 6

Conclusions

This chapter recaps our study on how ML can be applied to preventive healthcare having as a specific use case the COVID-19 disease. We highlight our main contributions and key achievements in applying ML to predictive models for healthcare. We also present some of the future work topics regarding the intersection of healthcare and ML.

6.1 Contributions

It's essential to recognize that this work represents just the beginning of a more extensive journey aimed at improving ML for healthcare applications. This dissertation has made a few contributions to preventive care and ML, in the context of the COVID-19 pandemic, which we describe next.

- **Presentation and Analysis of Datasets:** One contribution of this research lies in the presentation and analysis of COVID-19 datasets regarding patient characteristics. Through an evaluation of the available COVID-19 datasets, this dissertation ensures that the chosen datasets align with the research goals. This informed decision-making process, based on the understanding of the datasets' strengths and limitations, reinforces the relevance and applicability of the study's outcomes.
- **Nature of the Data:** This research highlights the significant implications of using *simple data* variables, such as age, sex, symptoms, and comorbidities, to predict the

likelihood of hospitalization for COVID-19 patients (Chapters 1 and 2). This approach demonstrates the potential of using readily available data for critical healthcare decisions, resource allocation, and patient management. This is important in the scope of preventive care.

- **Exploratory Data Analysis:** The exploratory data analysis conducted on the chosen dataset represents another valuable contribution. This EDA effort revealed insights into the relationships between the independent variables and the target variable. For example, we observed that individuals who eventually required hospitalization tended to be older males, experiencing prominent symptoms like difficulty breathing, along with the presence of cardiovascular comorbidities and diabetes. Additionally, our cluster analysis revealed distinct grouping patterns within symptoms and comorbidities. This analysis established a robust basis for our subsequent model development, underscoring the crucial role of comprehending the data before venturing into predictive modeling.

- **Addressing Class Imbalance:** One other contribution is the application of diverse resampling techniques to mitigate class imbalance within the dataset. With the positive class accounting for only 3.9% of the data, this imbalance is a common hurdle in medical data analysis. The study showcases how balanced datasets substantially enhance ML model performance (when using G-Means and AUC-ROC metrics), by employing methods like RUS, SMOTE, and SMOTEENN. This contribution boosts model robustness and offers insights for researchers tackling imbalanced data challenges or using TriCovB.

- **Final Result and Practical Application:** The final and perhaps most significant contribution is the development of an ML model ready to be integrated into an intelligent chatbot as a practical application of the research findings. The research has demonstrated how ML can be harnessed to provide actionable insights in a real-world context. This contribution not only showcases the practical applicability of the research but also provides a tangible tool that can be used to predict the need for hospitalization in COVID-19 patients, or potentially, in the context of other diseases.

In conclusion, this dissertation represents a humble yet meaningful contribution to the field of preventive care and ML. It is hoped that the findings will inspire further research in this area, ultimately leading to more robust and reliable predictive models for healthcare applications.

6.2 Future Work

The research presented in this thesis has laid a solid foundation for the application of ML techniques in the development of an intelligent chatbot for preventive care. The results obtained demonstrate the potential of ML in this domain and highlight the importance of addressing the challenges posed by imbalanced datasets.

However, like all research pursuits, there are many potential avenues for future work that could build upon the findings and insights gained from this thesis. The following list outlines a few representative topics and examples.

- **Exploration of Additional Datasets:** We chose the TriCovB dataset for its size and minimal missing data. Accessing COVID-19 patient data is challenging due to privacy and logistics. Future research may explore diverse datasets for validation and insights, potentially revealing region-specific factors and refining our methodology.

- **Incorporation of Additional Features:** The study primarily focused on vital features like age, sex, symptoms, and comorbidities due to data availability and importance. However, COVID-19's complexity suggests value in unexplored factors like health history, genetics, and socio-economics. Future research may enrich features with new data sources, potentially enhancing predictions, though careful selection and validation are crucial to avoid overfitting.

- **Use of Advanced Resampling Techniques:** This study applied common resampling techniques to tackle class imbalance effectively. While these methods enhance ML performance, advanced techniques like ADASYN and Borderline-SMOTE provide nuanced approaches. Emerging ensemble-based methods offer further potential. Future research can explore these techniques for COVID-19 patient hospitalization prediction, but their complexity and computational demands require careful assessment. Resampling choices should align with dataset characteristics, warranting systematic evaluation and benchmarking studies for optimal selection. While

this research demonstrated effective techniques, exploring advanced options holds promise for optimization, necessitating careful consideration of complexities and challenges.

- **Experimentation with Different Model Architectures:** In this study, ML model selection was guided by healthcare and COVID-19 prediction literature. The ML landscape offers various models like Support Vector Machines, Neural Networks, and advanced ensembles, each with unique strengths and weaknesses tied to dataset characteristics. Deep learning methods, including Convolutional and Recurrent Neural Networks, hold promise for diverse data types. Future research should systematically compare these models to determine their strengths and weaknesses. Optimizing hyperparameters, potentially through Bayesian or genetic algorithms, is crucial for model performance. While this study demonstrated model efficacy, exploring alternative models and methodologies is essential for improving COVID-19 patient hospitalization prediction.

- **Refinement of Hyperparameter Tuning:** Grid search systematically explores a hyperparameter subset, while random search efficiently handles numerous parameters. Bayesian optimization intelligently selects hyperparameters based on past data, ideal for limited evaluations. Future research should explore these advanced tuning methods, comparing their performance and resource requirements. Identifying the most effective hyperparameter tuning approach for predicting COVID-19 patient hospitalization requirements is crucial. While grid search worked well here, advanced methods may further enhance model performance.

- **Integration into an Intelligent Chatbot:** The project aims to integrate the classifier into an existing intelligent preventive care chatbot [105]. This integration would enable real-time predictions of COVID-19 patient hospitalization needs based on user inputs. The development of such a chatbot has the potential to revolutionize COVID-19 management by facilitating early intervention. Future work should focus on refining the chatbot for enhanced performance and user practicality.

Bibliography

[1] J. R. Gallaher, A. Yohann, C. Kajombo, A. Schneider, L. Purcell, and A. Charles, "Reallocation of hospital resources during covid-19 pandemic and effect on trauma outcomes in a resource-limited setting," *World Journal of Surgery*, vol. 46, pp. 2036–2044, 9 2022. [Online]. Available: https://link.springer.com/article/10.1007/s00268-022-06636-4 [Cited on page 1.]

[2] R. Filip, R. G. Puscaselu, L. Anchidin-Norocel, M. Dimian, and W. K. Savage, "Global challenges to public health care systems during the covid-19 pandemic: A review of pandemic measures and problems," *Journal of Personalized Medicine*, vol. 12, 8 2022. [Online]. Available: /pmc/articles/PMC9409667//pmc/articles/PMC9409667/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9409667/ [Cited on page 1.]

[3] S. Mantena and S. Keshavjee, "Strengthening healthcare delivery with remote patient monitoring in the time of covid-19," *BMJ Health Care Informatics*, vol. 28, p. 100302, 7 2021. [Online]. Available: /pmc/articles/PMC8300556/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8300556/ [Cited on page 1.]

[4] D. A. Hoffman, "Increasing access to care: telehealth during covid-19," *Journal of Law and the Biosciences*, vol. 7, pp. 1–15, 2020. [Online]. Available: /pmc/articles/PMC7337821//pmc/articles/PMC7337821/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7337821/ [Cited on page 2.]

[5] A. Dubey and A. Tiwari, "Artificial intelligence and remote patient monitoring in us healthcare market: a literature review," *Journal of Market Access Health Policy*, vol. 11, 2023. [Online].

Available: /pmc/articles/PMC10158563//pmc/articles/PMC10158563/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC10158563/ [Cited on page 2.]

[6] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: transforming the practice of medicine," *Future Healthcare Journal*, vol. 8, p. e188, 7 2021. [Online]. Available: /pmc/articles/PMC8285156//pmc/articles/PMC8285156/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8285156/ [Cited on page 2.]

[7] D. M. El-Sherif, M. Abouzid, M. T. Elzarif, A. A. Ahmed, A. Albakri, and M. M. Alshehri, "Telehealth and artificial intelligence insights into healthcare during the covid-19 pandemic," *Healthcare*, vol. 10, 2 2022. [Online]. Available: /pmc/articles/PMC8871559//pmc/articles/PMC8871559/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8871559/ [Cited on page 2.]

[8] "iCare4NextG Project," https://www.celticnext.eu/project-icare4nextg/. [Cited on page 2.]

[9] I. Ozkan and T. Teknoloji, "Project information." [Online]. Available: http://www.icare4nextg.eu/ [Cited on page 2.]

[10] E. E. Thomas, M. L. Taylor, A. Banbury, C. L. Snoswell, H. M. Haydon, V. M. G. Rejas, A. C. Smith, and L. J. Caffery, "Original research: Factors influencing the effectiveness of remote patient monitoring interventions: a realist review," *BMJ Open*, vol. 11, p. 51844, 8 2021. [Online]. Available: /pmc/articles/PMC8388293//pmc/articles/PMC8388293/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8388293/ [Cited on page 2.]

[11] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare Journal*, vol. 6, p. 94, 6 2019. [Online]. Available: /pmc/articles/PMC6616181//pmc/articles/PMC6616181/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/ [Cited on page 3.]

[12] A. Haleem, M. Javaid, R. P. Singh, and R. Suman, "Telemedicine for healthcare: Capabilities, features, barriers, and applications," *Sensors International*, vol. 2, p. 100117, 1 2021. [Online]. Available: /pmc/articles/PMC8590973//pmc/articles/PMC8590973/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8590973/ [Cited on page 3.]

[13] S. J. Russell and P. Norvig, *Artificial Intelligence: a modern approach*, 3rd ed. Pearson, 2009. [Cited on page 5.]

[14] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers, 2011. [Cited on page 6.]

[15] C. M. Bishop, "Pattern recogniton and machine learning," *Information Science and Statistics*, p. 738, 2006. [Online]. Available: https://www.springer.com/gp/book/9780387310732http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop-PatternRecognitionAndMachineLearning-Springer2006.pdf [Cited on pages 6 and 10.]

[16] E. Alpaydin, *Introduction to Machine Learning*. MIT Press, 2010. [Cited on pages 10 and 11.]

[17] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2019. [Cited on pages 6, 9, and 10.]

[18] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, pp. 1–21, 5 2021. [Online]. Available: https://link.springer.com/article/10.1007/s42979-021-00592-x [Cited on pages 6 and 7.]

[19] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2005. [Cited on page 6.]

[20] C. C. Aggarwal, *Data Mining: The Textbook*. Cham: Springer, 2015. [Cited on page 7.]

[21] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012. [Cited on page 7.]

[22] Željko Vujović, "Classification model evaluation metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, pp. 599–606, 15 2021. [Online]. Available: www.ijacsa.thesai.org [Cited on page 9.]

[23] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.ipm.2009.03.002 [Cited on page 9.]

[24] G. I. Webb, J. Fürnkranz, J. Fürnkranz, J. Fürnkranz, G. Hinton, C. Sammut, J. Sander, M. Vlachos, Y. W. Teh, Y. Yang, D. Mladeni, J. Brank, M. Grobelnik, Y. Zhao, G. Karypis, S. Craw, M. L. Puterman, and J. Patrick, "Decision tree," *Encyclopedia of Machine Learning*, pp. 263–267, 2011. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007/978-0-387-30164-8_204 [Cited on page 11.]

[25] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, "k-nearest neighbor classification," pp. 83–106, 2009. [Online]. Available: https://link.springer.com/chapter/10.1007/978-0-387-88615-2_4 [Cited on page 11.]

[26] B. F. French, H. C. Immekus, and H.-J. Yen, "Logistic regression," *Handbook of Quantitative Methods for Educational Research*, pp. 145–165, 2013. [Online]. Available: https://link.springer.com/chapter/10.1007/978-94-6209-404-8_7 [Cited on page 11.]

[27] P. Favaro and A. Vedaldi, "Adaboost," *Computer Vision*, pp. 16–19, 2014. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007/978-0-387-31439-6_663 [Cited on page 12.]

[28] C. Vens, "Bagging," *Encyclopedia of Systems Biology*, pp. 68–69, 2013. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007/978-1-4419-9863-7_602 [Cited on page 12.]

[29] V. K. Ayyadevara, "Gradient boosting machine," *Pro Machine Learning Algorithms*, pp. 117–134, 2018. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4842-3564-5_6 [Cited on page 12.]

[30] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 10 2001. [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324 [Cited on page 12.]

[31] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939785 [Cited on page 12.]

[32] S. S. Aljameel, I. U. Khan, N. Aslam, M. Aljabri, and E. S. Alsulmi, "Machine learning-based model to predict the disease severity and outcome in covid-19 patients," *Scientific Programming*, vol. 2021, 2021. [Cited on page 12.]

[33] K. Moulaei, M. Shanbehzadeh, Z. Mohammadi-Taghiabad, and H. Kazemi-Arpanahi, "Comparing machine learning algorithms for predicting covid-19 mortality," *BMC Medical Informatics and Decision Making*, vol. 22, pp. 1–12, 12 2022. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01742-0 [Cited on page 12.]

[34] S. S. Zakariaee, N. Naderi, M. Ebrahimi, and H. Kazemi-Arpanahi, "Comparing machine learning algorithms to predict covid-19 mortality using a dataset including chest computed tomography severity score data," *Scientific reports*, vol. 13, p. 11343, 12 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/37443373/ [Cited on pages 13 and 21.]

[35] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org. [Cited on page 13.]

[36] C. C. Aggarwal, *Neural Networks and Deep Learning: A Textbook*. Springer, 2018. [Cited on page 13.]

[37] K. Kumar, G. Sundar, and M. Thakur, "Information technology and computer science," *Information Technology and Computer Science*, vol. 6, pp. 57–68, 2012. [Online]. Available: http://www.mecs-press.org/ [Cited on pages 13 and 22.]

[38] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Beijing: O'Reilly, 2009. [Online]. Available: http://www.nltk.org/book [Cited on page 13.]

[39] P. J. Pronovost, M. D. Cole, and R. M. Hughes, "Remote patient monitoring during covid-19: An unexpected patient safety benefit," *JAMA*, vol. 327, pp. 1125–1126, 3 2022. [Online]. Available: https://jamanetwork.com/journals/jama/fullarticle/2789635 [Cited on page 14.]

[40] K. Bouabida, K. Malas, A. Talbot, M. Ève Desrosiers, F. Lavoie, B. Lebouché, N. Taghizadeh, L. Normandin, C. Vialaron, O. Fortin, D. Lessard, and M. P. Pomey, "Healthcare professional perspectives on the use of remote patient-monitoring platforms during the covid-19 pandemic: A cross-sectional study," *Journal of Personalized Medicine 2022, Vol. 12, Page 529*, vol. 12, p. 529, 3 2022. [Online]. Available: https://www.mdpi.com/2075-4426/12/4/529/htmhttps://www.mdpi.com/2075-4426/12/4/529 [Cited on page 14.]

[41] A. Alboksmaty, T. Beaney, S. Elkin, J. M. Clarke, A. Darzi, P. Aylin, and A. L. Neves, "Effectiveness and safety of pulse oximetry in remote patient monitoring of patients with covid-19: a systematic review," *The Lancet Digital Health*, vol. 4, pp. e279–e289, 4 2022. [Cited on page 14.]

[42] P. Kaur, A. A. Mack, N. Patel, A. Pal, R. Singh, A. Michaud, M. Mulflur, P. Kaur, A. A. Mack, N. Patel, A. Pal, R. Singh, A. Michaud, and M. Mulflur, "Unlocking the potential of artificial intelligence (ai) for healthcare," 4 2023. [Online]. Available: https://www.intechopen.com/online-first/87001undefined/online-first/87001 [Cited on page 14.]

[43] J. Oliver, M. Dutch, A. Rojek, M. Putland, and J. C. Knott, "Remote covid-19 patient monitoring system: a qualitative evaluation," *BMJ Open*, vol. 12, p. e054601, 5 2022. [Online]. Available: https://bmjopen.bmj.com/content/12/5/e054601https://bmjopen.bmj.com/content/12/5/e054601.abstract [Cited on page 14.]

[44] G. Saranya, N. Dineshkumar, A. S. Hariprasath, and G. Jeevanantham, "Design of iot enabled cloud assisted health monitoring system for covid-19 patients." Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1227–1232. [Cited on page 14.]

[45] V. Health, "Vivify health," https://www.vivifyhealth.com/. [Cited on page 15.]

[46] Philips, "ecarecoordinator clinical dashboard for ambulatory health," https://www.philips.com.au/healthcare/ecarecoordinator-clinical-dashboard-for-ambulatory-health, accessed: March 31, 2023. [Cited on page 15.]

[47] H. Bansal and R. Khan, "A review paper on human computer interaction," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 8, p. 53, 4 2018. [Cited on page 15.]

[48] A. Khanna, B. Pandey, K. Vashishta, K. Kalia, B. Pradeepkumar, and T. Das, "A study of today's a.i. through chatbots and rediscovery of machine intelligence," *International Journal of u- and e-Service, Science and Technology*, vol. 8, pp. 277–284, 7 2015. [Cited on page 15.]

[49] A. M. TURING, "I.—computing machinery and intelligence," *Mind*, vol. LIX, pp. 433–460, 10 1950. [Online]. Available: https://academic.oup.com/mind/article/LIX/236/433/986238 [Cited on page 15.]

[50] D. Yoneoka, T. Kawashima, Y. Tanoue, S. Nomura, K. Ejima, S. Shi, A. Eguchi, T. Taniguchi, H. Sakamoto, H. Kunishima, S. Gilmour, H. Nishiura, and H. Miyata, "Early sns-based monitoring system for the covid-19 outbreak in japan: a population-level observational study," *jstage.jst.go.jp*, vol. 30, pp. 362–370, 2020. [Online]. Available: https://www.jstage.jst.go.jp/article/jea/30/8/30_JE20200150/_article/-char/ja/ [Cited on pages 16 and 17.]

[51] U. Bharti, D. Bajaj, H. Batra, S. Lalit, S. Lalit, and A. Gangwani, "Medbot: Conversational artificial intelligence powered chatbot for delivering tele-health after covid-19," pp. 870–875, 7 2020. [Cited on pages 16, 17, 23, and 28.]

[52] C. Rodsawang, P. Thongkliang, T. Intawong, A. S. O. Journal, and undefined 2020, "Designing a competent chatbot to counter the covid-19 pandemic and empower risk communication in an emergency response system," *osirjournal.net*. [Online]. Available: http://www.osirjournal.net/index.php/osir/article/view/193 [Cited on pages 16 and 17.]

[53] G. Battineni, N. Chintalapudi, and F. Amenta, "Ai chatbot design during an epidemic like the novel coronavirus," *mdpi.com*. [Online]. Available: https://www.mdpi.com/733910 [Cited on pages 16, 17, and 18.]

[54] E. Meinert, M. Milne-Ives, and S. Surodina, "Agile requirements engineering and software planning for a digital health platform to engage the effects of isolation caused by social distancing: case study," *publichealth.jmir.org*. [Online]. Available: https://publichealth.jmir.org/2020/2/e19297/ [Cited on pages 16, 17, and 18.]

[55] T. Judson, A. Odisho, J. Y. J. of the ..., and undefined 2020, "Implementation of a digital chatbot to screen health system employees during the covid-19 pandemic," *academic.oup.com*. [Online]. Available: https://academic.oup.com/jamia/article-abstract/27/9/1450/5856745 [Cited on pages 16 and 18.]

[56] P. Amiri and E. Karahanna, "Chatbot use cases in the covid-19 public health response," *Journal of the American Medical Informatics Association*, vol. 29, pp. 1000–1010, 4 2022. [Online]. Available: https://dx.doi.org/10.1093/jamia/ocac014 [Cited on page 16.]

[57] M. Silva and J. Santos, "Ana: a brazilian chatbot assistant about covid-19," *ResearchGate*, 2020. [Online]. Available: https://www.researchgate.net/publication/342869406_Ana_a_brazilian_chatbot_assistant_about_covid_19 [Cited on pages 16, 17, and 18.]

[58] A. R. Dennis, A. Kim, M. Rahimi, and S. Ayabakan, "User reactions to COVID-19 screening chatbots from reputable providers," *J. Am. Med. Inform. Assoc.*, vol. 27, no. 11, pp. 1727–1731, Nov. 2020. [Cited on pages 16 and 17.]

[59] M. V. Pawar, A. M. Pawar, H. Bhapkar, J. Anuradha, R. Bachate, A. Sharma, S. Bhoyar, and N. Shardoor, "Artificial intelligence-based solutions for covid-19," *Data Science for COVID-19*, p. 167, 1 2022. [Online]. Available: /pmc/articles/PMC8988883//pmc/articles/PMC8988883/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8988883/ [Cited on page 18.]

[60] U. Albalawi and M. Mustafa, "Current artificial intelligence (ai) techniques, challenges, and approaches in controlling and fighting covid-19: A review," *International Journal of Environmental Research and Public Health*, vol. 19, 5 2022. [Online]. Available: /pmc/articles/PMC9140632//pmc/articles/PMC9140632/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9140632/ [Cited on page 18.]

[61] M. C. E. theory and moral practice, "Health care, capabilities, and ai assistive technologies," *Springer*. [Online]. Available: https://link.springer.com/article/10.1007/s10677-009-9186-2 [Cited on page 19.]

[62] T. Nadarzynski, O. Miles, A. Cowie, and D. Ridge, "Acceptability of artificial intelligence (ai)-led chatbot services in healthcare: A mixed-methods study," *https://doi.org/10.1177/2055207619871808*, vol. 5, 8 2019. [Online]. Available: https://journals.sagepub.com/doi/10.1177/2055207619871808 [Cited on page 19.]

[63] B. Beck, B. Shin, Y. Choi, S. Park, K. K. Computational, structural, and undefined 2020, "Predicting commercially available antiviral drugs that may act on the novel coronavirus (sars-cov-2) through a drug-target interaction deep learning model," *Elsevier*. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037020300490 [Cited on page 19.]

[64] A. Cossy-Gantner, S. Germann, N. R. Schwalbe, and B. Wahl, "Artificial intelligence (ai) and global health: how can ai contribute to health in resource-poor settings?" *BMJ Global Health*, vol. 3, p. e000798, 8 2018. [Online]. Available: https://gh.bmj.com/content/3/4/e000798https://gh.bmj.com/content/3/4/e000798.abstract [Cited on page 19.]

[65] M. H. Alsharif, Y. H. Alsharif, S. A. Chaudhry, and M. A. M. Albreem, "Artificial intelligence technology for diagnosing covid-19 cases: A review of substantial issues," *researchgate.net*. [Online]. Available: https://www.researchgate.net/profile/Abu-Jahid/publication/344356974_Artificial_intelligence_technology_for_diagnosing_COVID-19_cases_a_review_of_substantial_issues/links/5f6d37cc299bf1b53ef0a15c/Artificial-intelligence-technology-for-diagnosing-COVID-19-cases-a-review-of-substantial-issues.pdf [Cited on page 19.]

[66] J. Chen, K. Li, Z. Zhang, K. Li, P. Y. A. C. S. (CSUR), and undefined 2021, "A survey on applications of artificial intelligence in fighting against covid-19," *dl.acm.org*, vol. 54, 11 2021. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3465398 [Cited on page 19.]

[67] J. S. I. J. of Information Management and undefined 2020, "Considerations for development and use of ai in response to covid-19," *Elsevier*. [Online]. Available:

https://www.sciencedirect.com/science/article/pii/S026840122030949X [Cited on page 19.]

[68] A. Kumar, P. K. Gupta, and A. Srivastava, "A review of modern technologies for tackling covid-19 pandemic," *Diabetes Metabolic Syndrome: Clinical Research Reviews*, vol. 14, pp. 569–573, 7 2020. [Cited on page 19.]

[69] A. Salehi, P. Baglat, and G. Gupta, "Review on machine and deep learning models for the detection and prediction of coronavirus," *Elsevier*. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214785320347131 [Cited on pages 19 and 22.]

[70] E. Mbunge, "Integrating emerging technologies into covid-19 contact tracing: Opportunities, challenges and pitfalls," *Elsevier*. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1871402120303325 [Cited on page 19.]

[71] M. Khan, M. T. Mehran, Z. U. Haq, Z. Ullah, S. R. Naqvi, M. Ihsan, and H. Abbass, "Applications of artificial intelligence in covid-19 pandemic: A comprehensive review," *Expert Systems with Applications*, vol. 185, p. 115695, 12 2021. [Cited on page 19.]

[72] R. Vaishya and et al., "Artificial intelligence (ai) applications for covid-19 pandemic," *Diabetes Metabolic Syndrome: Clinical Research Reviews*, vol. 14, pp. 337–339, 7 2020. [Cited on pages 19 and 28.]

[73] O. Albahri, A. Zaidan, A. Albahri, and B. Zaidan, "Systematic review of artificial intelligence techniques in the detection and classification of covid-19 medical images in terms of evaluation and benchmarking," *Elsevier*. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187603412030558X [Cited on pages 19 and 22.]

[74] O. Gozes, M. . A. Frid-Adar, H. Greenspan, P. D. Browning, H. Zhang, W. Ji, A. Bernheim, and E. Siegel, "Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection patient monitoring using deep learning ct image analysis," *arxiv.org*. [Online]. Available: https://arxiv.org/abs/2003.05037 [Cited on pages 19 and 22.]

[75] M. Otoom, N. Otoum, M. A. Alzubaidi, Y. Etoom, and R. Banihani, "An iot-based framework for early identification and monitoring of covid-19 cases," *Biomedical Signal Processing and Control*, vol. 62, p. 102149, 9 2020. [Cited on page 20.]

[76] L. Wang, X. Chen, Y. Zhai, F. Zhu, H. Chen, Y. Wang, X. Su, S. Huang, L. Tian, W. Zhu, W. Sun, L. Zhang, Q. Han, J. Zhang, F. Pan, L. Chen, Z. Zhu, H. Xiao, Y. Liu, W. Chen, and T. Li, "A novel triage tool of artificial intelligence-assisted diagnosis aid system for suspected covid-19 pneumonia in fever clinics," *medrxiv.org*. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.03.19.20039099.abstract [Cited on pages 20 and 28.]

[77] L. Nguyen, N. Pham, L. Nguyen, and T. Nguyen, "Fruit-cov: An efficient vision-based framework for speedy detection and diagnosis of sars-cov-2 infections through recorded cough sounds," *Elsevier*. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422022308 [Cited on page 20.]

[78] L. N. Mahdy, K. A. Ezzat, H. H. Elmousalami, H. A. Ella, and A. E. Hassanien, "Automatic x-ray covid-19 lung image classification system based on multi-level thresholding and support vector machine," *medrxiv.org*. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.03.30.20047787.abstract [Cited on page 20.]

[79] Z. Meng, M. Wang, H. Song, S. Guo, Y. Zhou, W. Li, Y. Zhou, M. Li, X. Song, Y. Zhou, Q. Li, X. Lu, and B. Ying, "Development and utilization of an intelligent application for aiding covid-19 diagnosis," *medrxiv.org*. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.03.18.20035816.abstract [Cited on page 20.]

[80] R. Carrillo-Larco and M. Castillo-Cara, "Using country-level variables to classify countries according to the number of confirmed covid-19 cases: An unsupervised machine learning approach," *ncbi.nlm.nih.gov*. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7308996/ [Cited on page 20.]

[81] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, and Y. Zha, "Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of

covid-19 pneumonia using computed tomography," *Elsevier*. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867420305511

[82] X. Mei, H. Lee, K. Diao, M. Huang, B. Lin, C. Liu, and Z. Xie, "Artificial intelligence–enabled rapid diagnosis of patients with covid-19," *nature.com*. [Online]. Available: https://www.nature.com/articles/s41591-020-0931-3

[83] R. Magar, P. Yadav, and A. B. Farimani, "Potential neutralizing antibodies discovered for novel corona virus using machine learning," *nature.com*. [Online]. Available: https://www.nature.com/articles/s41598-021-84637-4 [Cited on page 20.]

[84] L. Huang, S. Ruan, T. D. T. F. I. C. on, and undefined 2021, "Covid-19 classification with deep neural network and belief functions," *dl.acm.org*, 7 2021. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3469678.3469719 [Cited on page 20.]

[85] S. Liang, H. Liu, Y. Gu, X. Guo, H. Li, L. Li, and Z. Wu, "Fast automated detection of covid-19 from medical images using convolutional neural networks," *nature.com*. [Online]. Available: https://www.nature.com/articles/s42003-020-01535-7

[86] Y. D. Zhang, S. C. Satapathy, S. Liu, and G. R. Li, "A five-layer deep convolutional neural network with stochastic pooling for chest ct-based covid-19 diagnosis," *Machine Vision and Applications*, vol. 32, 2 2021. [Cited on pages 20 and 29.]

[87] N. Ahmed, R. A. Michelin, W. Xue, S. Ruj, R. Malaney, S. S. Kanhere, A. Seneviratne, W. Hu, H. Janicke, and S. K. Jha, "A survey of covid-19 contact tracing apps," *IEEE Access*, vol. 8, pp. 134 577–134 601, 2020. [Cited on page 20.]

[88] C. for Disease Control and Prevention, "How to protect yourself others," 2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html [Cited on page 20.]

[89] A. Akinbi, M. Forshaw, and V. Blinkhorn, "Contact tracing apps for the covid-19 pandemic: a systematic literature review of challenges and future directions for neo-liberal societies," *Health Information Science and Systems*, vol. 9, pp. 1–15, 12 2021. [Online]. Available: https://link.springer.com/article/10.1007/s13755-021-00147-7 [Cited on page 21.]

[90] R. Abbas and K. Michael, "Covid-19 contact trace app deployments: Learnings from australia and singapore," *IEEE Consumer Electronics Magazine*, vol. 9, pp. 65–70, 9 2020. [Cited on page 21.]

[91] O. Shahid, M. Nasajpour, S. Pouriyeh, R. M. Parizi, M. Han, M. Valero, F. Li, M. Aledhari, and Q. Z. Sheng, "Machine learning research towards combating covid-19: Virus detection, spread prevention, and medical assistance," *Journal of Biomedical Informatics*, vol. 117, p. 103751, 5 2021. [Cited on page 21.]

[92] J. L. Raisaro, F. Marino, J. Troncoso-Pastoriza, R. Beau-Lejdstrom, R. Bellazzi, R. Murphy, E. V. Bernstam, H. Wang, M. Bucalo, Y. Chen, A. Gottlieb, A. Harmanci, M. Kim, Y. Kim, J. Klann, C. Klersy, B. A. Malin, M. Meán, F. Prasser, L. Scudeller, A. Torkamani, J. Vaucher, M. Puppala, S. T. Wong, M. Frenkel-Morgenstern, H. Xu, B. M. Musa, A. G. Habib, T. Cohen, A. Wilcox, H. M. Salihu, H. Sofia, X. Jiang, and J. P. Hubaux, "Scor: A secure international informatics infrastructure to investigate covid-19," *Journal of the American Medical Informatics Association*, vol. 27, pp. 1721–1726, 11 2020. [Online]. Available: https://academic.oup.com/jamia/article/27/11/1721/5869802 [Cited on page 21.]

[93] M. Hamza, A. A. Khan, and M. A. Akbar, "Toward a secure global contact tracing app for covid-19," *ACM International Conference Proceeding Series*, pp. 453–460, 6 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3530019.3531339 [Cited on page 21.]

[94] S. Technologies, "Home," 2021. [Online]. Available: https://sqreemtech.com/ [Cited on page 21.]

[95] S. Arora, A. Kumar, and S. Sambhav, "Analysing the effect of gender on mortality of covid-19 patients through cox-proportional hazard model," *2021 International Conference on Intelligent Technologies, CONIT 2021*, 6 2021. [Cited on pages 21 and 74.]

[96] N. Alballa and I. Al-Turaiki, "Machine learning approaches in covid-19 diagnosis, mortality, and severity risk prediction: A review," *Informatics in Medicine Unlocked*, vol. 24, p. 100564, 1 2021. [Cited on page 21.]

[97] M. Pourhomayoun and M. Shakibi, "Predicting mortality risk in patients with covid-19 using machine learning to help medical decision-making," *Smart*

health (Amsterdam, Netherlands), vol. 20, 4 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33521226/ [Cited on pages 21, 28, and 30.]

[98] R. Gupta, A. Ghosh, A. K. Singh, and A. Misra, "Clinical considerations for patients with diabetes in times of covid-19 epidemic," *Diabetes Metabolic Syndrome*, vol. 14, p. 211, 5 2020. [Online]. Available: /pmc/articles/PMC7102582/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7102582/ [Cited on page 22.]

[99] N. R. G. et al., "A fuzzy approach to support decision-making in the triage process for suspected covid-19 patients in brazil," *Applied Soft Computing*, vol. 129, p. 109626, 11 2022. [Cited on pages 22, 28, 29, 30, 32, and 33.]

[100] G. Deschrijver and E. E. Kerre, "Triangular norms and related operators in l*-fuzzy set theory," *Logical, Algebraic, Analytic and Probabilistic Aspects of Triangular Norms*, pp. 231–259, 1 2005. [Cited on page 22.]

[101] Z. Tang, W. Zhao, X. Xie, Z. Zhong, F. Shi, J. Liu, and D. Shen, "Severity assessment of coronavirus disease 2019 (covid-19) using quantitative features from chest ct images," 3 2020. [Online]. Available: https://arxiv.org/abs/2003.11988v1 [Cited on page 22.]

[102] N. Mehta and S. Shukla, "Pandemic analytics: How countries are leveraging big data analytics and artificial intelligence to fight covid-19?" *Sn Computer Science*, vol. 3, 1 2022. [Online]. Available: /pmc/articles/PMC8577168//pmc/articles/PMC8577168/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8577168/ [Cited on page 22.]

[103] A. H. Shamman, A. A. Hadi, A. R. Ramul, M. M. A. Zahra, and H. M. Gheni, "The artificial intelligence (ai) role for tackling against covid-19 pandemic," *Materials Today: Proceedings*, vol. 80, pp. 3663–3667, 1 2023. [Cited on page 22.]

[104] M. van der Schaar, A. M. Alaa, A. Floto, A. Gimson, S. Scholtes, A. Wood, E. McKinney, D. Jarrett, P. Lio, and A. Ercole, "How artificial intelligence and machine learning can help healthcare systems respond to covid-19," *Machine Learning*, vol. 110, pp. 1–14, 1 2021. [Online]. Available: https://link.springer.com/article/10.1007/s10994-020-05928-x [Cited on page 22.]

[105] E. Maia, P. Vieira, and I. Praça, "Empowering preventive care with geca chatbot," *Healthcare 2023, Vol. 11, Page 2532*, vol. 11, p. 2532, 9 2023. [Online]. Available: https://www.mdpi.com/2227-9032/11/18/2532/htmhttps://www.mdpi.com/2227-9032/11/18/2532 [Cited on pages 23 and 94.]

[106] H. Lei, W. Lu, A. Ji, E. Bertram, P. Gao, X. Jiang, and A. Barman, "Covid-19 smart chatbot prototype for patient monitoring," 3 2021. [Online]. Available: https://arxiv.org/abs/2103.06816v2 [Cited on page 23.]

[107] S. Schmeelk, A. Davis, Q. Li, C. Shippey, M. Utah, A. Myers, M. R. Turchioe, and R. M. Creber, "Monitoring symptoms of covid-19: Review of mobile apps," *JMIR mHealth and uHealth*, vol. 10, 6 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35609313/ [Cited on page 23.]

[108] H. Yu, Y. Guo, Y. Xiang, C. S. I. Cyber-Systems, ..., and undefined 2020, "Data-driven discovery of a clinical route for severity detection of covid-19 paediatric cases," *Wiley Online Library*, vol. 2, pp. 205–206, 12 2020. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-csr.2020.0037 [Cited on page 23.]

[109] B. S. Yelure and et al., "Remote monitoring of covid-19 patients using iot and ai." Institute of Electrical and Electronics Engineers Inc., 2022, pp. 73–80. [Cited on pages 23 and 28.]

[110] H.-S. Kim, S. E. Lee, H. A. Kim, H. Kim, Y. Lee, Y. H. Choi, E. J. Kim, H. Hwang, H. A. Kim, E. J. Kim *et al.*, "An easy-to-use machine learning model to predict the prognosis of patients with covid-19: Retrospective cohort study," *Journal of Medical Internet Research*, vol. 25, no. 1, Nov 2020. [Online]. Available: http://dx.doi.org/10.2196/30192 [Cited on page 23.]

[111] "Google colaboratory," https://colab.research.google.com/. [Cited on page 24.]

[112] sofiamalpique gecad. Github profile. [Online]. Available: https://github.com/sofiamalpique-gecad [Cited on page 24.]

[113] J. P. Cohen and et al., "Covid-19 image data collection," 2020. [Cited on page 28.]

[114] M. Khan and et al., "Applications of artificial intelligence in covid-19 pandemic: A comprehensive review," *Expert Systems with Applications*, vol. 185, p. 115695, 12 2021. [Cited on page 28.]

[115] E. Mbunge, "Integrating emerging technologies into covid-19 contact tracing: Opportunities, challenges and pitfalls," *Elsevier*. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1871402120303325 [Cited on page 28.]

[116] W. J. Gordon and et al., "Remote patient monitoring program for hospital discharged covid-19 patients background and significance," *Appl Clin Inform*, vol. 11, pp. 792–801, 2020. [Online]. Available: https://doi.org/ [Cited on page 28.]

[117] E. L. Wallace and et al., "Remote patient management for home dialysis patients," *Kidney International Reports*, vol. 2, no. 6, pp. 1009–1017, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468024917303170 [Cited on page 28.]

[118] C.-. D. Portal, "Covid-19 data portal," https://www.covid19dataportal.org/. [Cited on page 28.]

[119] Johns Hopkins University, "Covid-19 map - johns hopkins coronavirus resource center," https://coronavirus.jhu.edu/map.html, 2023. [Cited on pages 28 and 29.]

[120] S. K. Dey and et al., "Analyzing the epidemiological outbreak of covid-19: A visual exploratory data analysis approach," *Journal of Medical Virology*, vol. 92, no. 6, pp. 632–638, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jmv.25743 [Cited on pages 28 and 29.]

[121] A. N. Poudel, S. Zhu, N. Cooper, P. Roderick, N. Alwan, C. Tarrant, N. Ziauddeen, and G. L. Yao, "Impact of covid-19 on health-related quality of life of patients: A structured review," *PLoS ONE*, vol. 16, 10 2021. [Online]. Available: /pmc/articles/PMC8553121//pmc/articles/PMC8553121/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8553121/ [Cited on page 29.]

[122] B. Xu and et al., "Epidemiological data from the covid-19 outbreak, real-time case information," *Scientific Data*, vol. 7, 12 2020. [Cited on pages 29, 30, and 33.]

[123] J. Kim, "Coronavirus dataset," https://www.kaggle.com/datasets/kimjihoo/coronavirusdataset, 2020. [Cited on pages 29, 30, 31, and 33.]

[124] O. Albitar and et al., "Risk factors for mortality among covid-19 patients," *Diabetes Research and Clinical Practice*, vol. 166, p. 108293, 8 2020. [Cited on page 30.]

[125] T. Alafif and et al., "On the prediction of isolation, release, and decease states for covid-19 patients: A case study in south korea," *ISA Transactions*, vol. 124, pp. 191–196, 5 2022. [Cited on page 30.]

[126] N. AL-Rousan and H. AL-Najjar, "Data analysis of coronavirus covid-19 epidemic in south korea based on recovered and death cases," *Journal of Medical Virology*, vol. 92, no. 9, pp. 1603–1608, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jmv.25850 [Cited on page 30.]

[127] DataRobot, "Predicting days to recovery of covid-19 patients," https://www.datarobot.com/blog/predicting-days-to-recovery-of-covid-19-patients/, 2020. [Cited on page 30.]

[128] "A brief history of vaccination," https://www.who.int/news-room/spotlight/history-of-vaccination/a-brief-history-of-vaccination. [Cited on page 39.]

[129] S. X. Zhang, F. A. Marioli, R. Gao, and S. Wang, "A second wave? what do people mean by covid waves? – a working definition of epidemic waves," *Risk Management and Healthcare Policy*, vol. 14, p. 3775, 2021. [Online]. Available: /pmc/articles/PMC8448159//pmc/articles/PMC8448159/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8448159/ [Cited on page 39.]

[130] S. M. Moghadas, T. N. Vilches, K. Zhang, C. R. Wells, A. Shoukat, B. H. Singer, L. A. Meyers, K. M. Neuzil, J. M. Langley, M. C. Fitzpatrick, and A. P. Galvani, "The impact of vaccination on covid-19 outbreaks in the united states," *medRxiv*, 11 2020. [Online]. Available: /pmc/articles/PMC7709178//pmc/articles/PMC7709178/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7709178/ [Cited on page 39.]

[131] X. Liu, J. Huang, C. Li, Y. Zhao, D. Wang, Z. Huang, and K. Yang, "The role of seasonality in the spread of covid-19 pandemic," *Environmental Research*, vol. 195, p. 110874, 4 2021. [Online]. Available: /pmc/articles/PMC7892320//pmc/articles/PMC7892320/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7892320/ [Cited on page 40.]

[132] A. K. Weaver, J. R. Head, C. F. Gould, E. J. Carlton, and J. V. Remais, "Environmental factors influencing covid-19 incidence and severity," *Annual review of public health*, vol. 43, p. 271, 4 2022. [Online]. Available: /pmc/articles/PMC10044492//pmc/articles/PMC10044492/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC10044492/ [Cited on page 40.]

[133] E. D. Belay, J. Abrams, M. E. Oster, J. E. Giovanni, T. Pierce, L. Meng, J. Fuld, S. J. Salyer, and S. Godfred-Cato, "Trends in geographic and temporal distribution of us children with multisystem inflammatory syndrome during the covid-19 pandemic," *JAMA Pediatrics*, 2021. [Online]. Available: https://jamanetwork.com/journals/jamapediatrics/fullarticle/2777862 [Cited on page 40.]

[134] V. Dhanasekaran, S. G. Sullivan, K. M. Edwards, R. Xie, A. Khvorov, S. A. Valkenburg, and I. G. Barr, "Potential short- and long-term evolutionary dynamics of seasonal influenza," *Evolutionary Applications*, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/eva.13262 [Cited on page 40.]

[135] A. Hamidian Jahromi, "Analysis of reported case fatality rate and characteristics of covid-19 patients in italy," *J Biomed Res Environ Sci*, 2020. [Online]. Available: http://dx.doi.org/10.37871/jels1111 [Cited on page 43.]

[136] Y. Alimohamadi, H. H. Tola, A. Abbasi-Ghahramanloo, M. Janani, and M. Sepandi, "Case fatality rate of covid-19: a systematic review and meta-analysis," *Journal of preventive medicine and hygiene*, vol. 62, pp. E311–E320, 7 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34604571/ [Cited on page 43.]

[137] Y. Cao, A. Hiyoshi, and S. Montgomery, "Original research: Covid-19 case-fatality rate and demographic and socioeconomic influencers: worldwide spatial regression

analysis based on country-level data," *BMJ Open*, vol. 10, p. 43560, 11 2020. [Online]. Available: /pmc/articles/PMC7640588//pmc/articles/PMC7640588/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7640588/ [Cited on page 44.]

[138] J. A. W. Gold, K. K. Wong, C. M. Szablewski, P. R. Patel, J. Rossow, J. da Silva, P. Natarajan, S. B. Morris, R. N. Fanfair, J. Rogers-Brown *et al.*, "Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in new york city: prospective cohort study," *BMJ*, vol. 369, 2020. [Online]. Available: https://www.bmj.com/content/369/bmj.m1966 [Cited on page 45.]

[139] M. Reilev, K. B. Kristensen, A. Pottegård, L. C. Lund, J. Hallas, M. T. Ernst, C. F. Christiansen, H. T. Sørensen, N. B. Johansen, N. C. Brun *et al.*, "Characteristics and predictors of hospitalization and death in the first 11 122 cases with a positive rt-pcr test for sars-cov-2 in denmark: a nationwide cohort," *International Journal of Epidemiology*, vol. 49, no. 5, pp. 1468–1481, 2020. [Online]. Available: https://academic.oup.com/ije/article/49/5/1468/5893593 [Cited on page 45.]

[140] A. K. Clift, C. A. Coupland, R. H. Keogh, K. Diaz-Ordaz, E. Williamson, E. M. Harrison, A. Hayward, H. Hemingway, P. Horby, N. Mehta *et al.*, "Living risk prediction algorithm (qcovid) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study," *BMJ*, vol. 371, 2020. [Online]. Available: https://www.bmj.com/content/371/bmj.m3731 [Cited on page 45.]

[141] V. D. Teich, S. Klajner, F. A. S. de Almeida, A. C. B. Dantas, C. R. Laselva, M. G. Torritesi, T. R. Canero, O. Berwanger, L. V. Rizzo, E. P. Reis, and M. C. Neto, "Epidemiologic and clinical features of patients with covid-19 in brazil," *Einstein (Sao Paulo, Brazil)*, vol. 18, p. eAO6022, 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32813760/ [Cited on page 45.]

[142] I. Lakbar, D. Luque-Paz, J. L. Mege, S. Einav, and M. Leone, "Covid-19 gender susceptibility and outcomes: A systematic review," *PLoS ONE*, vol. 15, 11 2020. [Online]. Available: /pmc/articles/PMC7608911//pmc/articles/PMC7608911/?report=

abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7608911/ [Cited on pages 46 and 54.]

[143] N. B. Anderson, R. A. Bulatao, B. Cohen, Ethnicity, and H. in Later Life National Research Council (US) Panel on Race, "Genetic factors in ethnic disparities in health," 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK25517/ [Cited on page 46.]

[144] K. Shao and H. Feng, "Racial and ethnic healthcare disparities in skin cancer in the united states: A review of existing inequities, contributing factors, and potential solutions," *The Journal of Clinical and Aesthetic Dermatology*, vol. 15, p. 16, 7 2022. [Online]. Available: /pmc/articles/PMC9345197//pmc/articles/PMC9345197/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9345197/ [Cited on page 46.]

[145] M. Almagro, J. Coven, A. Gupta, and A. Orane-Hutchinson, "Racial disparities in frontline workers and housing crowding during covid-19: Evidence from geolocation data," 2020. [Online]. Available: https://www1.nyc.gov/site/doh/covid/covid-19-data. [Cited on page 47.]

[146] Y. Alimohamadi and et al., "Determine the most common clinical symptoms in covid-19 patients: a systematic review and meta-analysis," *Journal of Preventive Medicine and Hygiene*, vol. 61, p. E304, 10 2020. [Online]. Available: /pmc/articles/PMC7595075//pmc/articles/PMC7595075/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7595075/ [Cited on page 48.]

[147] B. Nogrady, "What the data say about asymptomatic covid infections," *Nature*, vol. 587, pp. 534–535, 11 2020. [Cited on page 49.]

[148] E. M. El-Ghitany, M. H. Hashish, A. G. Farghaly, E. A. Omran, N. A. Osman, and M. M. Fekry, "Asymptomatic versus symptomatic sars-cov-2 infection: a cross-sectional seroprevalence study," *Tropical Medicine and Health*, vol. 50, pp. 1–12, 12 2022. [Online]. Available: https://tropmedhealth.biomedcentral.com/articles/10.1186/s41182-022-00490-9 [Cited on page 49.]

[149] A. Sanyaolu and et al., "Comorbidity and its impact on patients with covid-19," *SN Comprehensive Clinical Medicine*, vol. 2, pp. 1069–1076, 8 2020. [Online]. Available: https://link.springer.com/article/10.1007/s42399-020-00363-4 [Cited on page 50.]

[150] J. B. Halter, N. Musi, F. M. F. Horne, J. P. Crandall, A. Goldberg, L. Harkless, W. R. Hazzard, E. S. Huang, M. S. Kirkman, J. Plutzky, K. E. Schmader, S. Zieman, and K. P. High, "Diabetes and cardiovascular disease in older adults: Current status and future directions," *Diabetes*, vol. 63, p. 2578, 2014. [Online]. Available: /pmc/articles/PMC4113072//pmc/articles/PMC4113072/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4113072/ [Cited on page 50.]

[151] V. Taneja, "Sex hormones determine immune response," *Frontiers in Immunology*, vol. 9, p. 386034, 8 2018. [Cited on page 54.]

[152] M. Dai, L. Tao, Z. Chen, Z. Tian, X. Guo, D. S. Allen-Gipson, R. Tan, R. Li, L. Chai, F. Ai, and M. Liu, "Influence of cigarettes and alcohol on the severity and death of covid-19: A multicenter retrospective study in wuhan, china," *Frontiers in Physiology*, vol. 11, p. 588553, 12 2020. [Online]. Available: /pmc/articles/PMC7756110//pmc/articles/PMC7756110/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7756110/ [Cited on page 54.]

[153] A. A. Schäfer, L. P. Santos, M. R. Quadra, S. C. Dumith, and F. O. Meller, "Alcohol consumption and smoking during covid-19 pandemic: Association with sociodemographic, behavioral, and mental health characteristics," *Journal of Community Health*, vol. 47, p. 588, 8 2022. [Online]. Available: /pmc/articles/PMC8951656//pmc/articles/PMC8951656/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8951656/ [Cited on page 54.]

[154] D. Azzolina, R. Comoretto, C. Lanera, P. Berchialla, I. Baldi, and D. Gregori, "Covid-19 hospitalizations and patients' age at admission: The neglected importance of data variability for containment policies," *Frontiers in Public Health*, vol. 10, p. 1002232, 11 2022. [Cited on page 54.]

[155] J. Casas-Rojo, J. Antón-Santos, J. Millán-Núñez-Cortés, C. Lumbreras-Bermejo, J. Ramos-Rincón, E. Roy-Vallejo, A. Artero-Mora, F. Arnalich-Fernández, J. García-Bruñén, J. Vargas-Núñez, S. Freire-Castro, L. Manzano-Espinosa, I. Perales-Fraile, A. Crestelo-Viéitez, F. Puchades-Gimeno, E. Rodilla-Sala, M. Solís-Marquínez, D. Bonet-Tur, M. Fidalgo-Moreno, E. Fonseca-Aizpuru, F. Carrasco-Sánchez, E. Rabadán-Pejenaute, M. Rubio-Rivas, J. Torres-Peña, and R. Gómez-Huelgas, "Clinical characteristics of patients hospitalized with covid-19 in spain: results from the semi-covid-19 registry," *Revista Clínica Española (English Edition)*, vol. 220, pp. 480–494, 11 2020. [Cited on page 54.]

[156] R. Meys, J. M. Delbressine, Y. M. Goërtz, A. W. Vaes, F. V. Machado, M. V. Herck, C. Burtin, R. Posthuma, B. Spaetgens, F. M. Franssen, Y. Spies, H. Vijlbrief, A. J. V. Hul, D. J. Janssen, M. A. Spruit, and S. Houben-Wilke, "Generic and respiratory-specific quality of life in non-hospitalized patients with covid-19," *Journal of Clinical Medicine 2020, Vol. 9, Page 3993*, vol. 9, p. 3993, 12 2020. [Online]. Available: https://www.mdpi.com/2077-0383/9/12/3993/htmhttps://www.mdpi.com/2077-0383/9/12/3993 [Cited on page 54.]

[157] J. Fabião, B. Sassi, E. F. Pedrollo, F. Gerchman, C. K. Kramer, C. B. Leitão, and L. C. Pinto, "Why do men have worse covid-19-related outcomes? a systematic review and meta-analysis with sex adjusted for age," *Brazilian Journal of Medical and Biological Research*, vol. 55, 2022. [Online]. Available: /pmc/articles/PMC8856598//pmc/articles/PMC8856598/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8856598/ [Cited on page 55.]

[158] N. T. Nguyen, J. Chinn, M. de Ferrante, K. A. Kirby, S. F. Hohmann, and A. Amin, "Male gender is a predictor of higher mortality in hospitalized adults with covid-19," *PLOS ONE*, vol. 16, p. e0254066, 7 2021. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0254066 [Cited on page 55.]

[159] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, pp. 129–137, 1982. [Cited on page 63.]

[160] D. Marutho, S. H. Handaka, E. Wijaya, and Muljono, "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news,"

Proceedings - 2018 International Seminar on Application for Technology of Information and Communication: Creative Technology for Human Life, iSemantic 2018, pp. 533–538, 11 2018. [Cited on page 63.]

[161] A. Elias, X. Vinicius, and L. Xavier, "Solving the minimum sum-of-squares clustering problem by hyperbolic smoothing and partition into boundary and gravitational regions." [Cited on page 63.]

[162] L. D. F. Costa, "Further generalizations of the jaccard index," 10 2021. [Online]. Available: https://arxiv.org/abs/2110.09619v3 [Cited on page 63.]

[163] A. K. Singh, R. Gupta, A. Ghosh, and A. Misra, "Diabetes in covid-19: Prevalence, pathophysiology, prognosis and practical considerations," *Diabetes Metabolic Syndrome*, vol. 14, p. 303, 7 2020. [Online]. Available: /pmc/articles/PMC7195120//pmc/articles/PMC7195120/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7195120/ [Cited on page 67.]

[164] Z. J. Gazzaz, "Diabetes and covid-19," *Open Life Sciences*, vol. 16, p. 297, 1 2021. [Online]. Available: /pmc/articles/PMC8010370//pmc/articles/PMC8010370/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8010370/ [Cited on pages 67 and 73.]

[165] P. Ssentongo, A. E. Ssentongo, E. S. Heilbrunn, D. M. Ba, and V. M. Chinchilli, "Association of cardiovascular disease and 10 other pre-existing comorbidities with covid-19 mortality: A systematic review and meta-analysis," *PLOS ONE*, vol. 15, p. e0238215, 8 2020. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0238215 [Cited on page 68.]

[166] L. D'Marco, M. J. Puchades, M. Romero-Parra, E. Gimenez-Civera, M. J. Soler, A. Ortiz, and J. L. Gorriz, "Coronavirus disease 2019 in chronic kidney disease," *Clinical Kidney Journal*, vol. 13, pp. 297–306, 6 2020. [Online]. Available: https://dx.doi.org/10.1093/ckj/sfaa104 [Cited on pages 68 and 73.]

[167] J. van Son, S. M. Oussaada, A. Şekercan, M. Beudel, D. A. Dongelmans, S. van Assen, I. A. Eland, H. S. Moeniralam, T. P. Dormans, C. A. van Kalkeren, R. A. Douma, D. Rusch, S. Simsek, L. Liu, R. S. Kootte, C. E.

Wyers, R. G. IJzerman, J. P. van den Bergh, C. D. Stehouwer, M. Nieuwdorp, K. W. ter Horst, and M. J. Serlie, "Overweight and obesity are associated with acute kidney injury and acute respiratory distress syndrome, but not with increased mortality in hospitalized covid-19 patients: A retrospective cohort study," *Frontiers in Endocrinology*, vol. 12, p. 1, 12 2021. [Online]. Available: /pmc/articles/PMC8713548//pmc/articles/PMC8713548/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8713548/ [Cited on page 68.]

[168] C. Kovesdy, S. Furth, C. Zoccali, P. K. T. Li, G. Garcia-Garcia, M. Benghanem-Gharbi, R. Bollaert, S. Dupuis, T. Erk, K. Kalantar-Zadeh, C. Osafo, M. C. Riella, and E. Zakharova, "Obesity and kidney disease: Hidden consequences of the epidemic," *Indian Journal of Nephrology*, vol. 27, p. 85, 3 2017. [Online]. Available: /pmc/articles/PMC5358165/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5358165/ [Cited on page 68.]

[169] J. Khateeb, E. Fuchs, and M. Khamaisi, "Diabetes and lung disease: An underestimated relationship," *The Review of Diabetic Studies : RDS*, vol. 15, p. 1, 2019. [Online]. Available: /pmc/articles/PMC6760893//pmc/articles/PMC6760893/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6760893/ [Cited on page 68.]

[170] F. A. Al-Muhanna, W. I. A. Albakr, A. V. Subbarayalu, C. Cyrus, H. A. Aljenaidi, L. A. Alayoobi, and O. Al-Muhanna, "Impact of covid-19 on kidney of diabetic patients," *Medicina*, vol. 58, 5 2022. [Online]. Available: /pmc/articles/PMC9143731//pmc/articles/PMC9143731/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9143731/ [Cited on page 68.]

[171] E. Liu, H. Lee, B. Lui, R. S. White, and J. D. Samuels, "Respiratory and nonrespiratory covid-19 complications in patients with obesity: recent developments," *Journal of Comparative Effectiveness Research*, vol. 11, pp. 371–381, 4 2021. [Online]. Available: /pmc/articles/PMC8757534//pmc/articles/PMC8757534/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8757534/ [Cited on page 69.]

[172] Y. Statsenko, F. A. Zahmi, T. Habuza, T. M. Almansoori, D. Smetanina, G. L. Simiyu, K. N.-V. Gorkom, M. Ljubisavljevic, R. Awawdeh, H. Elshekhali, M. Lee, N. Salamin, R. Sajid, D. Kiran, S. Nihalani, T. Loney, A. Bedson, A. Dehdashtian, and J. A. Koteesh, "Impact of age and sex on covid-19 severity assessed from radiologic and clinical findings," *Frontiers in Cellular and Infection Microbiology*, vol. 11, p. 777070, 2 2022. [Cited on page 72.]

[173] M. Farshbafnadi, S. K. Zonouzi, M. Sabahi, M. Dolatshahi, and M. H. Aarabi, "Aging covid-19 susceptibility, disease severity, and clinical outcomes: The role of entangled risk factors," *Experimental gerontology*, vol. 154, 10 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34352287/ [Cited on page 73.]

[174] K. R. Starke, D. Reissig, G. Petereit-Haack, S. Schmauder, A. Nienhaus, and A. Seidler, "The isolated effect of age on the risk of covid-19 severe outcomes: a systematic review with meta-analysis," *BMJ Global Health*, vol. 6, p. 6434, 12 2021. [Online]. Available: /pmc/articles/PMC8678541//pmc/articles/PMC8678541/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8678541/ [Cited on page 72.]

[175] P. G. Gibson, L. Qin, and S. H. Puah, "Covid-19 acute respiratory distress syndrome (ards): clinical features and differences from typical pre-covid-19 ards," *The Medical Journal of Australia*, vol. 213, p. 54, 7 2020. [Online]. Available: /pmc/articles/PMC7361309/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7361309/ [Cited on page 73.]

[176] A. Aslan, C. Aslan, N. M. Zolbanin, and R. Jafari, "Acute respiratory distress syndrome in covid-19: possible mechanisms and therapeutic management," *Pneumonia 2021 13:1*, vol. 13, pp. 1–15, 12 2021. [Online]. Available: https://pneumonia.biomedcentral.com/articles/10.1186/s41479-021-00092-9

[177] L. Gattinoni, D. Chiumello, and S. Rossi, "Covid-19 pneumonia: Ards or not?" *Critical Care*, vol. 24, pp. 1–3, 4 2020. [Online]. Available: https://ccforum.biomedcentral.com/articles/10.1186/s13054-020-02880-zhttp://creativecommons.org/publicdomain/zero/1.0/ [Cited on page 73.]

[178] "Symptoms of covid-19," https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html. [Cited on pages 73 and 74.]

[179] R. M. Pereira, D. Bertolini, L. O. Teixeira, C. N. Silla, and Y. M. Costa, "Covid-19 identification in chest x-ray images on flat and hierarchical classification scenarios," *Computer Methods and Programs in Biomedicine*, vol. 194, p. 105532, 10 2020. [Online]. Available: /pmc/articles/PMC7207172//pmc/articles/PMC7207172/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7207172/ [Cited on page 75.]

[180] K. Cassell, C. M. Zipfel, S. Bansal, and D. M. Weinberger, "Trends in non-covid-19 hospitalizations prior to and during the covid-19 pandemic period, united states, 2017–2021," *Nature Communications*, vol. 13, 12 2022. [Online]. Available: /pmc/articles/PMC9546751//pmc/articles/PMC9546751/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9546751/ [Cited on page 75.]

[181] S. Tian, W. Yang, J. M. L. Grange, P. Wang, W. Huang, and Z. Ye, "Smart healthcare: making medical care more intelligent," *Global Health Journal*, vol. 3, pp. 62–65, 9 2019. [Cited on page 75.]

[182] M. Kim and K. B. Hwang, "An empirical evaluation of sampling methods for the classification of imbalanced data," *PLoS ONE*, vol. 17, 7 2022. [Online]. Available: /pmc/articles/PMC9333262//pmc/articles/PMC9333262/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9333262/ [Cited on page 75.]

[183] M. Bach, A. Werner, and M. Palt, "The proposal of undersampling method for learning from imbalanced datasets," *Procedia Computer Science*, vol. 159, pp. 125–134, 1 2019. [Cited on page 75.]

[184] A. Estabrooks and N. Japkowicz, "A mixture-of-experts framework for learning from imbalanced data sets," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2189, pp. 34–43, 2001. [Online]. Available: https://link.springer.com/chapter/10.1007/3-540-44816-0_4 [Cited on page 75.]

[185] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance," *Information Sciences*, vol. 505, pp. 32–64, 12 2019. [Cited on page 76.]

[186] E. Mbunge, M. N. Sibiya, S. Takavarasha, R. C. Millham, G. Chemhaka, B. Muchemwa, and T. Dzinamarira, "Implementation of ensemble machine learning classifiers to predict diarrhoea with smoteenn, smote, and smotetomek class imbalance approaches," *2023 Conference on Information Communications Technology and Society, ICTAS 2023 - Proceedings*, 2023. [Cited on page 76.]

[187] J. I. Myung, D. R. Cavagnaro, and M. A. Pitt, "A tutorial on adaptive design optimization," *Journal of Mathematical Psychology*, vol. 57, pp. 53–67, 2013. [Cited on page 78.]

[188] J. Wu, X. Y. Chen, H. Zhang, L. D. Xiong, H. Lei, and S. H. Deng, "Hyperparameter optimization for machine learning models based on bayesian optimization," *Journal of Electronic Science and Technology*, vol. 17, pp. 26–40, 3 2019. [Cited on page 78.]

[189] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," 2022. [Online]. Available: https://github.com/ [Cited on page 78.]

[190] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," *Encyclopedia of Database Systems*, pp. 532–538, 2009. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9_565 [Cited on page 80.]

[191] J. Pachouly, S. Ahirrao, K. Kotecha, G. Selvachandran, and A. Abraham, "A systematic literature review on software defect prediction using artificial intelligence: Datasets, data validation methods, approaches, and tools," *Engineering Applications of Artificial Intelligence*, vol. 111, 5 2022. [Cited on page 80.]

[192] A. Gumilar, S. S. Prasetiyowati, and Y. Sibaroni, "Performance analysis of hybrid machine learning methods on imbalanced data (rainfall classification)," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, pp. 481–490, 7 2022. [Cited on page 89.]