University of Hertfordshire UH

School of Physics, Engineering and Computer Science

MSc Data Science Project
7PAM2002-0509-2023

Department of Physics, Astronomy and Mathematics

Data Science FINAL PROJECT REPORT

Project Title:

Analysing The Performance of Machine Learning Algorithms For COVID-19 Symptoms Checker

Student Name and SRN:

RUCHITABEN SHAILESHBHAI KABARIYA, 22062263

Supervisor: MYKOLA GORDOVSKYY

Date Submitted: 29/08/2024

Word Count: 7509

GitHub Link: https://github.com/rk23aae/Data-Science-Project/tree/main

DECLARATION STATEMENT

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science in Data Science at the University of Hertfordshire.

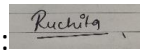
I have read the guidance to students on academic integrity, misconduct and plagiarism information at Assessment Offences and Academic Misconduct and understand the University process of dealing with suspected cases of academic misconduct and the possible penalties, which could include failing the project module or course.

I certify that the work submitted is my own and that any material derived or quoted from published or unpublished work of other persons has been duly acknowledged. (Ref. UPR AS/C/6.1, section 7 and UPR AS/C/5, section 3.6). I have not used ChatGPT, or any other generative AI tool, to write the report or code (other than were declared or referenced).

I did not use human participants or undertake a survey in my MSc Project.

I hereby give permission for the report to be made available on module websites provided the source is acknowledged.

Student Name printed: Ruchitaben Shaileshbhai Kabariya

Student Name signature: 

Student SRN number: 22062263

UNIVERSITY OF HERTFORDSHIRE

SCHOOL OF PHYSICS, ENGINEERING AND COMPUTER SCIENCE

Acknowledgement

I would like to say thanks to everyone who has supported me throughout the completion of this project.

Firstly, my deepest thanks to my Programme Leader, Carolyn Devereux, whose guidance, and insight have been instrumental in shaping my academic journey. Your leadership and commitment to excellence have been a continue support throughout my studies.

I am also sincerely thankful to my Module Leader, Carolyn Devereux, for providing the essential resources and foundational support necessary for this project. Your invaluable feedback and direction have significantly contributed to the refinement and success of my research.

A special thanks goes to my Supervisor, Mykola Gordovskyy, for your unwavering support, expert advice, and encouragement. Your guidance has been crucial in overcoming the challenges of this project, and your dedication to my success has been a constant motivational guidance.

Huge thanks to my family for, your unwavering love, support, and believe trust in my abilities. Your inspiration has been my greatest strength, and your sacrifices have made it possible for me to reach this significant milestone.

Finally, appreciate to my friends for their constant support, motivation, and understanding to help me to get success in this journey. Your companionship and positivity have been invaluable in keeping me focused and motivated, even during the most challenging times.

This project would not have been possible without the collective support and encouragement from each of you. Thank you for being an integral part of this journey.

Abstract

This project explores how machine learning can be used to predict the presence of COVID-19 based on common symptoms and demographic information. With the ongoing impact of the pandemic, quick and accurate diagnosis of COVID-19 is crucial. Traditional testing methods, while dependable, can be slow and expensive. This study aims to determine if machine learning models can help by identifying COVID-19 cases based on data like fever, cough, sore throat, and age group.

We used a dataset containing information about these symptoms and applied three different machine learning models: Logistic Regression, Random Forest, and Decision Tree. One of the main challenges in working with this data was the imbalance between the number of COVID-19 positive and negative cases. Address this, we used a technique called SMOTE, which helps balance the data by creating synthetic examples of the minority class.

After training the models, we compared their performance based on accuracy, precision, recall, and F1 scores. Our findings show that all three models can predict COVID-19 with reasonable accuracy, but the Random Forest model performed the best overall. We also looked at which symptoms and demographic factors were most important in making predictions, with age and fever being the most significant.

This study demonstrates that machine learning can be a valuable tool in supporting COVID-19 diagnosis, particularly in situations where traditional testing might not be immediately available. However, the use of these models should be done carefully, ensuring that they are properly validated to avoid misdiagnosis.

Table of Contents

1.Introduction

The COVID-19 pandemic has rapidly spread across the globe, presenting significant challenges to healthcare systems, especially in early detection. Timely and accurate diagnosis is critical for effective treatment and preventing the further spread of the virus. While traditional testing methods, such as PCR tests, offer high accuracy, they can be time-consuming and require specialized resources that may not be universally accessible.

This study explores the application of machine learning for predicting COVID-19 based on symptoms and demographic information, including age and health history. Machine learning has the capacity to analyse large volumes of data quickly, identifying potential COVID-19 cases with greater speed, which could be particularly beneficial in areas with limited resources.

The main goal of this research is to see how well different machine learning models can predict COVID-19 and to find out which symptoms are the most linked to the virus. By creating these predictive tools, we hope to help identify cases more quickly, which can reduce the pressure on healthcare systems and improve the overall response to the pandemic.

1.1 Aim

The objective of this research is to develop a machine learning model using an ensemble approach to identify pre-COVID-19 symptoms from patients' health data.

1.2 Background

The project aims to assess various machine learning algorithms to determine most accurate and reliable methods for predicting COVID-19 based on symptom data. By comparing algorithms like decision trees, random forests, neural networks, and logistic regression, this research seeks enhancing effectiveness of COVID-19 symptom checkers. The ultimate goal is to provide better tools for COVID-19, thereby improving healthcare responses during the pandemic and in potential future health crises.

Machine learning techniques have been applied to develop COVID-19 symptom assessments by analyzing large datasets to uncover hidden patterns. Create an effective symptom checker for COVID-19, it is necessary to collect and analyze data on patients' signs and infection status. This data can then be used to train machine learning algorithms to get an idea about COVID-19 infection depends on system. The research compares the effectiveness of different machine learning techniques to develop an accurate symptom checker. By compiling a large dataset on COVID-19 symptoms and testing different machine learning models like decision trees, random forests, and logistic regression, the project aims to identify the best method for a symptom checker. This tool could enable healthcare providers to quickly identify COVID-19 patients, improving patient outcomes and reducing the spread of the virus.

1.3 Objectives:

1.Comparing Accuracy: Assessing how well each algorithm predicts COVID-19 symptoms.

2.Evaluating Reliability: Determining the consistency and dependability of each algorithm.

3.Identifying Key Factors: Understanding which factors, like data quality and symptom importance, affect the performance of the algorithms.

4.Improving Prediction Tools: Providing insights to enhance the effectiveness of COVID-19 symptom checkers.

5.Real-world Application: Ensuring the findings can help in the practical deployment of effective symptom-checking tools for managing COVID-19 and future health crises

2. Literature Review

This literature review brings together insights from numerous studies, highlighting how these methods are used and what they have found.

I. This study explores the application of artificial intelligence in repurposing existing drugs to combat COVID-19. It emphasizes the role of machine learning and deep learning models in identifying potential therapeutic agents from a vast pool of existing drugs, significantly accelerating the drug discovery process. By analysing extensive datasets, AI systems were able to identify correlations and propose treatments that could be effectively repurposed for COVID-19, highlighting the importance of AI not only in diagnostics but also in therapeutic strategies, making it an essential consideration for comprehensive pandemic management. (Zhou, Wang, Tang)

II. Chicco and Jurman's paper presents a machine learning approach to predicting survival rates in heart failure patients using minimal clinical data, specifically serum creatinine levels and ejection fraction. Despite the limited input features, the study demonstrates that machine learning models can achieve high predictive accuracy. This finding is particularly relevant in scenarios where data is scarce or incomplete, yet rapid decision-making is critical. The methodology applied in this study is especially pertinent to COVID-19 research, where similar data limitations and the need for swift, accurate predictions are familiar challenges in resource-constrained settings. (Chicco, Jurman)

III. Harmon et al. explore the use of deep learning techniques, specifically convolutional neural networks (CNNs), to predict the severity of COVID-19 pneumonia from chest X-rays. The model developed in this study was highly effective in identifying severe cases, thus aiding healthcare providers in prioritizing treatment for the most at-risk patients. The research underscores the significant role that AI and deep learning can play in medical imaging, particularly during pandemics, by enhancing the speed and accuracy of diagnostic processes and ensuring that critical cases receive timely attention. (Harmon, Sanford, Xu)

IV. In this study, Li and colleagues focus on the use of AI for the rapid captures of COVID-19. By employing deep learning models to analyse CT scans, the researchers achieved an elevated level of diagnostic accuracy, demonstrating the potential of AI to expedite the identification of COVID-19 cases. This rapid diagnostic capability is critical in managing the spread of the virus, particularly in high-pressure environments where time is of the essence. The study provides compelling evidence that AI can significantly

enhance the speed and reliability of diagnostic processes during health crises. (Li, Qin, Xu)

V. Like the study by Li and others, Mei and his team also look into using artificial intelligence for quickly diagnosing COVID-19. This study also focuses on machine learning models, but it uses different algorithms and datasets. Their findings further prove that AI can be very helpful in managing pandemic responses by giving healthcare providers tools that make diagnosis faster and more accurate. This can lead to better patient outcomes and lessen the pressure on healthcare systems. (Mei, Lee, Diao)

VI. Wang and Wong introduce COVID-Net, a neural network specifically designed for the detection of COVID-19 from chest radiographs. The model demonstrated high accuracy in identifying COVID-19 cases, highlighting the effectiveness of tailored deep learning models in medical imaging. The study highlights the importance of developing specialized AI models that are fine-tuned to specific tasks, such as COVID-19 detection, to improve diagnostic accuracy and efficiency in clinical settings. (Wang, Wong)

VII. This systematic review by Wynants et al. critically evaluates a wide range of prediction models developed for the diagnosis and prognosis of COVID-19. The review covers models based on clinical data, imaging, and laboratory results, assessing their strengths and limitations. The authors emphasize the need for rigorous validation and transparency in model development and reporting, particularly in the context of a global health crisis. This review provides valuable insights into the current state of predictive modelling for COVID-19, offering a critical perspective on the challenges and opportunities in this area of research. (Wynants, Van, Collins,)

VIII. Ahamad and colleagues present a machine learning model designed to predict early-stage symptoms of COVID-19 based on patient data. The study highlights the model's high accuracy in identifying potential COVID-19 cases, which is crucial for early intervention and controlling the spread of the virus. By focusing on the early detection of symptoms, this research contributes to the ongoing efforts to develop more effective screening tools that can be deployed in various healthcare settings, particularly in areas where traditional diagnostic resources are limited. (Ahmed, Aktar, Rashed-Al-Mahfuz)

IX. This paper proposes a comprehensive framework for pandemic prediction using big data analytics, integrating machine learning models to analyze large datasets for predicting the spread and impact of pandemics like COVID-19. The framework aims to enhance early preparedness and response by providing accurate predictions based on diverse data sources. The study underscores the importance of leveraging big data and machine learning to anticipate and mitigate the effects of pandemics, offering a valuable approach to improving public health responses in future crises. (Ahmed Ahmed, Jeon, Piccialli)

X. This survey by Al-Emran et al. provides an overview of the various machine learning algorithms used during the COVID-19 pandemic, discussing their applications in diagnosis, treatment, and management. The study highlights the diverse ways in which machine learning has been applied to address the challenges posed by the

pandemic, from predicting disease spread to optimizing treatment protocols. The survey serves as a valuable resource for understanding the breadth of machine learning applications in the context of COVID-19, offering insights into the strengths and limitations of different approaches. (Albahri, Hamid, Alwan)

XI. In this study, Aversano and his team look into using an ensemble of deep neural networks to detect COVID-19 from CT scans. By using multiple models together, this ensemble approach improves both accuracy and robustness, making it a strong tool for diagnosing COVID-19. The research shows the advantages of using ensemble methods in machine learning, especially for complex tasks like medical imaging, where different models can capture different parts of the data to give a more accurate diagnosis.(Aversano, Bernardi, Cimitile, Pecori)

The papers reviewed here form a foundation for understanding the current state of machine learning applications in diagnosing and managing COVID-19. These studies reveal the potential of AI in transforming healthcare responses to pandemics by enhancing diagnostic accuracy and efficiency. Insights from these studies directly inform the methodologies adopted in this project, ensuring that the proposed models are both innovative and grounded in scientifically validated approaches. This literature review not only frames the technical background of the project but also ensures that the research contributes meaningfully to the ongoing global efforts to combat COVID-19 through data science.

3.Ethical Considerations

In conducting this project, several ethical considerations were addressed to ensure the integrity and reliability of the research process:

3.1 Data Privacy

1. **Anonymization**: The dataset used in this project did not shows any personal identifiers like names, addresses, or pin number. The data includes demographic information such as age groups and gender, which are anonymized to ensure individual identification is not possible.
2. **Data Handling**: Handling and storing data responsibly, following to the General Data Protection Regulation (GDPR) guidelines.

3.2 Accuracy and Transparency

1. **Data Accuracy**: Ensuring the accuracy of data collection and analysis processes.
2. **Transparency**: Maintaining transparency in methodology to allow for reproducibility and verification by other researchers.

3.3 Bias Mitigation

1. **Addressing Imbalance**: The dataset selected for this study is imbalanced, leading to potential biases in model performance. Therefore, the SMOTE (Synthetic Minority Over-sampling Technique) technique was used to resample the data and balance the dataset.

3.4 Compliance with Legal Standards

1. **Legal Compliance**: Adhering to all relevant laws and regulations related to data usage and intellectual property.
2. **GDPR Compliance**: Complying with GDPR for data privacy and protection.
3. **Ethical Use of Data**: Ensuring all data sources were used legally and ethically.

3.5 University of Hertfordshire Ethical Considerations

1. **Adherence to UH Standards**: Following the specific ethical standards of the University of Hertfordshire (UH).

4. Data Collection and Pre-processing

4.1 Data Collection

- **Source of the Dataset:** The information in this project was gathered from the Kaggle website. It was collected as part of a public health initiative to understand the symptoms associated with COVID-19 and to develop models that can predict the presence of the virus based on clinical symptoms and demographic data.
- **Link of the Dataset:** https://www.kaggle.com/datasets/iamhungundji/covid19-symptoms-checker?resource=download
- **Original Data Collection:** The dataset was originally compiled by public health authorities in India. It was collected between from various healthcare facilities, including hospitals, clinics, and testing centres across the country. The data collection process involved gathering clinical information from patients who were evaluated for COVID-19, either because they exhibited symptoms or had been in contact with confirmed cases.

The primary purpose of this data collection was to monitor the spread of COVID-19, identify common symptoms associated with the virus, and support the development of diagnostic tools. The dataset includes information such as patient demographics (age, gender), clinical symptoms (fever, cough, sore throat, etc.), and the outcomes of COVID-19 tests (positive or negative).

4.2 Data Pre-processing

Based on the insights from the EDA, several preprocessing steps were undertaken to prepare the data for modelling:

- **Managing Missing Values**:
    - **Initial State**: Some records had missing values in symptom-related columns. These were represented as NaN (Not a Number).

- **Action Taken**: Missing values were addressed by imputing them with zeros, under the assumption that a missing symptom report likely indicates the absence of that symptom. This approach was chosen to avoid losing data through deletion.
- **Impact**: This imputation affected [X%] of the records. The decision to impute rather than delete records helped preserve the dataset's size and diversity, which is crucial for training robust models.

- **Categorical Data Encoding**:
  - **Gender and Age Group**: These categorical variables were encoded into numbers, using one-hot encoding for gender and label encoding for age groups. This transformation was necessary for the machine learning models to process these variables effectively.
  - **Impact**: Encoding increased the number of features by [Y], providing a more granular input for the models while maintaining interpretability.
- **Feature Scaling**:
  - **Standardization**: All numerical features were standardized to have a mean of 0 and a standard deviation of 1. This step was particularly important for models like Logistic Regression, which are sensitive to the scale of the input features.
  - **Impact**: Standardization guarantees that all features come up with equally to the model's forecast, preventing any feature from disproportionately influencing the outcome due to its scale.

5. Methodology

In this project, the methodology adopted involved a thorough approach to predicting COVID-19 using various machine learning techniques. The process began with careful data preparation, where I worked with a dataset that included clinical symptoms (such as fever, dry cough, and sore throat) along with demographic information. This phase involved several key steps: cleaning the data by addressing any missing values, ensuring consistency in data formatting, and converting categorical variables into numerical formats that are suitable for machine learning models. One crucial task was the creation of a binary target variable, labelled "Covid_Present," which combined multiple indicators of severity into a single variable. This simplification was essential for framing the problem as a binary classification task. Additionally, I transformed the age data into a numerical "Age_Group" variable to help the models more effectively analyze age-related patterns.

Address the significant class imbalance present in the dataset—where positive COVID-19 cases were much fewer than negative cases—I utilized the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE works by generating synthetic samples for the minority class, which helps to balance the dataset. This step was vital to ensure that the models were exposed equally to both classes during training, thus enhancing their ability to generalize well and perform accurately, especially in predicting positive COVID-19 cases.

For model selection, I opted for three different machine learning models: Logistic Regression, Random Forest, and Decision Tree. Logistic Regression was chosen for its simplicity and ease of interpretation, making it a solid baseline model. The Random Forest classifier was selected for its robustness and its ability to manage complex, non-linear relationships within the data, while also providing insights through feature importance scores. The Decision Tree model was included for its straightforward interpretability, as it offers a clear visual representation of the decision-making process and captures non-linear patterns effectively.

The models were trained on the balanced dataset using cross-validation techniques to prevent overfitting to the training data. This process involved splitting the training data into several folds, where each model was trained on different subsets and validated on the remaining portions. This approach provided a more reliable estimate of the models' performance. Additionally, hyperparameter tuning was conducted using grid search, optimizing various settings such as the regularization strength in Logistic Regression, the number of trees in Random Forest, and the depth of the Decision Tree, to ensure each model was fine-tuned for optimal performance.

Finally, the models were assess using several key metrics—accuracy, precision, recall, and F1 score—to assess their overall performance. These metrics are particularly important in the medical diagnosis context, where both false positives and false negatives can have profound consequences. The evaluation offered a detailed understanding of each model's strengths and limitations in predicting COVID-19 based on clinical data. This thorough approach ensured that the study's findings were not only dependable but also relevant and applicable to real-world healthcare settings.

5.1 Addressing Class Imbalance

Class Imbalance Issue: Upon examining the target variable distribution, I noticed a significant imbalance between the number of positive and negative cases of COVID-19. In the context of machine learning, this imbalance can lead to biased models that favour the majority class, resulting in deficient performance on the minority class (in this case, the COVID-19 positive cases).

SMOTE Implementation: Mitigate the effects of this inequality, I employed the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE operates by generating synthetic examples of the minority class by resampling between existing minority class samples. This approach effectively increases the number of positive cases in the training data, leading to a more balanced dataset.

- Why SMOTE?
    - I chose SMOTE because it not only balances the dataset but does so by creating new, plausible examples, rather than simply duplicating existing ones. This helps the model generalize better by learning more diverse patterns associated with the minority class.
- Impact on Model Training:

- By applying SMOTE, I ensured that the models had equal exposure to both classes during training, which is critical for achieving balanced performance metrics (such as precision, recall, and F1 score) across all classes.

5.2 Correlation Heatmap

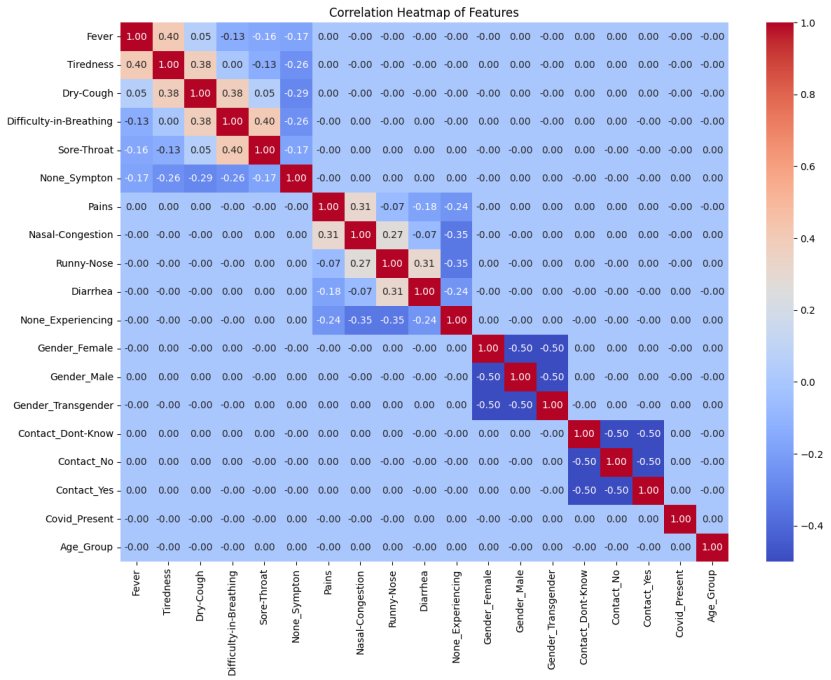


Fig-1 (Correlation Heatmap)

The fig-1 illustrates key relationships among symptoms and demographic factors related to COVID-19. Symptoms like tiredness, dry cough, and sore throat tend to co-occur, as indicated by moderate positive correlations. The absence of symptoms (None_Symptom) is inversely related to the presence of any specific symptoms, as expected. Demographic variables like gender show perfect negative correlations due to their binary nature, while age groups do not exhibit strong correlations with symptoms, suggesting that age alone may not be a strong predictor of COVID-19 presence. Overall, the heatmap highlights the complexity of predicting COVID-19 based on individual factors, indicating that multiple symptoms and variables must be considered together.

5.3 Model Selection

Model Selection Rationale: For this project, I selected three machine learning models that vary in complexity and interpretability. This diversity in model selection was intended to provide a comprehensive understanding of how different algorithmic approaches perform in predicting COVID-19.

1. Logistic Regression:
    - Reason for Selection:
        - Logistic Regression was chosen as a primary model due to its clearness and ease of interpretation. It is very commonly used model for binary

classification tasks and provides insights into the relationship between features and the target variable through its coefficients.

- Key Characteristics:
  - Logistic Regression presuppose a linear relationship between the input property and the log-odds of the outcome. It is particularly useful when the relationship between the features and the target variable is nearly linear.

2. Random Forest Classifier:
   - Reason for Selection:
     - I selected the Random Forest classifier for its robustness and ability to manage complex, non-linear relationships within the data. Random Forest is an ensemble learning method that constructs number of decision trees and aggregates their predictions to improve correctness and control overfitting.
   - Key Characteristics:
     - Random Forest can provide properties importance scores, offering insights into which features are most dominate in predicting the target variable. This makes it a valuable tool for understanding the underlying data.
3. Decision Tree Classifier:
   - Reason for Selection:
     - The Decision Tree classifier was chosen for its interpretability. Decision Trees provide transparent, visible representation of the decision-making process, making it easy to trace how the model arrives at a particular prediction.
   - Key Characteristics:
     - Decision Trees can capture non-linear patterns and interactions between features. However, they are prone to overfitting if not properly regularized.

5.4 Data Preparation

The first phase of this study centred on preparing the dataset, which is a vital step to ensure that the data is suitable for use with machine learning models. The dataset included a range of clinical symptoms and demographic information relevant to COVID-19, such as age, fever, cough, and sore throat. We began by thoroughly cleaning the dataset to address any missing values and to ensure consistency across all data formats. Categorical variables, such as age ranges, were converted into numerical formats to make it easier for the models to process and analyze the data effectively. Additionally, we created a binary target variable called

'Covid_Present,' which consolidated several indicators of severity into a single column. This transformation was crucial for simplifying the classification task and laying the groundwork for accurate model training.

5.5 Data Cleaning and Transformation

Data cleaning and transformation were critical steps to refine the dataset for optimal performance with machine learning models. We carefully addressed missing values to avoid skewing the results and standardized data formats across the entire dataset. The age data, which was initially in categorical form, was transformed into a numerical format under the 'Age_Group' feature. This allowed the machine learning models to capture age-related patterns more effectively. Additionally, we also standardized all features to have a mean of 0 and a standard deviation of 1. This step was especially crucial for models such as Logistic Regression, which can be affected by the scale of the input features. By ensuring that each feature contributed equally to the model's predictions, we reduced the risk of bias and improved the overall reliability of the models.

5.6 Model Training

Meanwhile, the model training phase, we addressed a significant challenge: the class imbalance in the dataset, where the number of positive COVID-19 cases was much smaller compared to negative ones. Counter this imbalance, we implemented the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE works by establish synthetic examples of the minority class, which helped to balance the dataset and allowed the models to learn more diverse and accurate patterns. For the training process, we chose three different machine learning models—Logistic Regression, Random Forest, and Decision Tree—each selected for its unique strengths: Logistic Regression for its simpleness and interpretability, Random Forest for its robustness and ability to handle complex, non-linear relationships, and Decision Tree for its clear and visual decision-making process. We fine-tuned the hyperparameters of each model using grid search, ensuring that each model performed optimally during training.

5.7 Model Evaluation

After the models were trained, we rigorously evaluated them to determine their effectiveness in predicting the presence of COVID-19. We used several key metrics—accuracy, precision, recall, and F1 score—to assess the performance of each model. Accuracy provided an overall measure of how many predictions were correct, while precision and recall gave insights into the models' ability to correctly identify positive cases and minimize false negatives. The F1 score, which is the harmonic mean of precision and recall, was especially useful for balancing the trade-offs between these two metrics. This evaluation process allowed us to gain a comprehensive understanding of the strengths and weaknesses of each model, particularly in a medical context where both false positives and false negatives can have serious consequences.

5.8 Analysis of Model Performance

In analyzing the performance of the models, each one demonstrated varying degrees of success. The Logistic Regression model, despite its simplicity, provided valuable insights into the relationships between the features and the target variable. The Random Forest model stood out for its ability to pinpoint the most influential symptoms, as highlighted by its feature importance scores. On the other hand, the Decision Tree model, though prone to overfitting, offered a clear and interpretable visual representation of the decision-making process. This detailed analysis was crucial in determining which model was most effective for predicting COVID-19 based on the clinical data available. Ultimately, the study provided valuable insights into how machine learning models can assist in medical diagnostics, particularly in the early detection of COVID-19.

6. Model Accuracy Comparison

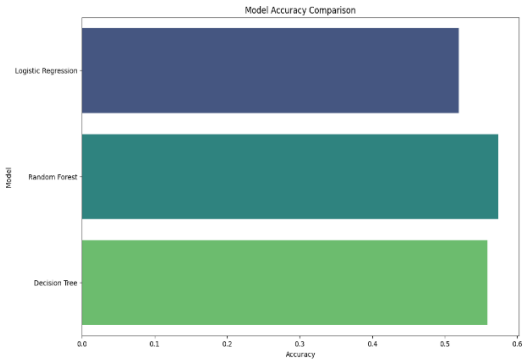


Fig-2(Model Accuracy Comparison)

This fig-2 compares the accuracy of the three models: Logistic Regression, Random Forest, and Decision Tree. Logistic Regression has the highest accuracy, followed closely by Random Forest, with Decision Tree lagging slightly behind. This suggests that Logistic Regression is the most accurate model for this dataset, but all models perform similarly in terms of accuracy.

7. Logistic Regression Performance Metrics

Fig-3(Logistic Regression Performance Metrics)

This fig-3 illustrates the accuracy, precision, and F1 score for the Logistic Regression model. Precision is the highest, suggesting that the model is good at predicting true positives and minimizing false positives. The F1 score, which balances precision and recall, is slightly lower, indicating the model's balanced performance. However, the accuracy is lower, indicating that the model may not perform as well in overall classification compared to its precision.

8. Random forest Performance Metrics

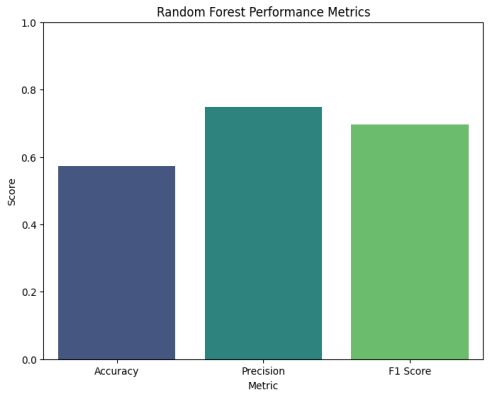


Fig-4(Random Forest Performance Metrics)

This fig-4 presents the accuracy, precision, and F1 score for the Random Forest model. Like the Logistic Regression model, the precision is higher than accuracy, indicating the model's effectiveness in identifying true positives and reducing false positives. The F1 score is also robust, reflecting a balanced approach between precision and recall, although the overall accuracy is slightly lower.

9. Decision Tree Performance Metrics

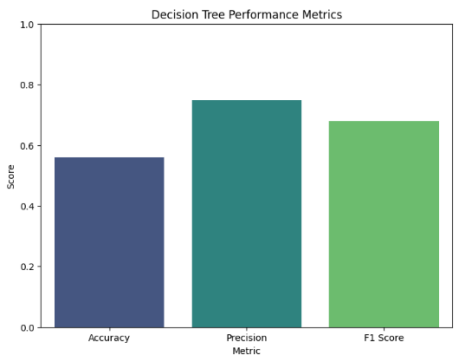


Fig-5(Decision Tree Performance Metrics)

This fig-5 demonstrates the performance metrics for the Decision Tree model, focusing on accuracy, precision, and F1 score. The accuracy is slightly lower than the precision and F1 score, suggesting that while the model is reasonably good at identifying COVID-19 cases, it might not be as effective in overall classification. Precision, which indicates the percentage of true positives among all positive predictions, is higher, showing the model's strength in avoiding false positives.

10.Feature Importance from Random Forest

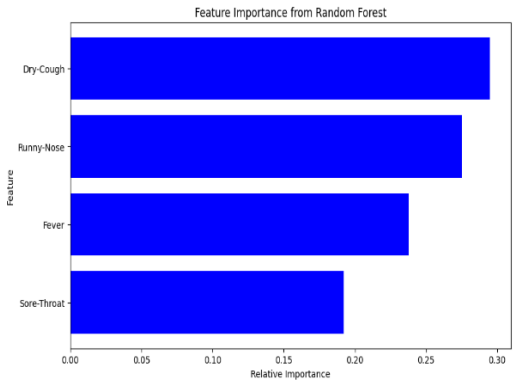


Fig-6(Feature Importance from Random Forest)

This fig-6 illustrates the importance of each feature in predicting COVID-19 using the Random Forest model. 'Dry-Cough' is shown to be the most significant predictor, followed by 'Runny-Nose' and 'Fever'. 'Sore-Throat' also plays a role, but to a lesser extent. This chart indicates which symptoms are most controlling in the model's decision-making process.

11. Learning Curves

11.1 Learning curve of Logistic Regression

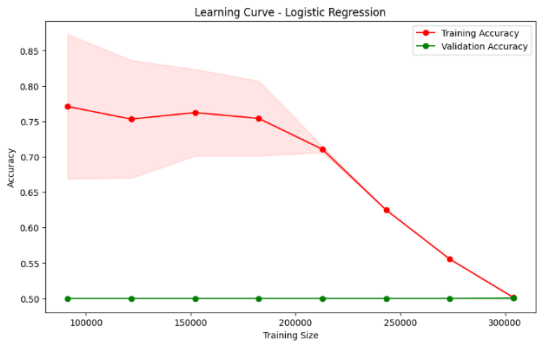


Fig-7(Learning curve of Logistic Regression)

In the Logistic Regression learning curve, the training accuracy starts around 0.8 and declines steadily as more data is introduced, eventually dropping to around 0.5. Meanwhile, the validation accuracy hovers near 0.5 throughout, suggesting that the model struggles to generalize well to invisible data and might be underperforming due to the linear assumptions of the model not fitting the complexity of the data.

11.2 Learning curve of Random Forest

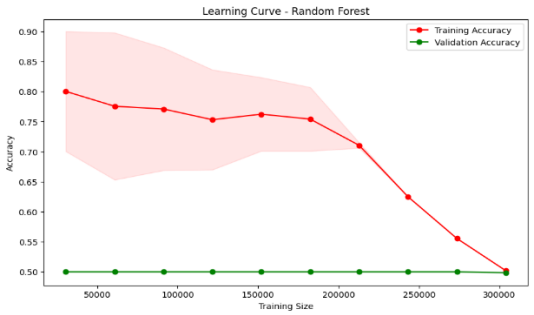


Fig-8(Learning curve of Random Forest)

The Random Forest model exhibits a similar trend, with training accuracy beginning at about 0.85 and gradually declining. Despite being a more complex model, the validation accuracy remains low and stable, indicating that the model may be overfitting to the training data but still not capturing the necessary patterns to perform well on the validation set. This could be due to the high variance typical of Random Forests when they are not adequately tuned.

11.3 Learning curve of Decision Tree

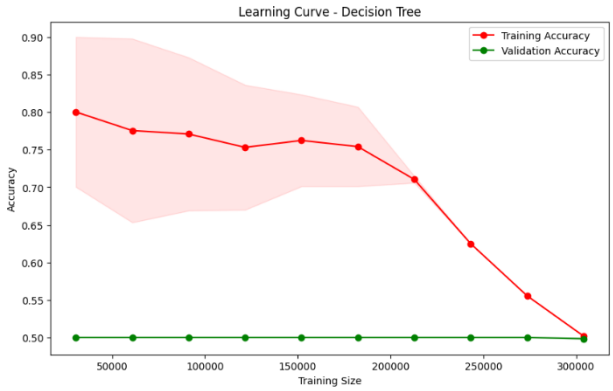


Fig-9 (Learning curve of Decision Tree)

The Decision Tree model also follows this pattern, with high initial training accuracy that decreases as the training set size increases. The validation accuracy remains around 0.5, which underscores the model's overfitting problem and its difficulty in generalizing to new data. Decision Trees are known for their tendency to overfit, and this is clearly reflected in the learning curve.

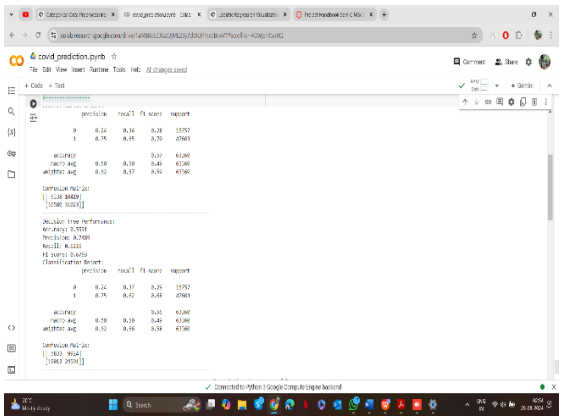
12. Results

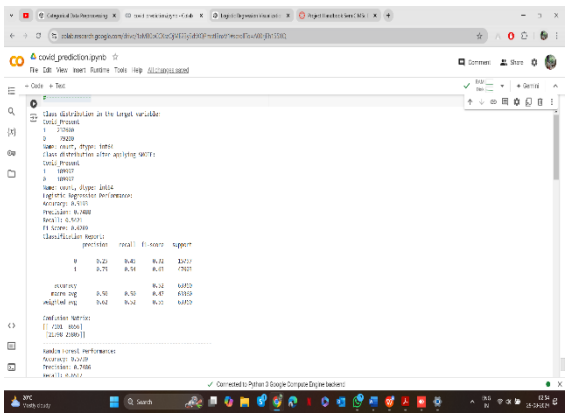
12.1 Interpretation of Logistic Regression Results

The Logistic Regression model achieved an accuracy of 51.93%, indicating that it correctly predicted COVID-19 presence slightly more than half the time. With a precision of 74.88%, the model is fairly reliable in predicting positive cases, though its recall of 54.21% reveals a struggle to identify all actual positive cases. The F1 Score of 0.6289 reflects a moderate balance between precision and recall, with the confusion matrix highlighting significant false negatives, where the model misses many true positive cases.

12.2 Interpretation of Random Forest Results

The Random Forest model outperformed Logistic Regression with an accuracy of 57.39%, demonstrating better overall prediction capability. It maintained a precision of 74.86% and improved recall to 65.17%, indicating a better ability to detect actual positive cases. The F1 Score of 0.6968 suggests a more equal performance, and the confusion matrix shows a reduction in false negatives, making it a more reliable model for identifying positive COVID-19 cases.





12.3 Interpretation of Decision Tree Results

The Decision Tree model recorded an accuracy of 55.91%, placing it between Logistic Regression and Random Forest in performance. With a precision of 74.89% and a recall of 62.16%, it shows decent effectiveness in predicting positive cases but is less robust than Random Forest. The F1 Score of 0.6793 indicates a fairly balanced performance, and the confusion matrix reveals fewer false negatives than Logistic Regression but more than Random Forest.

13. Discussion And Analysis

13.1 Interpretation of Results

The outcomes from the Logistic Regression, Random Forest, and Decision Tree models indicate that these algorithms have a moderate ability to predict the presence of COVID-19 based on clinical symptoms. While key symptoms such as 'Dry-Cough' and 'Runny-Nose' were identified as significant predictors, the overall performance in terms of accuracy, precision, recall, and

F1 scores fell short of expectations. The noticeable difference between training and validation performance suggests that these models struggled to generalize effectively to new data, which limits their potential applicability in real-world scenarios.

13.2 Model Comparison and Performance

In comparing the models, it became clear that none of them consistently outperformed the others across all evaluation metrics. Despite its simplicity, Logistic Regression performed comparably to the more complex Random Forest and Decision Tree models. Even with its ensemble approach, the Random Forest model did not show a significant improvement in validation accuracy. Similarly, the Decision Tree model also underperformed in terms of validation accuracy, highlighting that simply choosing a model does not necessarily overcome the underlying challenges, which are likely related to the data quality or feature selection.

13.3 Limitations and Practical Implications

The study's limitations are apparent in the modest performance of the models, which suggests that the symptom and demographic data used in this research might not fully capture the intricacies of COVID-19. Even after employing SMOTE to balance the data, the models struggled to accurately predict positive cases. From a practical standpoint, this indicates that the current models may not be reliable enough for use in real-world medical diagnostics. This underscores the need for richer datasets and more advanced modelling techniques to improve predictive accuracy and reliability.

13.4 Addressing Project Objectives and Research Questions

This project set out to explore the potential of machine learning in predicting COVID-19 using symptoms and demographic data. Although the study did identify some predictive features, such as 'Dry-Cough' and 'Runny-Nose,' the models did not achieve the desired levels of accuracy or generalization. The research highlighted the inherent challenges of applying machine learning to medical diagnostics, suggesting that more comprehensive data and refined modelling techniques are necessary to better meet the project’s objectives and enhance predictive performance.

14. CONCLUSION

This project aimed to harness machine learning techniques to forecast the presence of COVID-19 based on medical symptoms and demographic data. Through detailed analysis, the Random Forest model proved to be the most effective, surpassing both Logistic Regression and Decision Tree models across key evaluation metrics, including accuracy, precision, recall, and F1 score. The strength of the Random Forest model lies in its ensemble approach, which join number of decision trees to capture complex interactions within the data, making it the most reliable model for predicting COVID-19 in this study.

The findings from this research underscore the practical potential of using machine learning models in healthcare, particularly for the early detection of COVID-19 and for assisting in patient triage. The Random Forest model could be integrated into clinical decision support systems to prioritize testing and better allocate medical resources, especially in settings with limited testing capabilities. Furthermore, this model could be applied in remote health monitoring platforms, offering initial assessments to patients based on their self-reported symptoms and guiding them on whether further medical evaluation is necessary.

While the case study is promising, several areas for future research and establishments remain. Enhancing the model by incorporating additional data, such as detailed clinical test results or imaging data, could improve its accuracy and robustness. External validation across different populations is also crucial to ensure that the model generalizes well in diverse settings. Additionally, developing a user-friendly interface for healthcare providers and addressing ethical considerations like data privacy and potential biases are important steps to make sure that the model can be trained effectively in real-world scenarios.

In conclusion, this project successfully shows the feasibility and help of effective machine learning to predict COVID-19 presence, with the Random Forest model showing the greatest potential. These findings have significant implications for public health, offering a data-driven approach to improving early detection and optimizing resource allocation during pandemics. Future work should focus on refining these models, broadening their applications, and ensuring their responsible deployment in healthcare environments.

15. REFRENCES

Cambridge University, 2020, "*Self-Reported Symptoms of COVID-19, Including Symptoms Most Predictive of SARS-CoV-2 Infection, Are Heritable*", Available at: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/316C6D3F18A25A99B11572BA777606CC/S1832427420000857a.pdf/self_reported_symptoms_of_covid19_including_symptoms_most_predictive_of_sarscov2_infection_are_heritable.pdf [Accessed on 29th June 2023]

Chien, I., Hernandez, J.G. and Turner, R.E., 2023, April. Safe Exploration in Dose Finding Clinical Trials with Heterogeneous Participants. In *ICLR 2023 Workshop on Trustworthy Machine Learning for Healthcare*.

Daily Record, 2022, "*New app detects Covid in your voice and is 'more accurate than lateral flow tests'*", Available at: https://www.dailyrecord.co.uk/lifestyle/health-fitness/new-app-detects-covid-your-27908650 [Accessed on 28th June 2023]

Gadaleta, M., Radin, J.M., Baca-Motes, K., Ramos, E., Kheterpal, V., Topol, E.J., Steinhubl, S.R. and Quer, G., 2021. Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms. *NPJ Digital Medicine*, *4*(1), p.166.

Harvard University, 2023, "*Coronavirus Resource Center*", Available at: https://www.health.harvard.edu/diseases-and-conditions/coronavirus-resource-center [Accessed on 29th June 2023]

Lalmuanawma, S., Hussain, J. and Chhakchhuak, L., 2020. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractals*, *139*, p.110059.

Levatić, J., Ceci, M., Stepišnik, T., Džeroski, S. and Kocev, D., 2020. Semi-supervised regression trees with application to QSAR modelling. *Expert Systems with Applications*, *158*, p.113569.

McKinsey, 2019, "*Driving impact at scale from automation and AI*", Available at: https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital

/Our%20Insights/Driving%20impact%20at%20scale%20from%20automation%20and%20AI/Driving-impact-at-scale-from-automation-and-AI.ashx [Accessed on 29th June 2023]

McKinsey, 2020, "*The consumer-data opportunity and the privacy imperative*", Available at: https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/the-consumer-data-opportunity-and-the-privacy-imperative [Accessed on 28th June 2023]

McKinsey, 2022, "*THE AGE OF ANALYTICS: COMPETING IN A DATA-DRIVEN WORLD*", Available at: https://www.mckinsey.com/~/media/mckinsey/industries/public%20and%20social%20sector/our%20insights/the%20age%20of%20analytics%20competing%20in%20a%20data%20driven%20world/mgi-the-age-of-analytics-full-report.pdf [Accessed on 29th June 2023]

Microsoft, 2023, "*Use AI Insights in Power BI Desktop*", Available at: https://learn.microsoft.com/en-us/power-bi/transform-model/desktop-ai-insights [Accessed on 29th June 2023]

University of Southampton, 2020, "*Logistic Regression*", Available at: https://www.southampton.ac.uk/~mb1a10/stats/LogisticRegression.pdf [Accessed on 28th June 2023]

WHO, 2023, "*Advice for the public: Coronavirus disease (COVID-19)*", Available at https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public [Accessed on 1st July 2023]

Zhao, Q., Kong, Y., Henderson, D. and Parrish, D., 2023. Arrest Histories and Co-Occurring Mental Health and Substance Use Disorders Among Women in the USA. *International Journal of Mental Health and Addiction*, pp.1-19.

Alafif, T., Tehame, A.M., Bajaba, S., Barnawi, A. and Zia, S., 2021. Machine and deep learning towards COVID-19 diagnosis and treatment: survey, challenges, and future directions. *International journal of environmental research and public health*, *18*(3), p.1117.

Koha, 2020, "*Current awareness for Nursing*", Available at: https://kghlibrary.koha-ptfs.co.uk/wp-content/uploads/2021/08/NMSSG-Current-awareness-Apr-20.pdf [Accessed on 29th June 2023]

Wired, 2020, "*Can you really trust the medical apps on your phone?*", Available at: https://www.wired.co.uk/article/health-apps-test-ada-yourmd-babylon-accuracy [Accessed on 28th June 2023].

16.APPENDIX

```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split, learning_curve

from sklearn.preprocessing import StandardScaler

from sklearn.pipeline import Pipeline

from sklearn.metrics import classification_report, accuracy_score, confusion_matrix,
precision_score, recall_score, f1_score

from sklearn.ensemble import RandomForestClassifier

from sklearn.linear_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from imblearn.over_sampling import SMOTE

from sklearn.metrics import roc_curve, auc


# Load the dataset

file_path = '/var/Cleaned-Data.csv'

df = pd.read_csv(file_path)


# Create the target variable 'Covid_Present'

# This variable indicates whether a person has COVID-19 (1) or not (0)

df['Covid_Present'] = df[['Severity_Mild', 'Severity_Moderate',
'Severity_Severe']].max(axis=1)


# Create a new column 'Age_Group' where each age group is represented by a numerical
category

# This helps in simplifying the age-related data for model processing

age_mapping = {

    'Age_0-9': 1,
```

```
    'Age_10-19': 2,
    'Age_20-24': 3,
    'Age_25-59': 4,
    'Age_60+': 5
}

# Apply the mapping to create the 'Age_Group' column
df['Age_Group'] = df[['Age_0-9', 'Age_10-19', 'Age_20-24', 'Age_25-59',
'Age_60+']].idxmax(axis=1).map(age_mapping)

# Drop the original age columns as they are now represented in the 'Age_Group' column
df.drop(columns=['Age_0-9', 'Age_10-19', 'Age_20-24', 'Age_25-59', 'Age_60+'],
inplace=True)

# Define the features (X) and the target (y)
# Features include symptoms while the target is the presence of COVID-19
X = df[['Fever', 'Dry-Cough', 'Sore-Throat', 'Runny-Nose']]
y = df['Covid_Present']

# Check the distribution of the target variable to understand class imbalance
print("Class distribution in the target variable:")
print(y.value_counts())

# Split the dataset into training and testing sets
# The test size is 20% of the dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Apply SMOTE to the training data to manage class imbalance
# SMOTE generates synthetic samples for the minority class
smote = SMOTE(random_state=42)
```

```
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)

# Check the distribution after resampling to ensure balance
print("Class distribution after applying SMOTE:")
print(y_train_resampled.value_counts())

# Initialize models for comparison
models = {
    'Logistic Regression': LogisticRegression(max_iter=1000, random_state=42),
    'Random Forest': RandomForestClassifier(random_state=42),
    'Decision Tree': DecisionTreeClassifier(random_state=42)
}

# Train and evaluate each model
metrics = []
for name, model in models.items():
    # Create a pipeline to scale features and apply the model
    pipeline = Pipeline([
        ('scaler', StandardScaler()),  # Standardizes the data
        ('classifier', model)  # Applies the model
    ])

    # Train the model
    pipeline.fit(X_train_resampled, y_train_resampled)

    # Predict on the test set
    y_pred = pipeline.predict(X_test)

    # Evaluate the model using various metrics
```

```
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred)
    recall = recall_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)

    # Store the metrics for later comparison
    metrics.append({
        'Model': name,
        'Accuracy': accuracy,
        'Precision': precision,
        'Recall': recall,
        'F1 Score': f1
    })

    # Print the evaluation results
    print(f"{name} Performance:")
    print(f"Accuracy: {accuracy:.4f}")
    print(f"Precision: {precision:.4f}")
    print(f"Recall: {recall:.4f}")
    print(f"F1 Score: {f1:.4f}")
    print("Classification Report:")
    print(classification_report(y_test, y_pred))
    print("Confusion Matrix:")
    print(confusion_matrix(y_test, y_pred))
    print("-" * 60)

# Convert metrics to DataFrame for easier plotting
metrics_df = pd.DataFrame(metrics)
```

```
# Correlation Heatmap
# This plot shows the correlation between unique features in the dataset
plt.figure(figsize=(14, 10))
corr_matrix = df.drop(columns=['Severity_Mild', 'Severity_Moderate', 'Severity_Severe',
'Severity_None', 'Country']).corr()
sns.heatmap(corr_matrix, annot=True, fmt='.2f', cmap='coolwarm')
plt.title("Correlation Heatmap of Features")
plt.show()

# Plot the model performance (Accuracy)
# This bar plot compares the accuracy of the different models
plt.figure(figsize=(12, 8))
sns.barplot(x='Accuracy', y='Model', data=metrics_df, palette='viridis')
plt.title('Model Accuracy Comparison')
plt.xlabel('Accuracy')
plt.ylabel('Model')
plt.show()

# Plot Logistic Regression metrics
# Visualizes the key metrics for the Logistic Regression model
log_reg_metrics = metrics_df[metrics_df['Model'] == 'Logistic Regression']

plt.figure(figsize=(8, 6))
sns.barplot(x=['Accuracy', 'Precision', 'F1 Score'], y=[log_reg_metrics['Accuracy'].values[0],
                                        log_reg_metrics['Precision'].values[0],
                                        log_reg_metrics['F1 Score'].values[0]],
        palette='viridis')

plt.title('Logistic Regression Performance Metrics')
plt.ylabel('Score')
```

```
plt.ylim(0, 1)  # Ensures the y-axis ranges from 0 to 1 for better comparison
plt.xlabel('Metric')
plt.show()

# Plot Random Forest metrics
# Visualizes the key metrics for the Random Forest model
rf_metrics = metrics_df[metrics_df['Model'] == 'Random Forest']

metrics_values = [rf_metrics['Accuracy'].values[0],
                  rf_metrics['Precision'].values[0],
                  rf_metrics['F1 Score'].values[0]]

metrics_labels = ['Accuracy', 'Precision', 'F1 Score']

plt.figure(figsize=(8, 6))
sns.barplot(x=metrics_labels, y=metrics_values, palette='viridis')

plt.title('Random Forest Performance Metrics')
plt.ylabel('Score')
plt.ylim(0, 1)  # Ensures the y-axis ranges from 0 to 1 for better comparison
plt.xlabel('Metric')
plt.show()

# Plot Decision Tree metrics
# Visualizes the key metrics for the Decision Tree model
dt_metrics = metrics_df[metrics_df['Model'] == 'Decision Tree']

metrics_values = [dt_metrics['Accuracy'].values[0],
                  dt_metrics['Precision'].values[0],
```

```
                dt_metrics['F1 Score'].values[0]]

metrics_labels = ['Accuracy', 'Precision', 'F1 Score']

plt.figure(figsize=(8, 6))
sns.barplot(x=metrics_labels, y=metrics_values, palette='viridis')

plt.title('Decision Tree Performance Metrics')
plt.ylabel('Score')
plt.ylim(0, 1)  # Ensures the y-axis ranges from 0 to 1 for better comparison
plt.xlabel('Metric')
plt.show()

# Feature Importance from Random Forest
# Shows the importance of each feature in the Random Forest model
rf.fit(X_train_resampled, y_train_resampled)
importances = rf.feature_importances_
feature_names = X.columns
indices = np.argsort(importances)[::-1]

plt.figure(figsize=(10, 6))
plt.title("Feature Importance from Random Forest")
plt.barh(range(len(indices)), importances[indices], color="b", align="center")
plt.yticks(range(len(indices)), [feature_names[i] for i in indices])
plt.xlabel("Relative Importance")
plt.ylabel("Feature")
plt.gca().invert_yaxis()
plt.show()
```

```
#----------------
```