A FINAL REPORT ON

HEART DISEASE PREDICTION USING MACHINE LEARNING

PREPARED BY TETTEH CHINELO NKIRUKA C.

IN FULFILMENT OF DATA SCIENCE FINAL PROJECT ASSESSMENT

ON 7TH FEBRUARY, 2025

1. ABSTRACT

This report presents a comprehensive analysis of heart disease prediction using machine learning techniques. The study involves data preprocessing, exploratory data analysis, model training, evaluation, and comparison of multiple classifiers. Key findings, insights, and recommendations are provided to enhance predictive accuracy and real-world applicability,

2. INTRODUCTION

Heart disease remains one of the leading causes of mortality worldwide, accounting for millions of deaths annually. It encompasses various cardiovascular conditions, including coronary artery disease, heart failure, and arrhythmias, often resulting from factors such as poor lifestyle habits, genetic predisposition, and underlying health conditions like hypertension and diabetes.

Early detection of heart disease is critical in preventing severe complications and improving patient outcomes. Traditional diagnostic methods, such as electrocardiograms, stress tests, and blood tests, require medical expertise and may not always provide timely interventions. This is where machine learning (ML) plays a crucial role.

By leveraging ML algorithms, we can analyze complex medical data to identify patterns and risk factors associated with heart disease. Predictive models can assist healthcare professionals in early diagnosis, risk assessment, and personalized treatment plans, ultimately reducing hospitalizations and mortality rates. This study aims to explore the effectiveness of different ML models in predicting heart disease based on patient data, evaluating their performance, and identifying the most reliable approach for real-world applications.

3. OBJECTIVE

To develop a machine learning model to predict the likelihood and presence of heart disease in patients using clinical and lifestyle data.

4.0 METHODOLOGY (USING PYTHON LANGUAGE FOR CODING IN A JUPYTER NOTEBOOK)

4.1 Importing of all necessary libraries for data preprocessing, analysis, visualizations and model development

4.2 Loading of dataset into a pandas dataframe.

4.3 Dataset Summary & Preprocessing

4.3.1 Dataset Overview

Dataset Information

https://drive.google.com/file/d/1rLAiGfOcS2YcgwAeZb8RTUCE_MgPtpCo/view?usp=drive_link

The dataset contains 13 features related to patient demographics, medical history, and test results. The target variable is to determine whether a patient has heart disease (binary classification) - 1 indicates the presence of heart disease and 0 indicates no heart disease.

The descriptions of the attributes in each column include:

- **Age:** Age of the patient in years.
- **Sex:** Gender of the patient (1 = male; 0 = female).
- **Chest Pain Type (cp)**: Type of chest pain experienced by the patient, categorized into four types: {0: Typical angina, 1: Atypical angina, 2: Non-anginal pain, 3: Asymptomatic}
- **Resting Blood Pressure (trestbps)**: Resting blood pressure of the patient, measured in mmHg.
- **Serum Cholesterol (chol)**: Serum cholesterol levels in mg/dL.
- **Fasting Blood Sugar (fbs)**: Whether fasting blood sugar is > 120 mg/dL (1 = true; 0 = false).
- **Resting Electrocardiographic Results (restecg)**: Results of resting ECG, categorized into {0: Normal, 1: Having ST-T wave abnormality (e.g., T wave inversions and/or ST elevation or depression of > 0.05 mV), 2: Showing probable or definite left ventricular hypertrophy}
- **Maximum Heart Rate Achieved (thalach)**: The highest heart rate achieved during exercise.
- **Exercise-Induced Angina (exang)**: Presence of exercise-induced angina (1 = yes; 0 = no).
- **Oldpeak**: ST depression induced by exercise relative to rest.
- **The Slope of The Peak Exercise St Segment (slope)**: Slope of the peak exercise ST segment, categorized into: {0: Upsloping, 1: Flat, 2: Downsloping}
- **Number Of Major Vessels (ca):** Number of major vessels (0–3) colored by fluoroscopy.
- **Thal**: A blood disorder type categorized as: {0: Normal, 1: Fixed defect, 2: Reversible defect.}

4.3.2 Data Preprocessing Steps

- Data inspection for statistical summary, information, datatypes, shape, columns and duplicates.
- Handling Missing Values**:** Checked for any missing or null values.
- Continuous features were either standardized or normalized using StandardScaler. This ensured that features were on a comparable scale, which is especially important for distance-based or gradient-based algorithms.

- Feature Selection: Pearson Correlation Analysis revealed that some features had very poor correlation scores hence were irrelevant for prediction.
- Chi-Squared Test: For categorical variables, the chi-squared test helped identify which features were statistically associated with the target (heart disease). Features with low chi-squared scores were considered less informative and were candidates for removal.
- Overall, these preprocessing and feature selection steps helped narrow the focus to a set of features that were most relevant for predicting heart disease, thereby potentially improving model performance and interpretability.

4.4. Exploratory Data Analysis (EDA)

4.4.1 Visualizations and Insights

- **Distribution of Target Variable:** Showed a balanced/unbalanced distribution of heart disease cases.
- **Correlation Heatmap:** Identified key features with strong correlations to heart disease.
- **Histograms:** Analyzed the distribution of numerical features.
- **Pairplots** and **scatter plots** were used.
- **Feature Importance:** Visualized the most influential features in model prediction.

4.5. Model development and evaluations.

- **Train-Test Split:** Splitting the dataset into 80% training and 20% testing data.

4.6 Model Comparisons & Performance Metrics Evaluations

Three classification models were evaluated:

- Logistic Regression
- Decision Tree
- Random Forest

The performance metrics used to test the models include accuracy, precision, recall, F1-Score, ROC-AUC.

4.7 Evaluation Metrics Visualization:

- **Bar plots** were used to compare accuracy, precision, recall and F1-Score.
- **Confusion matrix** was used visualize true positives, false positives, true negatives and false negatives.
- **ROC curve** was used to find out the true positive and false positive rates.

5.0 Results

5.1 Model comparisons and performance metrics

Performance Metrics

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Logistic Regression	0.803	0.833	0.781	0.806	0.13
Decision Tree	0.770	0.800	0.750	0.774	0.78
Random Forest	0.836	0.893	0.781	0.833	0.92

- **Logistic Regression performance**: Provided a baseline level of performance. The model's accuracy, precision, recall and F1-Score were moderate, suggesting that it was able to capture some relationships in the data.
- **Random Forest Performance**: Generally showed superior performance across all metrics was best at distinguishing between patients with and without heart disease.
- **Decision Tree Performance:** The model's accuracy, precision, recall and F1-Score suggested that it was able to capture some relationships in the data, but not as much as the random forest model.
- Overall summary of findings: The Random Forest model consistently outperformed the other two models across most evaluation metrics. This suggests that ensemble methods are particularly effective for this dataset, likely due to their ability to model complex, non-linear relationships and interactions.
- Logistic Regression served as a strong baseline but was limited by its linear nature.
- Feature Selection Impact: Removing redundant or uninformative features (as identified by correlation analysis and the chi-squared test) likely contributed to improved model performance by reducing noise and overfitting.
- Fine-tuning the best-performing model (Random Forest in this case) with techniques like hyperparameter optimization using GridSearchCV) potentially led to even better performance.

6. Key Takeaways & Recommendations

6.1 Key Takeaways

- Random Forest is the best-performing model with excellent balance across all metrics.
- Feature importance analysis suggests that chest pain type is the strongest predictor of heart disease.

- Higher recall models (like Decision Tree) may be preferred in scenarios where missing a true case is costly.

6.2 Recommendations

- **Further Model Optimization:** Additional hyperparameter tuning and ensemble techniques (e.g., stacking) could improve performance.
- **Class Imbalance Handling:** If needed, techniques like SMOTE or adjusting classification thresholds can improve recall.
- **Integration into Healthcare Systems:** Deploying this model in a real-world setting (e.g., hospital decision support systems) could help with early disease detection.

7. Conclusion

This study demonstrated the effectiveness of machine learning in predicting heart disease. The **Random Forest model showed superior performance**, making it the most reliable choice for deployment in clinical settings. Future improvements and real-world validation can further enhance its impact in preventive healthcare.

Appendix: Visualizations

(Include plots of confusion matrices, ROC curves, and feature importance charts here)

6. Key Takeaways & Recommendations

6.1 Key Takeaways

- **Random Forest is the best-performing model** with excellent balance across all metrics.
- **Feature importance analysis** suggests that factors like cholesterol level, blood pressure, and age are strong predictors of heart disease.
- **Higher Recall models (like Decision Tree) may be preferred** in scenarios where missing a true case is costly.

6.2 Recommendations

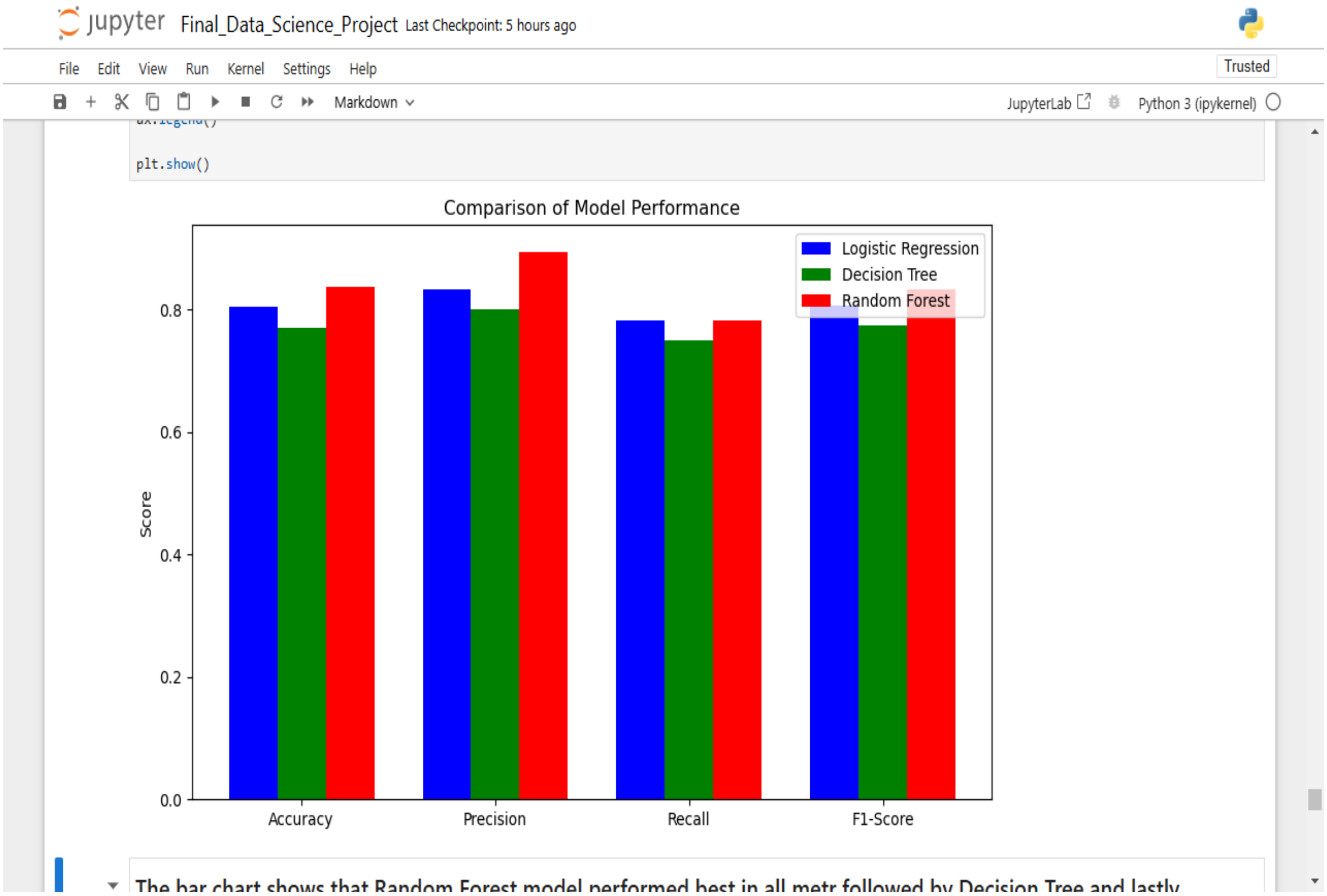
- **Further Model Optimization:** Additional hyperparameter tuning and ensemble techniques (e.g., stacking) could improve performance.
- **Class Imbalance Handling:** If needed, techniques like SMOTE or adjusting classification thresholds can improve recall.
- **Integration into Healthcare Systems:** Deploying this model in a real-world setting (e.g., hospital decision support systems) could help with early disease detection.

7. Conclusion

This study demonstrated the effectiveness of machine learning in predicting heart disease. The **Random Forest model showed superior performance**, making it the most reliable choice for deployment in clinical settings. Future improvements and real-world validation can further enhance its impact in preventive healthcare.

Appendix: Visualizations

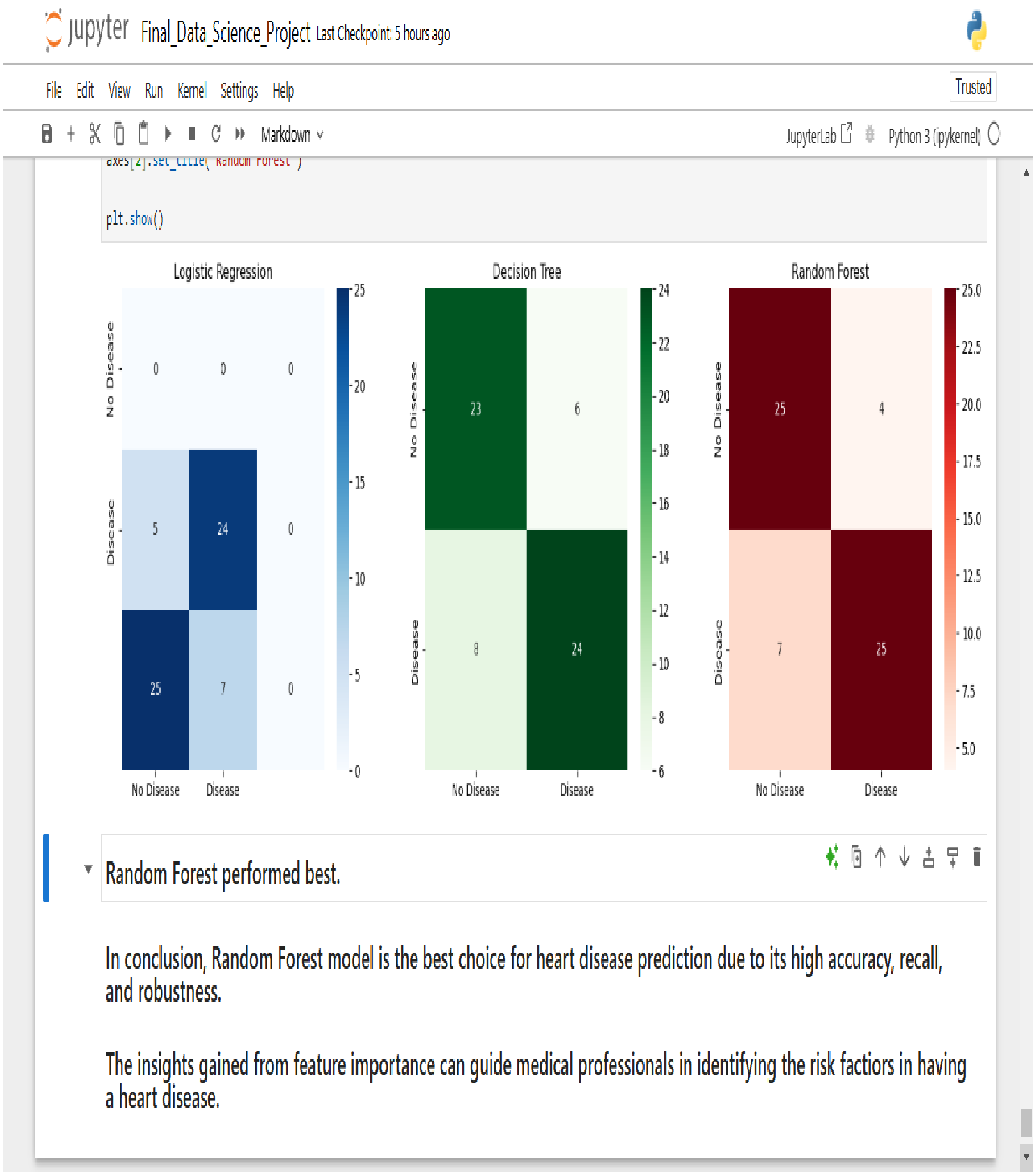
1. COMPARISON OF MODEL PERFORMANCES ON DIFFERENT EVALUATION METRICS



```
ax.legend()

plt.show()
```

The bar chart shows that Random Forest model performed best in all metr followed by Decision Tree and lastly

2. CONFUSION MATRIX FOR THE 3 MODELS



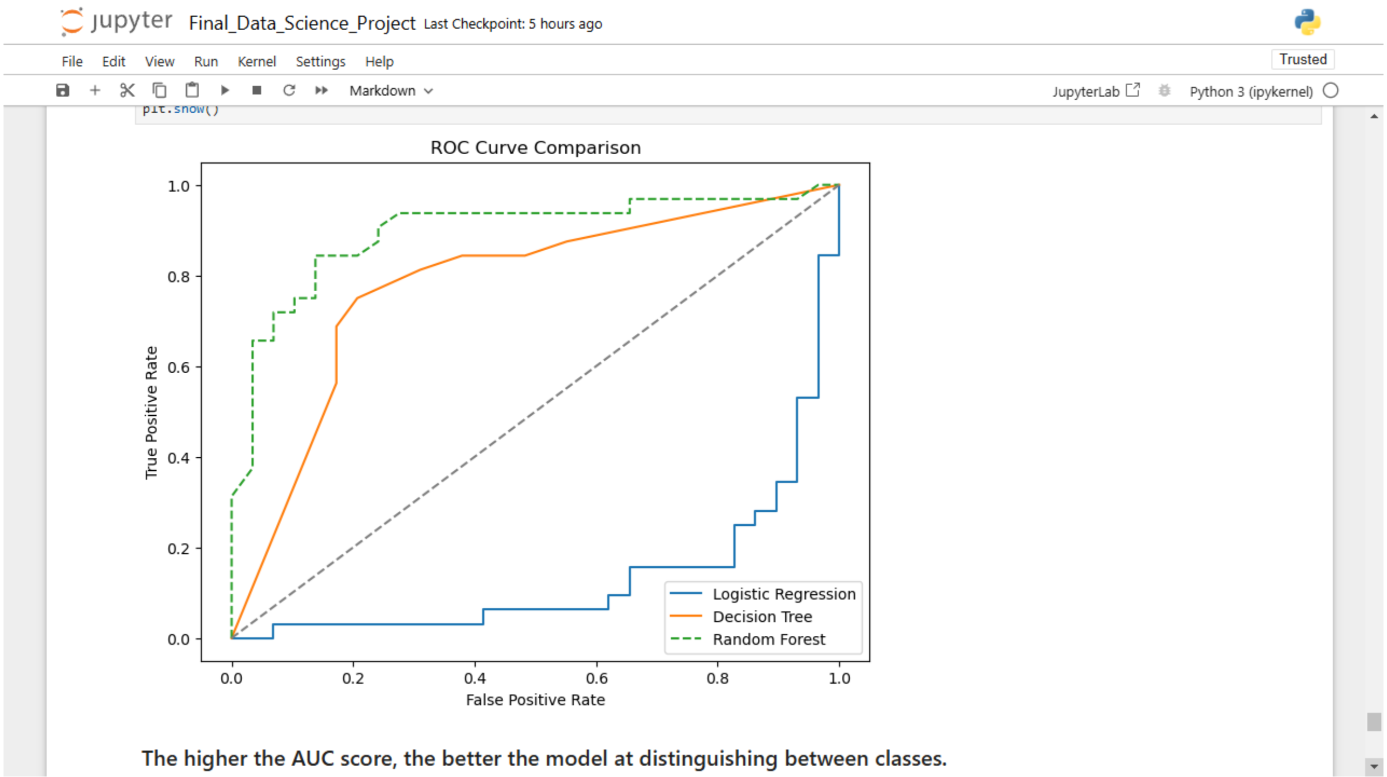
```
axes[2].set_title('Random Forest')

plt.show()
```

Random Forest performed best.

In conclusion, Random Forest model is the best choice for heart disease prediction due to its high accuracy, recall, and robustness.

The insights gained from feature importance can guide medical professionals in identifying the risk factors in having a heart disease.

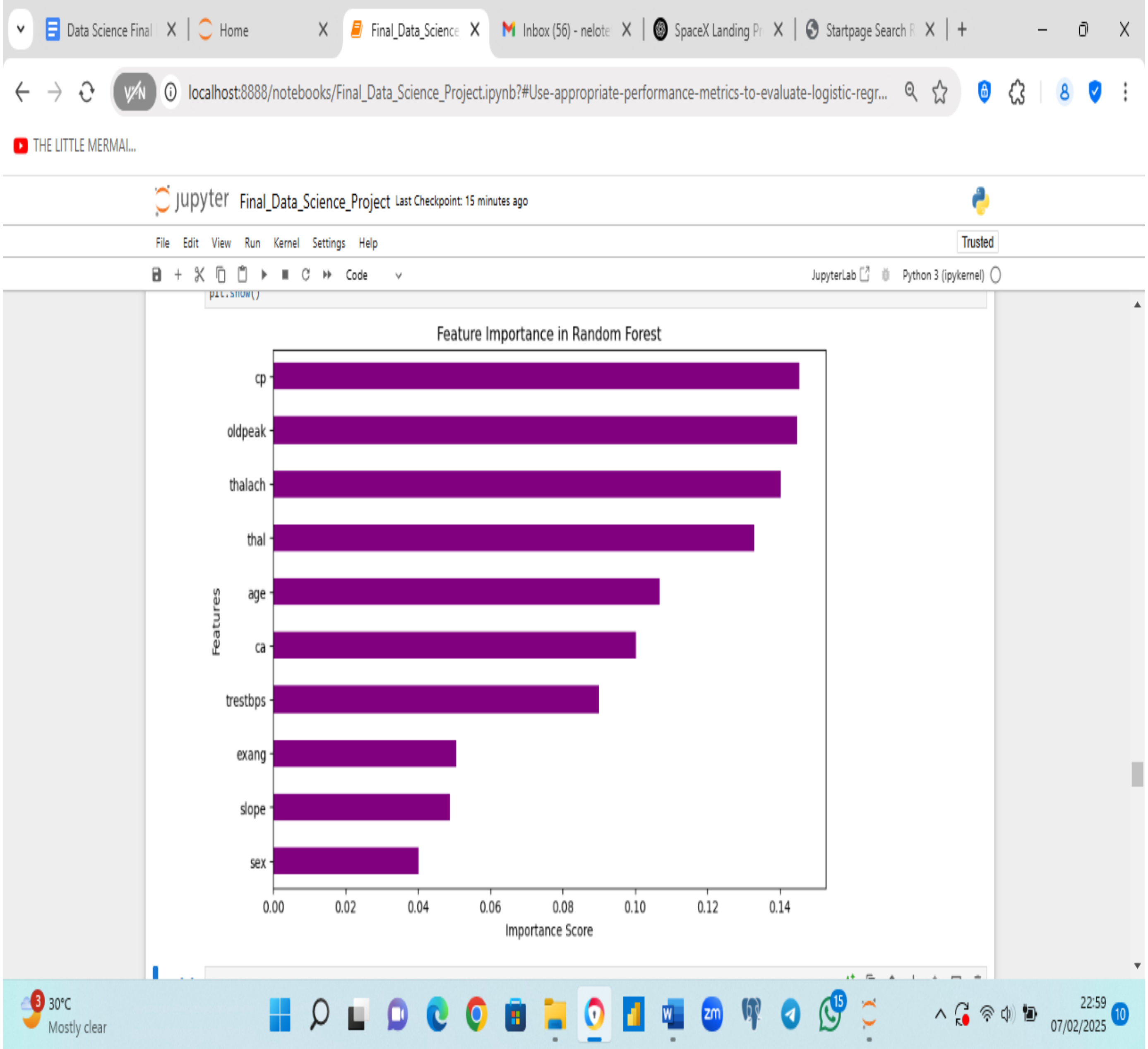
3. ROC-CURVE FOR THE 3 MODELS



```
plt.show()
```

The higher the AUC score, the better the model at distinguishing between classes.

4. FEATURE IMPOPRTANCE IN RANDOM FOREST



```
plt.show()
```