

**CSC460: Machine Learning by Dr. Abbass Rammal**

**Done by: Abdul Rahman Al Zaatari, Kamel Ibrahim**

**Project: Heart Disease Predictive Model**

****

Table of Contents

[**Topic** 3](#_Toc172471474)

[**Objective** 3](#_Toc172471475)

[**Background Information** 3](#_Toc172471476)

[**Cleveland Heart Disease Dataset** 3](#_Toc172471477)

[**Medical and Demographic Features** 4](#_Toc172471478)

[**Data Preprocessing:** 6](#_Toc172471479)

[**Code** 9](#_Toc172471480)

[**Data Preprocessing by us:** 9](#_Toc172471481)

[**Choosing and Training of Models** 13](#_Toc172471482)

[**Parameter Tuning (Optimization) and PCA** 14](#_Toc172471483)

[**Comparing Accuracies and Reaching Ensemble** 14](#_Toc172471484)

[**Evaluating the Models** 15](#_Toc172471485)

[**Making Predictions** 15](#_Toc172471486)

[**Results:** 16](#_Toc172471487)

[**Conclusion:** 17](#_Toc172471488)

[**Acknowledgements:** 18](#_Toc172471489)

# **Topic**

This project focuses on predicting heart disease using a supervised learning approach. The problem is a classification task where the goal is to determine whether a patient has heart disease (binary classification: presence or absence of heart disease).

# **Objective**

The objective of the model is to accurately predict the presence of heart disease in patients based on various medical and demographic features.

# **Background Information**

Heart disease, also known as cardiovascular disease, encompasses a range of conditions affecting the heart, including coronary artery disease, arrhythmias (irregular heartbeats), and heart defects. These conditions can lead to severe health issues, including heart attacks, strokes, and even death. Early and accurate prediction of heart disease is crucial for effective treatment and prevention.

# **Cleveland Heart Disease Dataset**

The dataset used in this project is sourced from the Cleveland Heart Disease dataset available on Kaggle. This dataset is widely used in the medical research community to study heart disease. It consists of patient records collected from the Cleveland Clinic Foundation. Each record includes various medical and demographic features that can help in predicting the presence of heart disease.

The Cleveland Heart Disease dataset is considered reliable and has been extensively used in academic research for developing machine learning models for heart disease prediction. The dataset contains 303 patient records with 14 attributes (features) that provide valuable information about each patient.

Source: <https://www.kaggle.com/datasets/ritwikb3/heart-disease-cleveland>

# **Medical and Demographic Features**

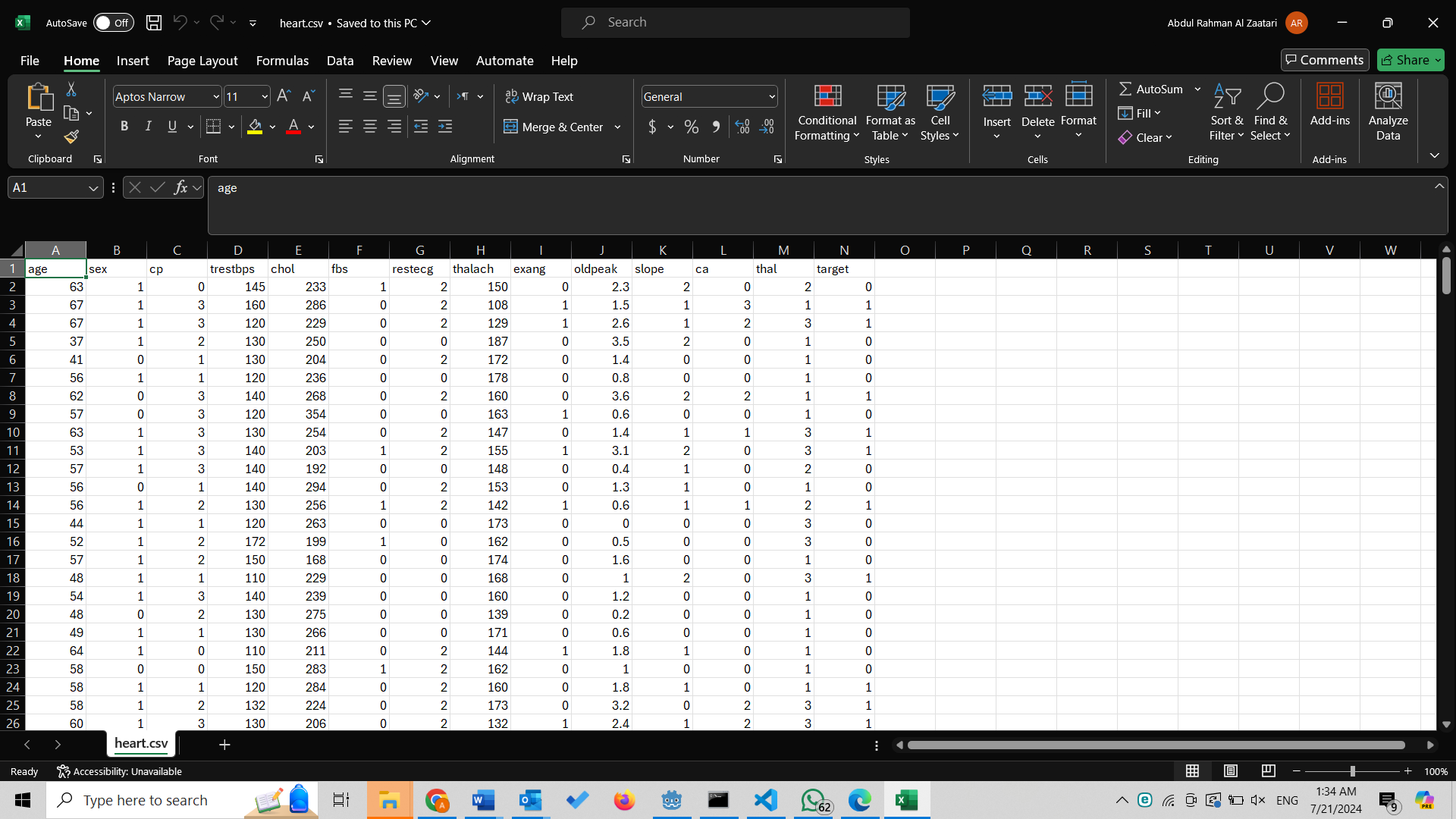
Here are the features included in the dataset, explanations are provided for people with no experience in the medical field:

1. **Age**: The age of the patient.
2. **Sex**: The sex of the patient (1 = male; 0 = female).
3. **CP (Chest Pain Type)**: The type of chest pain experienced by the patient.
   * 0: Typical angina (chest pain related to the heart).
   * 1: Atypical angina (chest pain not related to the heart).
   * 2: Non-anginal pain (not related to heart issues).
   * 3: Asymptomatic (no chest pain).
4. **Trestbps (Resting Blood Pressure)**: The blood pressure of the patient while at rest (measured in mm Hg).
5. **Chol (Serum Cholesterol)**: The cholesterol level in the patient's blood (measured in mg/dl).
6. **FBS (Fasting Blood Sugar > 120 mg/dl)**: Whether the patient’s blood sugar level is greater than 120 mg/dl after fasting (1 = yes; 0 = no).
7. **Restecg (Resting Electrocardiographic Results)**: Results from an electrocardiogram (ECG) test performed while the patient is at rest.
   * 0: Normal results.
   * 1: Abnormal results (ST-T wave abnormalities).
   * 2: Probable or definite left ventricular hypertrophy (enlarged heart).
8. **Thalach (Maximum Heart Rate Achieved)**: The highest heart rate achieved by the patient during physical activity.
9. **Exang (Exercise Induced Angina)**: Whether the patient experienced angina (chest pain) during exercise (1 = yes; 0 = no).

**Oldpeak (ST Depression Induced by Exercise Relative to Rest)**: A measure of abnormal heart activity during exercise compared to rest.

1. **Slope (Slope of the Peak Exercise ST Segment)**: The slope of the peak exercise ST segment.
   * 0: Upsloping (indicating increasing heart stress).
   * 1: Flat (indicating stable heart stress).
   * 2: Downsloping (indicating decreasing heart stress).
2. **Ca (Number of Major Vessels Colored by Fluoroscopy)**: The number of major blood vessels visible in a fluoroscopy image (0-3).
3. **Thal (Thalassemia)**: A blood disorder status determined by a blood test.
   * 1: Normal blood flow.
   * 2: Fixed defect (problem with blood flow that does not change).
   * 3: Reversible defect (problem with blood flow that can change).
4. **Target**: The diagnosis of heart disease (0 = absence; 1 = presence).

**A look at the dataset**:



# **Data Preprocessing:**

The Cleveland Heart Disease dataset underwent several specific preprocessing steps to prepare the data for analysis and modeling. Below are the detailed steps based on various research sources:

1. Handling Missing Values:

- The dataset did not have significant missing values. In cases where there were missing values, researchers typically used median imputation to fill these gaps, ensuring that no critical data was lost during preprocessing.

2. Removing Duplicates:

- Duplicate entries were identified and removed to ensure the integrity of the dataset. This step ensures that each patient record is unique, which is crucial for accurate modeling.

3. Encoding Categorical Variables:

- The categorical variables in the dataset, such as chest pain type (CP), resting electrocardiographic results, slope of the peak exercise ST segment (Slope), and thalassemia (Thal), were converted into numerical values using one-hot encoding. This method creates binary columns for each category, allowing the machine learning algorithms to process these features effectively.

4. Feature Scaling:

- Numerical features such as age, resting blood pressure (Trestbps), serum cholesterol (Chol), maximum heart rate achieved (Thalach), and ST depression induced by exercise relative to rest (Oldpeak) were standardized. Standardization involved scaling these features to have a mean of 0 and a standard deviation of 1, which helps in improving the performance of various machine learning algorithms.

5. Exploratory Data Analysis (EDA):

- EDA was conducted to understand the distribution and relationships between the features. For instance, histograms and bar plots were used to visualize the distribution of age, gender, chest pain type, and other attributes across patients with and without heart disease. This analysis highlighted that 68.32% of the records were male and 31.68% female, and provided insights into the distribution of chest pain types and fasting blood sugar levels among the patients.

6. Addressing Class Imbalance:

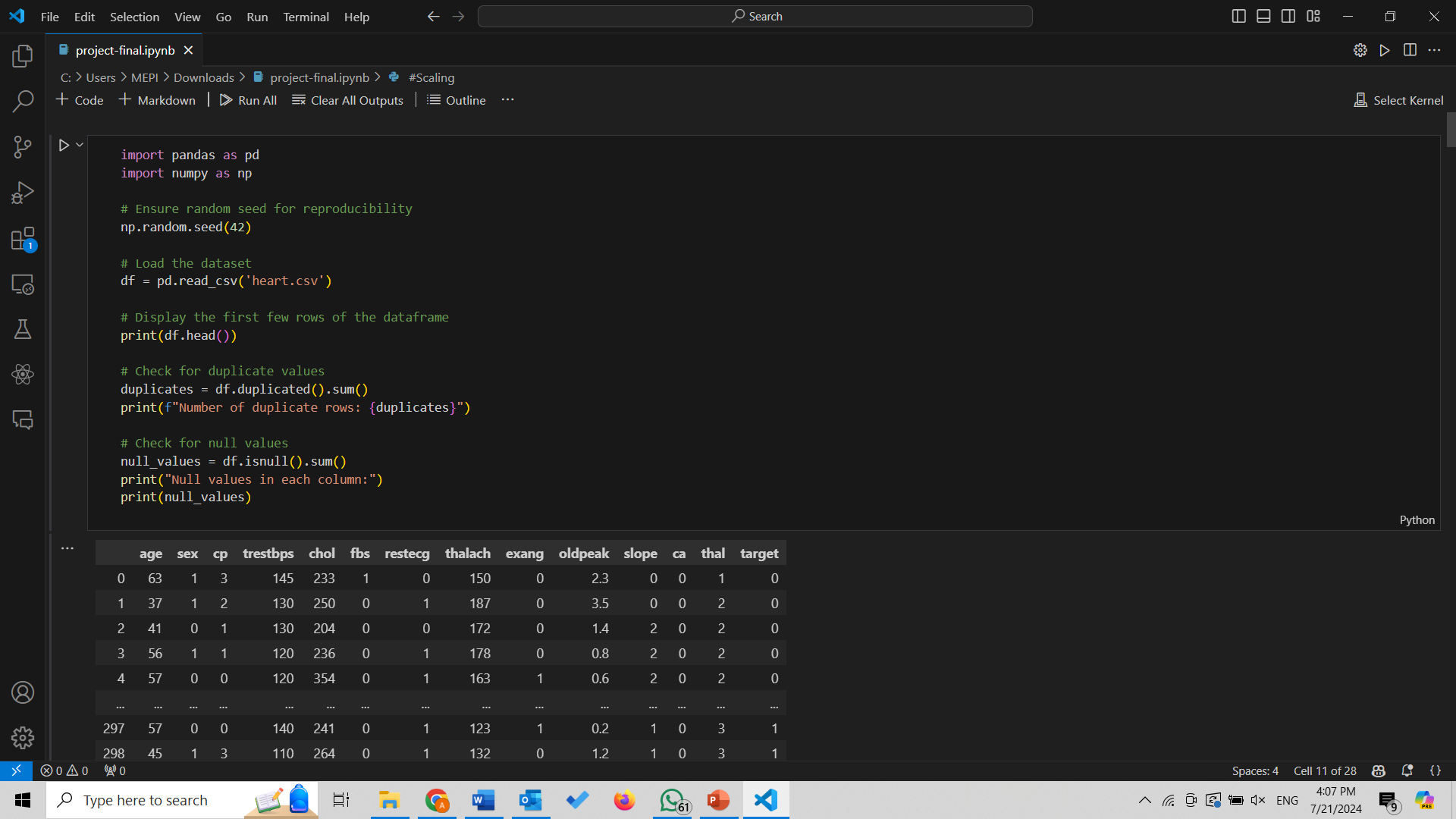
- The dataset had an imbalance in the target variable, with 138 records indicating no heart disease and 165 records indicating the presence of heart disease. To address this imbalance, techniques such as Synthetic Minority Over-sampling Technique (SMOTE) were applied. SMOTE generates synthetic samples of the minority class to create a balanced dataset, which helps in preventing the model from being biased towards the majority class.

# **Code**

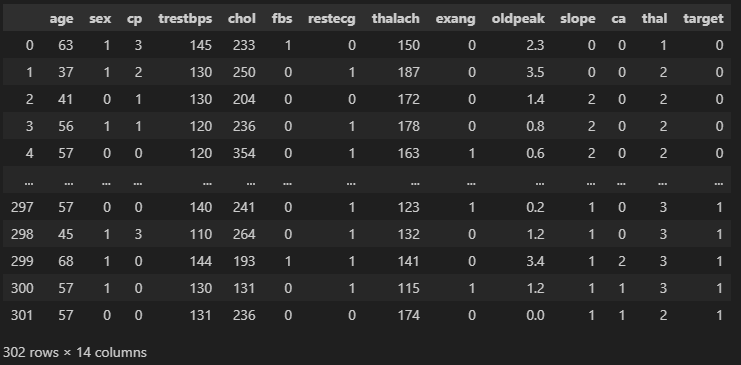
## **Data Preprocessing by us:**

Data preprocessing is a critical step in the machine learning workflow, ensuring the data is clean and suitable for model training. The process begins with loading the dataset, which in this case is a CSV file containing heart disease data. This dataset is read into a pandas DataFrame for further processing and analysis.

Code:

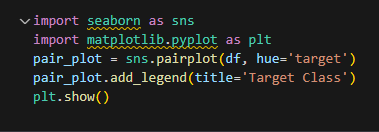


Results:

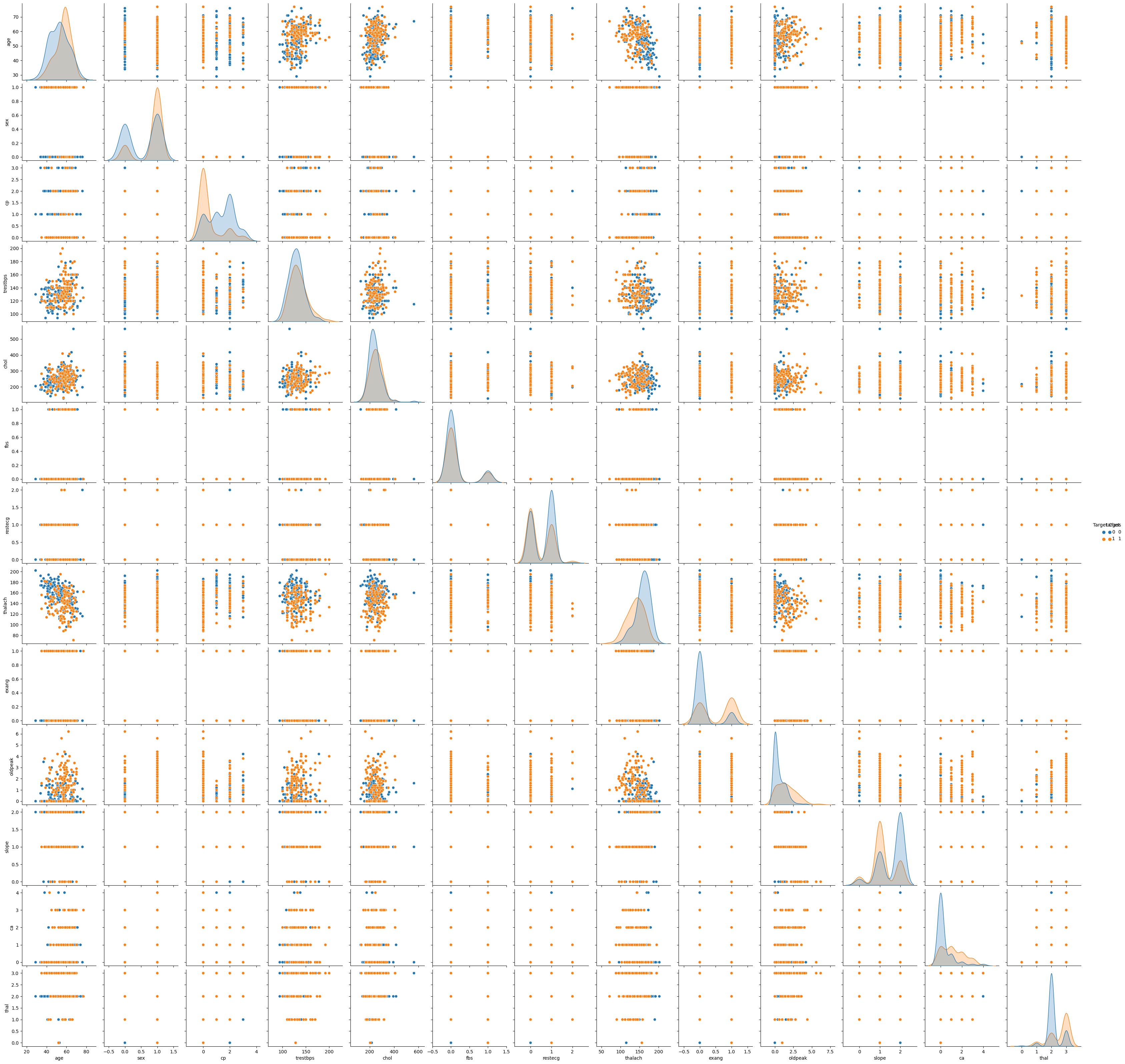


Exploratory Data Analysis (EDA) is performed to understand the data's structure and relationships. Pair plots are created using Seaborn to visualize the distribution and relationships between features, colored by the target variable.

Code:



Results:

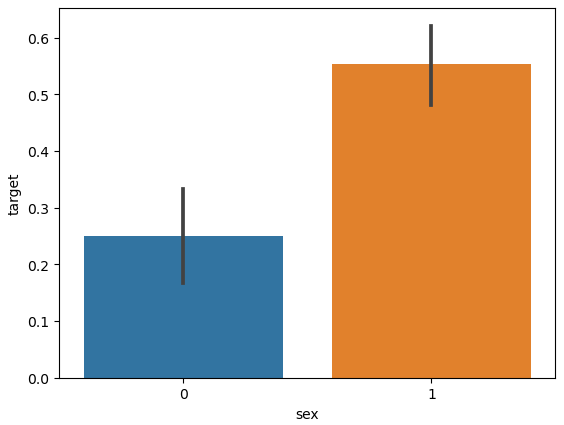


Next, a bar plot is generated to explore the relationship between the 'sex' feature and the target variable.

Code:

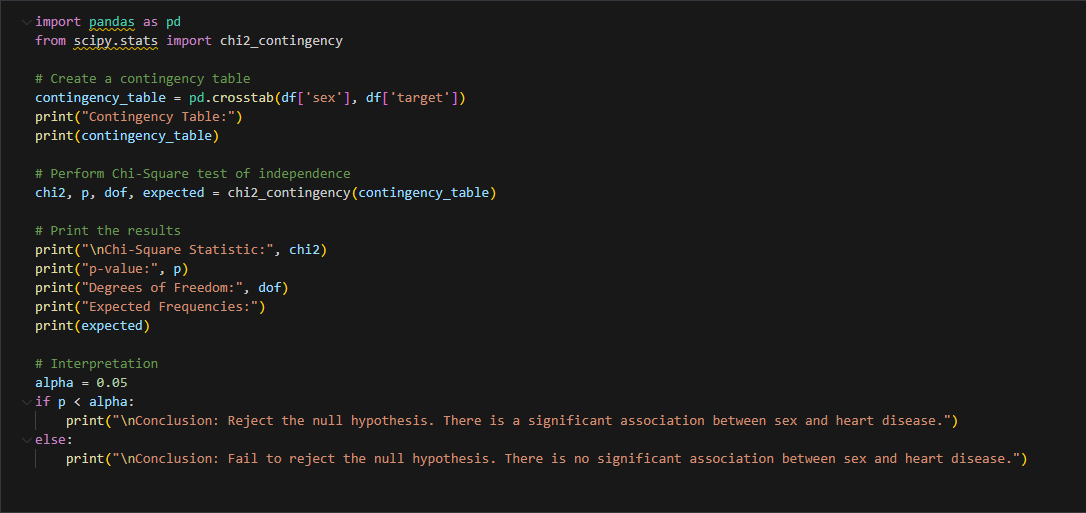


Result:

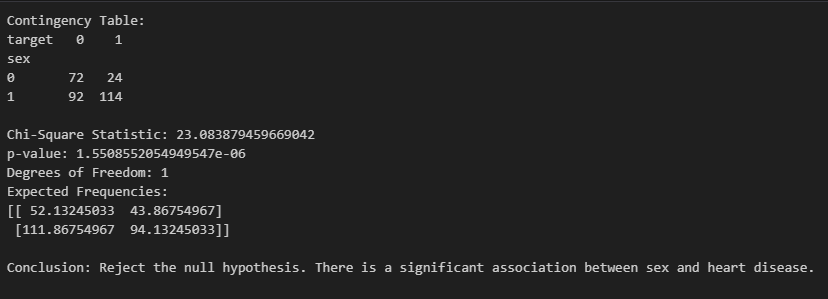


Statistical tests are conducted to examine associations between categorical variables. A Chi-Square test of independence assesses the relationship between 'sex' and the target variable.

Code:

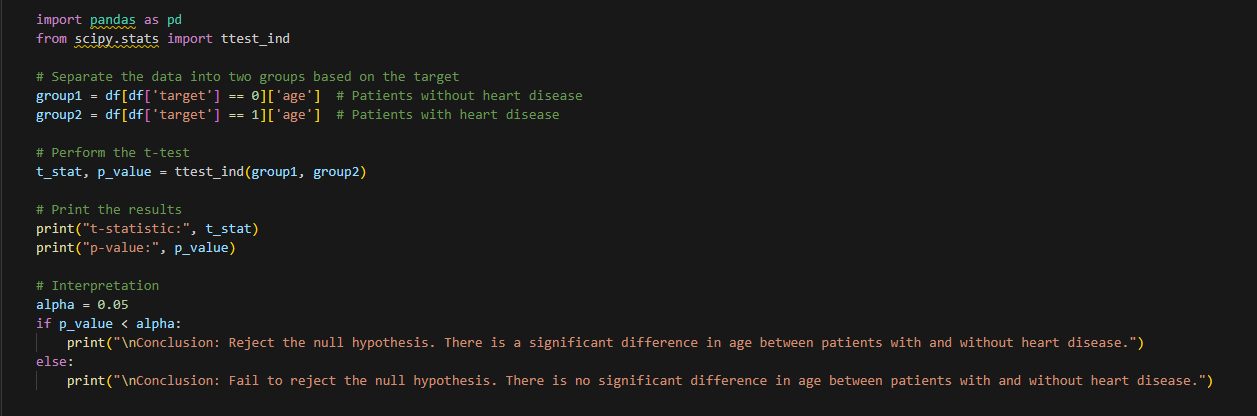


Result:

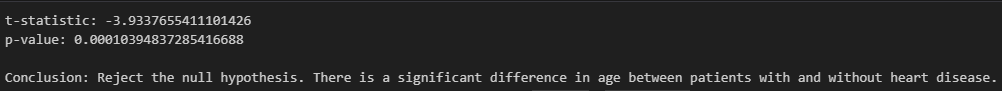


A t-test compares the ages of patients with and without heart disease to determine if there is a significant difference.

Code:



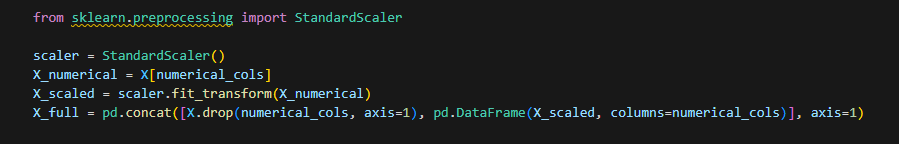
Result:

The numerical features are identified for scaling using the `StandardScaler` to ensure they are on a similar scale.

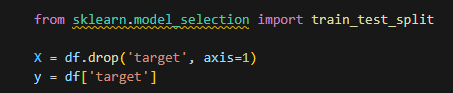
Code:

A screen shot of a computer code

Description automatically generated



Finally, the data is split into training and testing sets.





## **Choosing and Training of Models**

A variety of machine learning models are chosen to train on the data, including RandomForestClassifier, XGBClassifier, LGBMClassifier, SVC, and LogisticRegression. The models are trained on the training data using the fit method, learning the patterns in the data.

A computer screen with white text

Description automatically generated

This process is repeated for each model to ensure they all contribute to the final prediction.

Code:

A screenshot of a computer

Description automatically generated

## **Parameter Tuning (Optimization) and PCA**

Parameter tuning is executed using GridSearchCV and RandomizedSearchCV to find the optimal hyperparameters for each model. GridSearchCV performs an exhaustive search over specified parameter values, while RandomizedSearchCV samples a specified number of parameter settings from a given distribution. PCA is also applied to the different models, and accuracies are compared.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

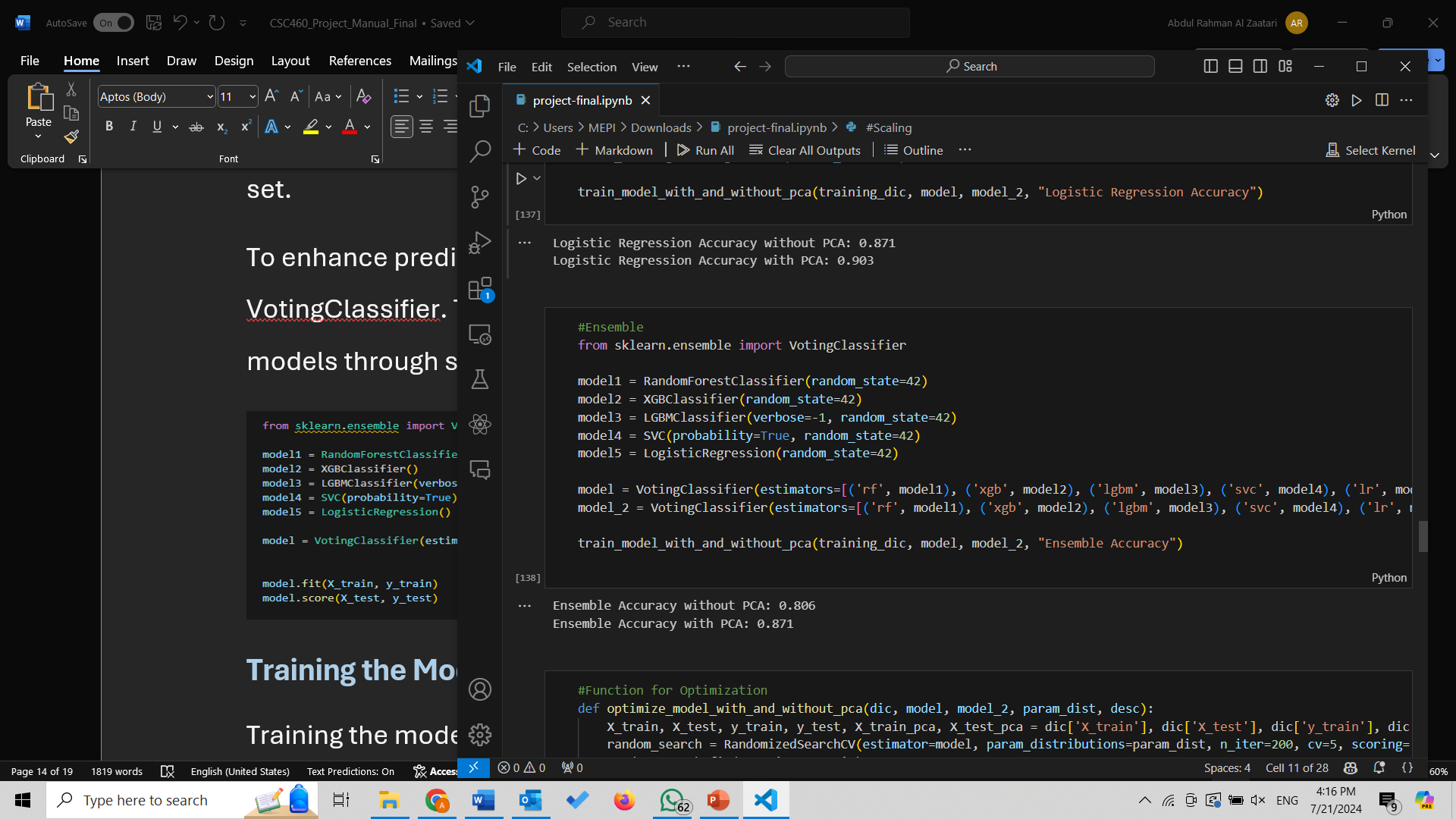
Description automatically generated

The best parameters found through the search are used to train the model, ensuring it has the optimal configuration for performance.

## **Comparing Accuracies and Reaching Ensemble**

The accuracy of each model is calculated using the score method on the test set.

To enhance prediction performance, an ensemble model is created using VotingClassifier. This model combines the predictions of the individual models through soft voting, which averages the predicted probabilities.



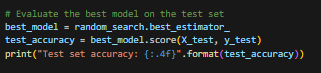
## **Evaluating the Models**

Model evaluation is performed using the score method, which calculates the accuracy of the model on the test set. This provides a direct measure of how well the model generalizes to unseen data.

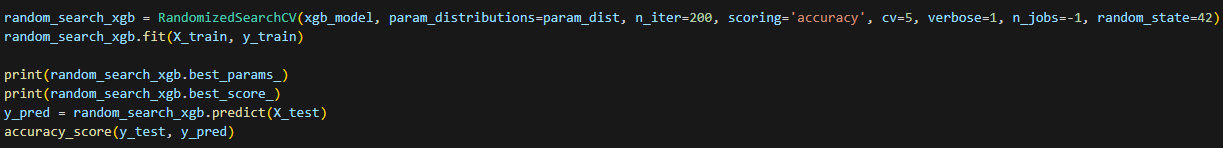
Additionally, cross-validation scores from parameter tuning offer insights into the model's performance across different data splits.

## **Making Predictions**

Predictions are made on the test set using the predict method, and the accuracy of these predictions is assessed using accuracy\_score, providing a quantitative measure of the model's performance.



The final ensemble model, after parameter tuning, is also used to make predictions, and its accuracy is compared against individual models to validate the improvement in performance.



# **Results:**

Those were the final results of the different models:  
  
Figure 1:

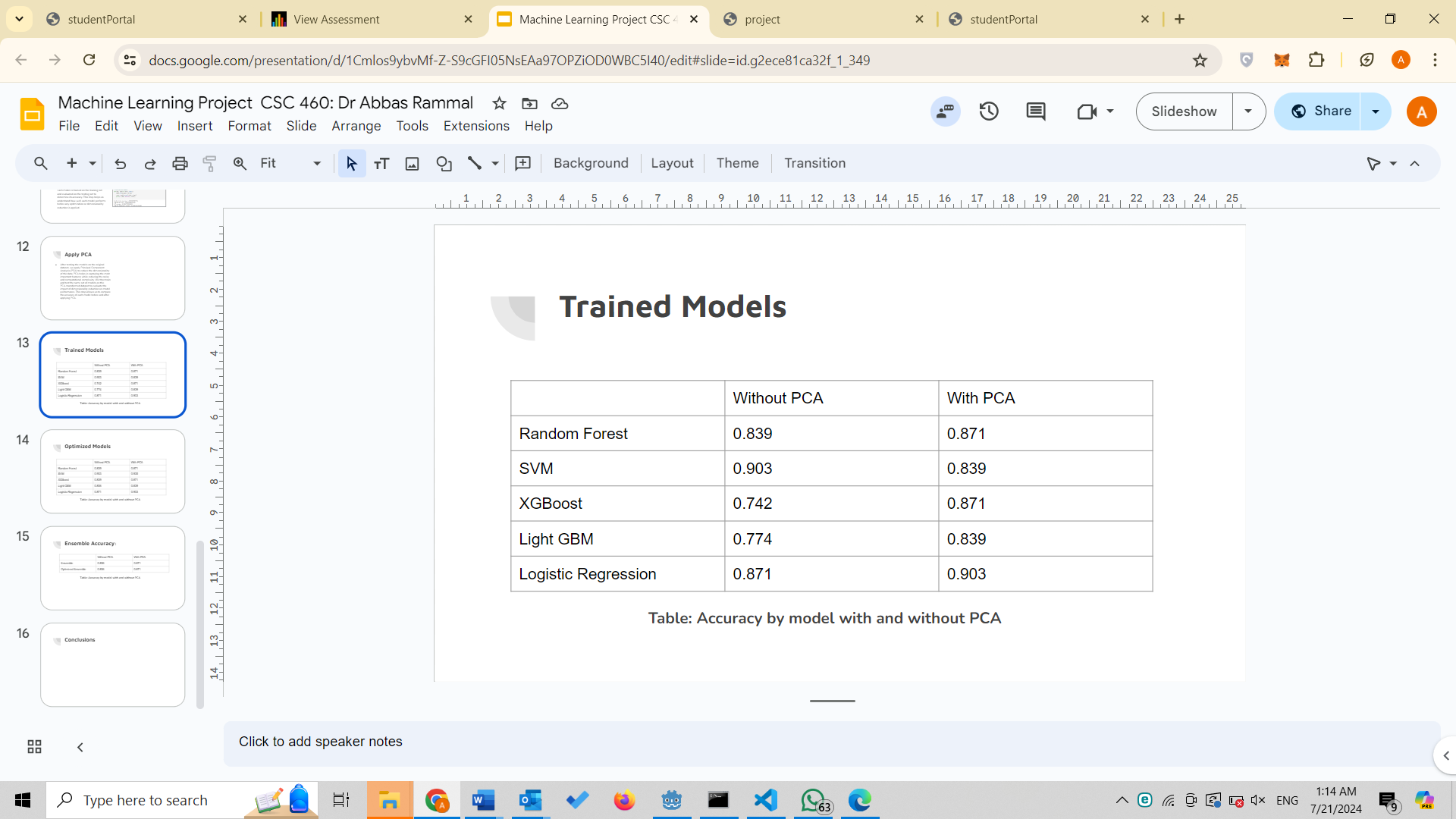


Figure 2:

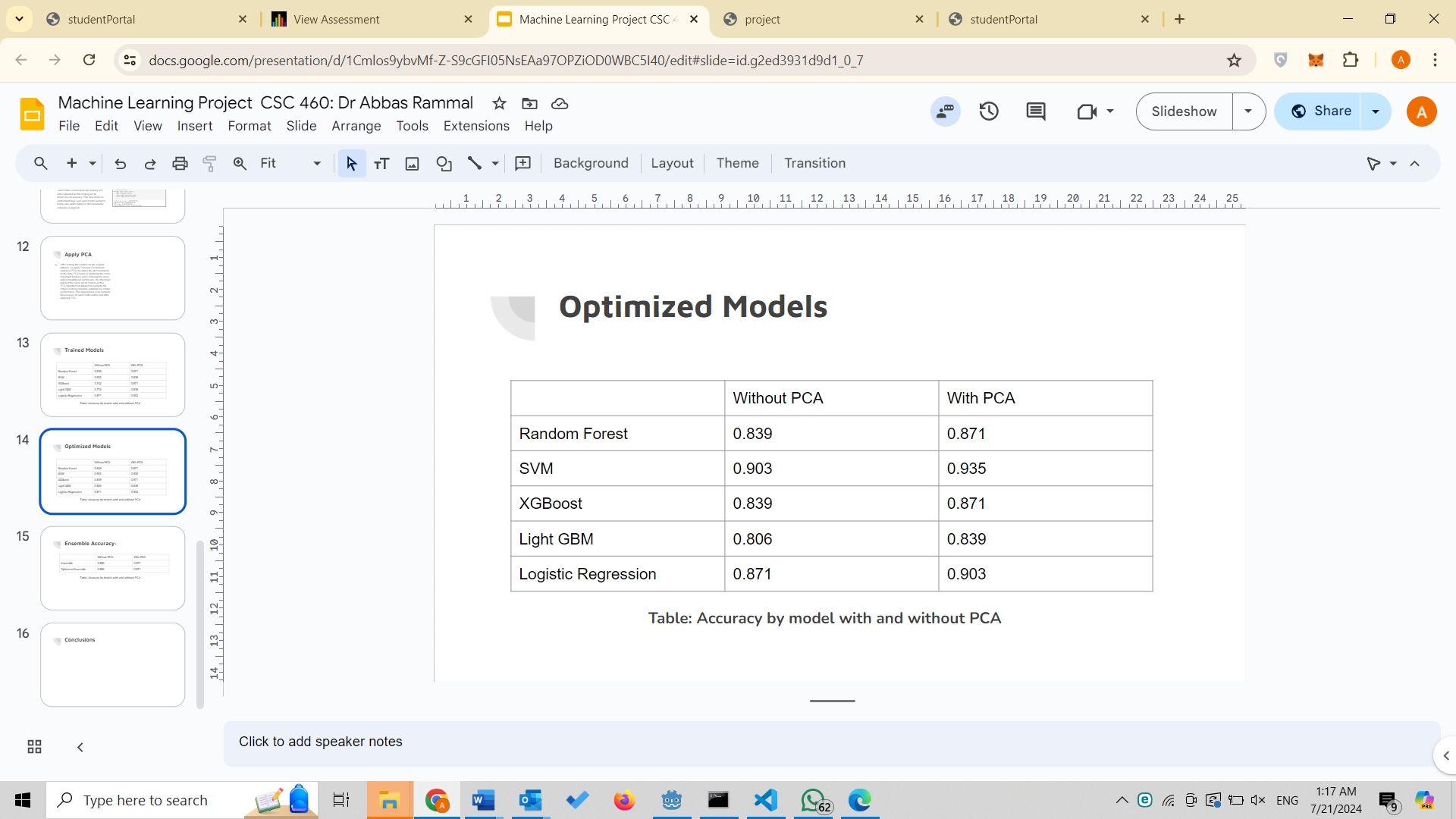
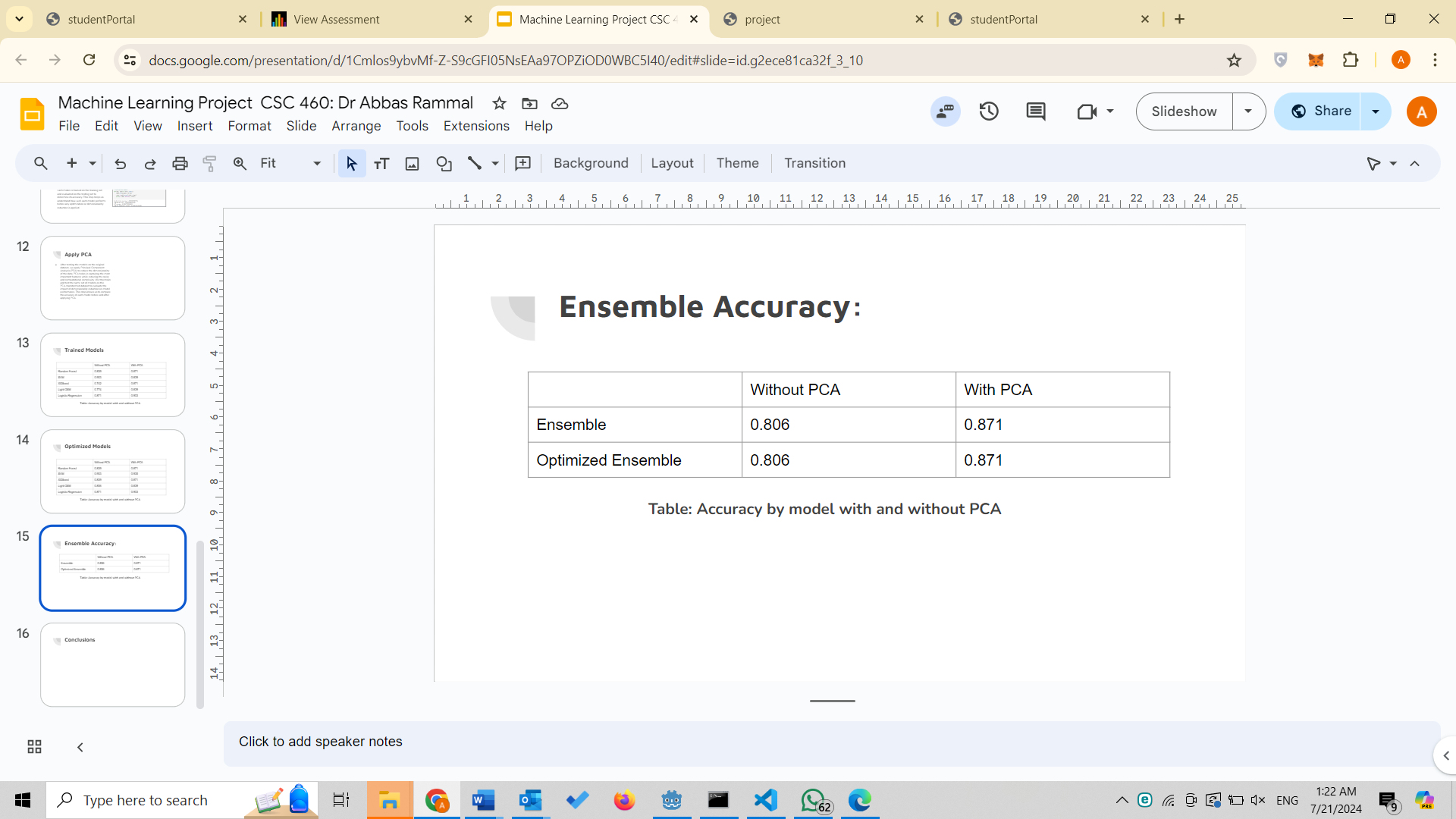


Figure 3:



**The model with the Highest Accuracy is optimized SVC with PCA, an outstanding accuracy of 0.935!**

**Conclusion:**This project demonstrates a comprehensive approach to building a predictive model for heart disease diagnosis using machine learning techniques. Starting with thorough data preprocessing and exploratory data analysis, we ensured the data was clean and ready for model training. Multiple models, including RandomForestClassifier, XGBClassifier, LGBMClassifier, SVC, and LogisticRegression, were trained and evaluated to identify the best-performing algorithms. Through rigorous parameter tuning using GridSearchCV and RandomizedSearchCV, we optimized the hyperparameters of these models to enhance their predictive accuracy. Ultimately, we constructed an ensemble model using VotingClassifier, leveraging the strengths of individual models to achieve superior performance. Although the ensemble model, validated through extensive evaluation, demonstrated improved accuracy in predicting heart disease, the best model was shown to be optimized SVC with PCA, being able to give an accuracy of 0.935!

# **Acknowledgements:**

We would like to extend our heartfelt gratitude to Dr. Abbass Rammal for his invaluable guidance and support throughout this courseS. His expertise and insights were instrumental in navigating the complexities of machine learning and data analysis. Additionally, we would like to thank the Lebanese American University (LAU) for providing the resources and environment conducive to academic and research excellence. The support and opportunities offered by LAU have been fundamental to the successful completion of this project.