Introduction/Background

🌟 Introduction / Background

Cardiovascular diseases (CVDs) affect the heart and blood vessels, leading to conditions like **heart attacks** and **strokes**. Heart disease is one of the most known and deadly diseases in the world [1]. That's why **early detection** is critical, and **machine learning** has shown promise in improving CVD diagnosis [2].

We propose using the publicly available **Cardiovascular Disease dataset**, with:

- **70,000 patient records**
- **11 features**, including **age, cholesterol**, and **blood pressure**

By applying machine learning techniques, we aim to identify patterns to predict **CVD presence** and enhance early detection efforts [3]. The dataset is available here 📊.

CVD early detection often fails with many symptoms going unnoticed, leading to late diagnoses and preventable complications. The lack of accessible and affordable preventative care leaves **underserved populations** at higher risk. Current prevention efforts are limited by the time and expertise needed for accurate early detection. We propose an **ML model** to analyze patient data and predict **CVD risks**, enabling earlier intervention, improving resource efficiency, and empowering individuals to make informed health decisions 💡.

Previous
References

Next
Problem Definition

Problem and Motivation

Problem & Motivation 💡

CVD early detection often fails with many symptoms going unnoticed, leading to late diagnoses and preventable complications. The lack of accessible and affordable **preventative care** leaves **underserved populations** at higher risk. Current prevention efforts are limited by the time and expertise needed for accurate early detection.

We propose an **ML model** to analyze patient data and predict **CVD risks**, enabling earlier intervention, improving resource efficiency, and empowering individuals to make informed health decisions 🌍✨.

Previous
Introduction

Next
Methods

Machine Learning Methods 🤖

🛠️ Data Preprocessing Methods

Our group used various data cleaning methods to adjust the database into a more model-friendly format. Below are the techniques we applied:

🗓️ Age Conversion

We adjusted the data to account for **leap years**, ensuring the accuracy of our dataset. While this change does not affect the results, it adds precision to the data representation.

🔢 Feature Encoding

- Initially applied **one-hot encoding** to separate the features of **Gender, Cholesterol**, and **Glucose** into individual columns. This approach simplifies clustering models like **K-Means**, which benefit from well-separated data [8].
- Later, we discovered this was unsuitable for **LogReg**, where distance matters (e.g., between normal and high cholesterol). Instead, we encoded these features as a range (1, 2, 3) for better compatibility.
- Gender was encoded as **0 (female)** or **1 (male)** for simplicity, while still treating it as categorical rather than binary.

🚫 Data Filtering

We applied filters to ensure **realistic values** for:

- **Height, Weight**, and **Systolic/Diastolic Blood Pressures**.
- Added human-based limits (e.g., Diastolic Blood Pressure set between **60 and 140**).
  Some rows contained impossible values (e.g., blood pressure over **1000**), which we excluded.

✅ Data Validation

In cases where the **Diastolic pulse** was higher than the **Systolic pulse**, we swapped the values (if valid). This ensures the dataset conforms to expected physiological norms, preventing issues during model training.

🗑️ Duplicate and Missing Row Deletions

- **Duplicates** were removed as unnecessary.
- Rows with missing values would have been dropped, though the dataset fortunately had none.

🔄 Binary Conversion

The **Smoking, Alcohol**, and **Active** fields, originally represented as **0s and 1s**, were converted to **true/false** values to optimize processing and model performance.

🔪 Scaling

We scaled features such as **Age, Height, Weight, and Systolic/Diastolic Blood Pressures** to enhance compatibility with models like **LogReg**.

🛠️ Feature Engineering

We engineered new features:

- **BMI** (Body Mass Index).
- **Pulse Pressure** (difference between Systolic and Diastolic pressures).

These features were added based on proposal suggestions to improve the dataset's predictive power.

title: Machine Learning Methods 🤖 description: Detailed description of the ML models used in the heart disease predictor project.

🤖 Machine Learning Methods

📈 Logistic Regression

For our first methodology, we implemented **Logistic Regression** on the cleaned data. **LogReg** is a good choice for checking if someone has a **CVD** because:

- **LogReg is specifically designed for binary classification problems**, where the goal is to predict a binary problem. In our project, we identify the two outcomes as such:
  1. **The person has the condition**, or
  2. **The person does not have the condition**.
- **LogReg outputs probabilities** (values between 0 and 1) that represent the likelihood of an observation belonging to a certain class (e.g., the probability that someone has the condition). This makes it well-suited for decision-making where you might want to know not just the predicted class, but also the **confidence level of the prediction**. For instance:
  - A confidence level of **40%** may mean the doctors may want to do further screening and/or discuss preventive measures to ensure the patient's well-being.
  - A confidence level of **90%** means the doctor definitely wants to start taking steps for the patient's health.
- **LogReg uses the logistic function** to map any real-valued number into a probability. The logistic function outputs values between **0 and 1**, which works great for classification because it means we can interpret the output as a probability.
- **LogReg models are relatively easy to interpret**, especially when considering the coefficients. They represent the change in the odds of the outcome for a change in a variable we could modify, holding all other variables constant. This allows us to understand the influence of each feature on the outcome. For example, if we wanted to see how **glucoses levels** were correlated with **alcohol**, we can use a visualizer like a **heatmap** to see how close the two variables are.

Implementation 💻

We implemented **two codes** for our logistic regression to predict cardiovascular disease (CVD), but with distinct methodologies to improve model performance and validation.

- **The first code** uses **PCA** with two components for dimensionality reduction, enabling basic visualization but without handling class imbalance or hyperparameter tuning.
- **In contrast, the second code** uses **SMOTE** to address class imbalance, applies **GridSearchCV** for hyperparameter tuning, and includes an **SGD classifier** to compare model performance. Additionally, **PCA** in the second code dynamically retains **95% variance** for efficiency.

These variations provide a more comprehensive evaluation of logistic regression and alternative techniques, aiding in model validation and robustness.

Part A - Logistic Regression with Fixed PCA for CVD Prediction

This code performs logistic regression to predict cardiovascular disease (CVD) using various health-related features from a dataset. Initially, it preprocesses the data by:

- **Removing unnecessary columns**.
- **Splitting** it into features and target variables.
- **Converting boolean columns**.
- **Standardizing the data** with **StandardScaler** to prepare it for model training.

For dimensionality reduction, it applies **PCA (Principal Component Analysis)**, reducing the data to **two components**, which aids in visualization and helps compare model performance with and without dimensionality reduction.

The code then trains **two logistic regression models**:

1. **One on the data reduced by PCA**.
2. **Another on the complete set of standardized features**.

To evaluate model performance, it calculates several metrics, including:

- **Accuracy**
- **Precision**
- **Recall**
- **F1 score**
- **Confusion Matrix**
- **ROC-AUC score**

For visualization, the code plots:

- **PCA results**
- **Heatmaps** to show feature correlations
- **Confusion matrices**
- **Receiver-Operating-Characteristics (ROC) Curves**

This makes it easier to interpret results.

In addition, the code conducts **feature analysis** by calculating and plotting feature importance using the logistic regression coefficients, odds ratios, and permutation importance, helping identify the impact of individual features on the prediction of cardiovascular disease.

Part B - Advanced CVD Prediction with SMOTE, Tuned PCA, and SGD 🏋️

This code uses logistic regression and an **SGD (Stochastic Gradient Descent) classifier** to predict cardiovascular disease (CVD) based on a range of health-related features. It begins by:

- **Loading a dataset**, defining the target variable (**cardio**), and excluding unnecessary columns.
- The selected features include **age, gender, blood pressure measures, and lifestyle indicators**.
- After ensuring boolean columns are integer-based, the data is **split into training and test sets**.

To handle class imbalance, **SMOTE (Synthetic Minority Oversampling Technique)** is applied to the training set, creating a balanced sample for the minority class. The features are then:

- **Standardized** using **StandardScaler**.
- Followed by **PCA** to reduce dimensionality while retaining **95% of the variance**.

A logistic regression model is trained with **hyperparameter tuning** using **GridSearchCV** to optimize the regularization parameter. Model performance is assessed on the test set, calculating:

- **Accuracy**
- **Precision**
- **Recall**
- **F1 score**
- **Confusion Matrix**

Visualization via a **heatmap** and **ROC curve** provides deeper insights.

The code then trains an **SGD classifier** with a logistic loss function as an alternative, followed by similar performance evaluation and visualization steps, including:

- **Accuracy**
- **Precision**
- **Recall**
- **F1 score**
- **ROC-AUC scores**

Additional visualizations include:

- **Heatmaps** of feature correlations
- **Confusion matrices** for both models
- **ROC curves**

This provides a comprehensive view of model performance and feature relationships.

title: Decision Tree 🌳 description: Detailed description of the Decision Tree model used in the heart disease predictor project.

🌳 Decision Tree

The second model this project utilizes is a **decision tree model** to predict the presence of **cardiovascular disease** based on key medical and lifestyle features. Its ability to provide clear decision-making paths and handle complex interactions between features makes it an effective choice for this task.

- The model's strength lies in its **interpretability**, as its decision-making process can be visualized and understood by stakeholders without technical expertise. This is particularly valuable in a medical context where **trust** in the model is critical.
- **Non-linear relationships** among features are captured naturally, which is important since cardiovascular disease risk factors often interact in complex ways.
- **No need for feature scaling** simplifies the preprocessing pipeline while preserving the natural units of measurement for medical data.
- Handling **mixed data types** allows the model to work seamlessly with datasets containing both numerical and categorical features, such as blood pressure levels, cholesterol categories, and lifestyle habits like smoking or alcohol consumption.
- **Focus on relevant features** makes the model robust to irrelevant or redundant information in the dataset.
- **Flexibility** makes it effective as both a standalone model and as part of ensemble methods like **Random Forests** or **Gradient Boosting**, which can enhance predictive performance.

🛠️ Implementation Process

The implementation began with **preprocessing the data** to prepare it for model training. Relevant features were selected based on their importance in predicting cardiovascular disease, including factors like **age, gender, blood pressure**, and **lifestyle habits**. Boolean features were explicitly converted into integer representations to ensure compatibility with the decision tree model. The dataset was then **split into training and testing subsets** using an **80-20 ratio** to train the model and evaluate its performance on unseen data.

Once the preprocessing was complete, a **decision tree classifier** was initialized with specific hyperparameters. The tree depth was limited to **five levels** to prevent overfitting and ensure the model remained interpretable. Additionally, a **minimum number of 20 samples per split** was enforced to ensure statistically meaningful decisions at each node. The model was then **trained on the training data** to learn the patterns and relationships between the features and the target variable, which indicated the presence of cardiovascular disease.

After training, the model's performance was evaluated on the test data using standard classification metrics. **Predictions were made for the test set**, and metrics such as **accuracy, precision, recall**, and **F1-score** were calculated. A **confusion matrix** was generated to provide a detailed breakdown of true and false predictions. The decision tree was **visualized** to interpret its decision paths, and a **heatmap of the confusion matrix** was created for better visual representation of the classification results. This comprehensive approach ensured both robust model evaluation and interpretability.

🌲 Random Forest

The final model for this project uses a **random forest** to predict the presence of **cardiovascular disease**. A random forest is an **ensemble learning method** that combines multiple decision trees to make predictions and in the context of CVD, it is a good choice for the following reasons:

- **By aggregating the outputs from a diverse set of decision trees**, it can reduce the likelihood of overfitting compared to a single decision tree. This helps to ensure that the model doesn't overly rely on patterns that might only be present in a subset of the data, leading to better generalization and improved accuracy.
- **Differences in patient characteristics** often result in noise and variability. Random Forests average the predictions of multiple trees, meaning that outliers or noisy data points are less likely to overly influence the model's decision-making process.
- **They can easily capture non-linear relationships**, making them more effective than linear models when dealing with features that may have complex relationships. This allows to better model the interactions between multiple features, leading to improved predictive performance.
- **They are inherently good at capturing interactions between features without the need for explicit feature engineering**. This ability to model interactions implicitly is advantageous when dealing with medical data, where multiple risk factors often work in tandem to influence outcomes.

🛠️ Implementation Process

The data was first **preprocessed** in order to prepare it for the specific model. The data was then **split into training and testing datasets** with an **80-20 ratio**. Once complete, a **random forest classifier** was implemented with a depth of **5** to compare results to that of the decision tree, and depths of **3, 6, 9, and 12** were also explored to see how the model behaved.

After training, the model's performance was evaluated on the test data. Metrics such as **accuracy, precision, recall**, and **F1-score** were calculated. The results also allowed for a visualization of **permutation feature importance**, which gives a better idea on what features contribute to the results the most. Overall these are a robust set of results to draw conclusions from.

Results and Discussion

Results 🏆

Logistic Regression

Both **Part A** and **Part B** reveal similar feature relationships in predicting **CVD**. Strong correlations appear between **systolic (ap_hi)** and **diastolic (ap_lo)** blood pressure, as well as between **blood pressure** and **pulse pressure**. Minor differences in correlation strengths suggest variations due to different preprocessing methods, like **SMOTE** in Part B. **Blood pressure** and **cholesterol** emerge as key predictors, while **lifestyle factors** show weaker correlations.

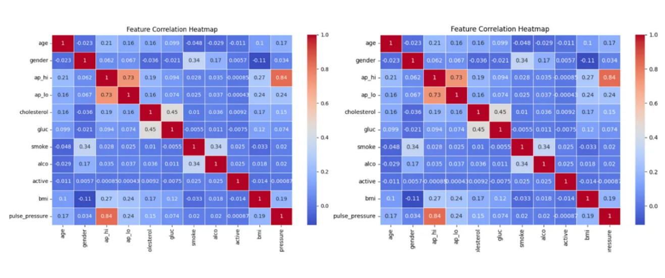


Figure 1. Feature Correlation Heatmap (Part A to the left, Part B to the right)

Part A 🔷

The results highlight the model's performance in predicting **CVD**. **PCA** shows data distribution, with **blood pressure correlations** prominent. **Confusion matrices** and an **AUC** of **0.79** indicate moderate accuracy. Key predictors are **ap_hi**, **ap_lo**, **cholesterol**, and **age**, marking them as crucial for assessing **CVD risk**.

Figure 2. PCA of Training Data

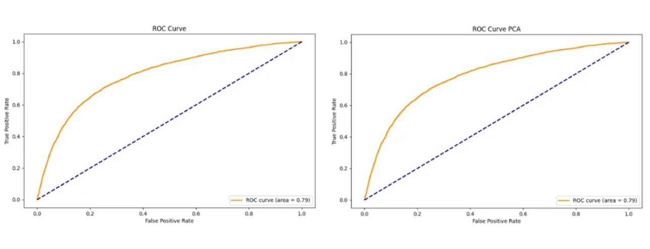


Figure 3. ROC Curve (without, left and with PCA, right)

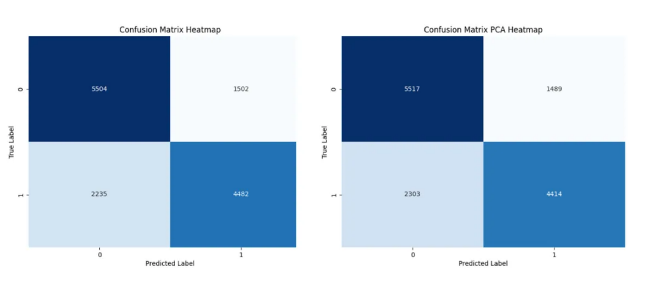


Figure 4. Confusion Matrix Heatmap (without, left and with PCA, right)

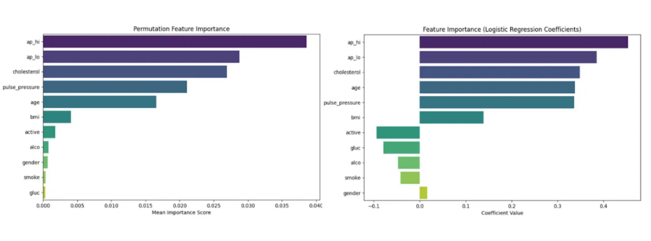


Figure 5. Permutation and Feature Importance

Part B 🔶

Part B results show **logistic regression** and **SGD classifier** models effectively predicting **CVD**. The **feature correlation heatmap** highlights strong relationships, particularly among **blood pressure indicators**, with **cholesterol** and **age** as key **CVD predictors**. **Confusion matrices** demonstrate similar classification performance for both models, while **ROC curves** show moderate discriminatory power, with **AUCs** of **0.79** for logistic regression and **0.78** for SGD. **Blood pressure** and **cholesterol** emerge as primary factors in assessing **CVD risk**.

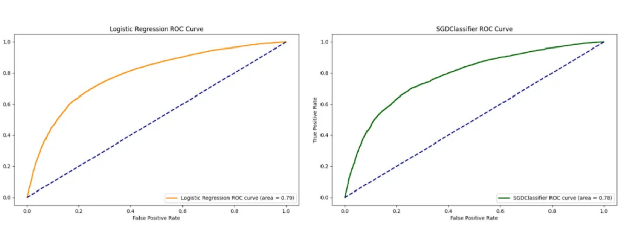


Figure 6. Confusion Matrix Heatmap Part B

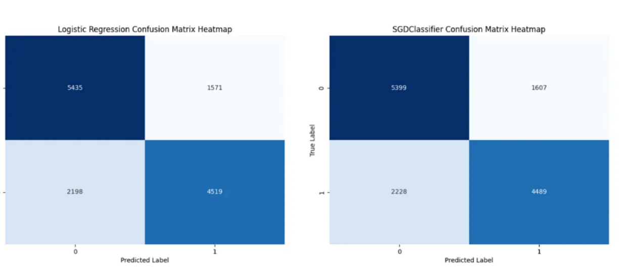


Figure 7. ROC Curve Part B

Decision Tree

The decision tree model achieved an **accuracy** of **72.84%**, demonstrating its ability to correctly classify the majority of instances in the dataset. The **precision** was **75.15%**, indicating that when the model predicted the presence of cardiovascular disease, **75%** of the predictions were correct. The **recall** was **66.50%**, suggesting that the model successfully identified about two-thirds of the actual cases of cardiovascular disease. Finally, the **F1-score**, a balance between precision and recall, was **70.56%**, reflecting the model's overall effectiveness in handling both false positives and false negatives. The **confusion matrix heatmap** visually highlights these results, showing the distribution of **true positives**, **true negatives**, **false positives**, and **false negatives**. Additionally, the **decision tree visualization** provides clear insights into the splitting criteria and feature importance used to make predictions.

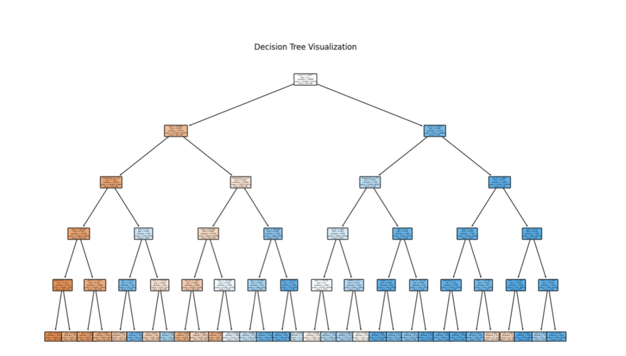


Figure 8. Decision Tree Visualization

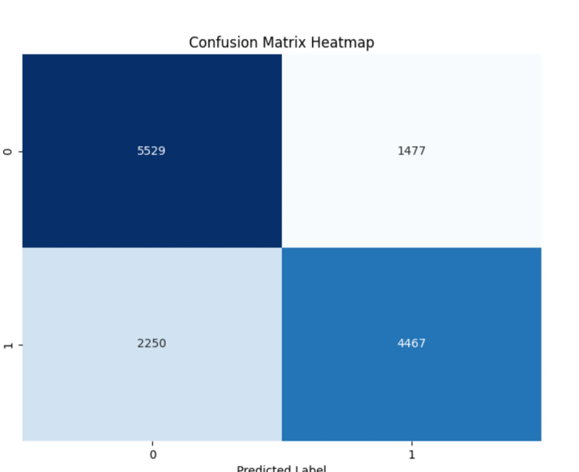


Figure 9. Decision Tree Confusion Matrix Heatmap

Random Forest

The random forest model was evaluated at different **tree depths (3, 6, 9, and 12)** with and without **Principal Component Analysis (PCA)** for dimensionality reduction. At a depth of **3**, the model achieved an **accuracy** of **72%** for both PCA and non-PCA, with **precision** values of **76% (PCA)** and **75% (non-PCA)**. **Recall** was **64%** for both settings, resulting in an **F1-score** of **70%** for PCA and **69%** for non-PCA. Increasing the depth to **6** led to an **accuracy** of **73%**, with precision slightly higher for non-PCA (**76%**) compared to PCA (**75%**). The recall increased to **66%** for both configurations, and the F1-score remained at **70%**. At depths **9** and **12**, both PCA and non-PCA models reached similar performance levels, with an **accuracy** of **73%**, precision around **75-76%**, recall at **66-68%**, and an **F1-score** of **71%**. The **confusion matrix heatmaps** and **ROC curves** were also plotted for each depth which can be observed below:

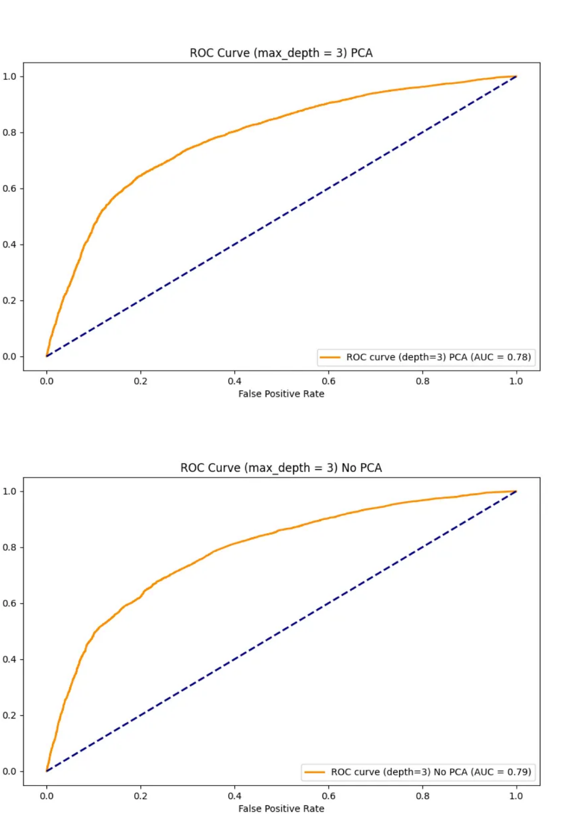


Figure 10. Comparison of ROC Curves for depth 3 with and without PCA

```
PCA, as shown in Figure III, did not change the results of testing that much, aff
```

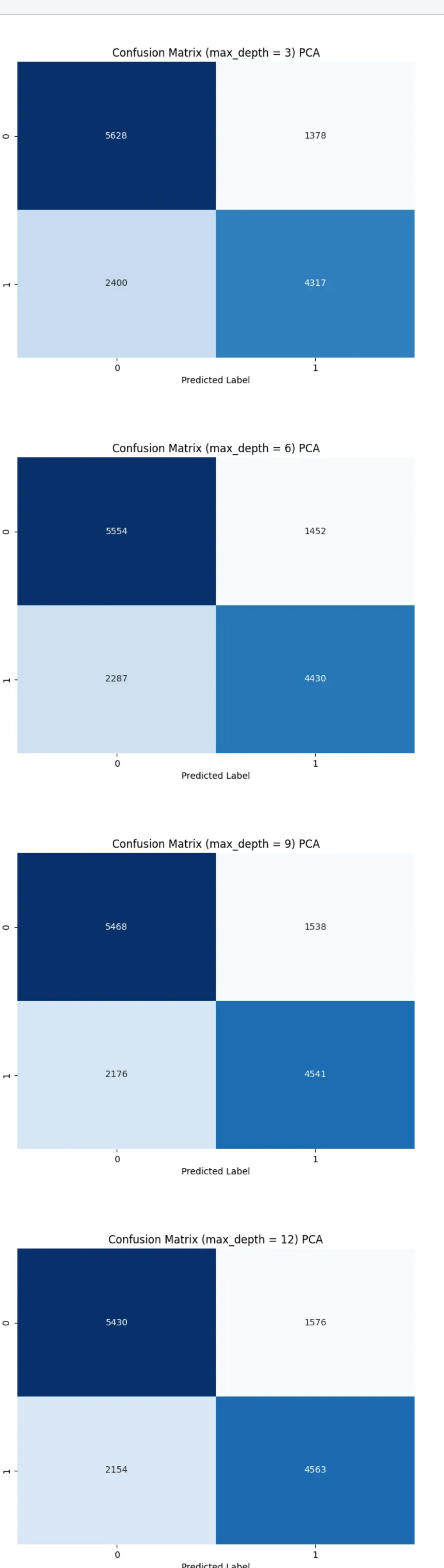


Figure 11. Comparison of Confusion Matrices for varying depths without PCA

An interesting note is that with PCA, the model's feature importances change from prioritizing systolic blood pressure to heavily prioritizing age, as shown below:

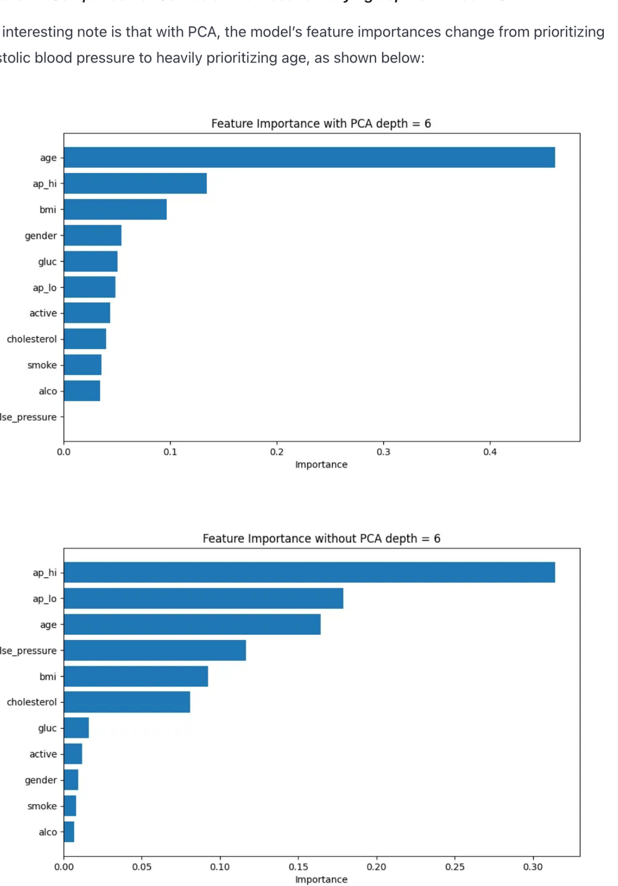


Figure 12 Comparsion of Feature Importances of Depth 6 Forests

Discussions

Logistic Regression

Logistic Regression with Fixed PCA for CVD Prediction 📊

The logistic regression models show similar performance, with the model without PCA performing slightly better (**accuracy 0.73, precision 0.75, recall 0.67, F1 score 0.71**) compared to the PCA model (**accuracy 0.72, precision 0.75, recall 0.66, F1 score 0.70**). The confusion matrices are also similar, indicating consistent classification results.

The odds ratios reveal that **ap_hi (1.57)** and **ap_lo (1.47)** are the strongest predictors of CVD, followed by **cholesterol (1.42)** and **age (1.40)**, indicating that higher values increase CVD risk. In contrast, factors like **smoke (0.96), alco (0.95), gluc (0.92)**, and **active (0.91)** show weaker or potentially protective effects. This suggests that **blood pressure and cholesterol** are more influential in predicting CVD than lifestyle factors.

Advanced CVD Prediction with SMOTE, Tuned PCA, and SGD 🚀

The results indicate similar performance between the **logistic regression** and **SGD classifier** models for predicting cardiovascular disease (CVD). The logistic regression model achieved an accuracy of **0.73**, with a precision of **0.74**, recall of **0.67**, and F1 score of **0.71**. Its confusion matrix shows **5,435 true negatives, 1,571 false positives, 2,198 false negatives**, and **4,519 true positives**.

The **SGD classifier**, on the other hand, had a slightly lower accuracy of **0.72**, with the same precision of **0.74**, recall of **0.67**, and a marginally lower F1 score of **0.70**. Its confusion matrix shows **5,399 true negatives, 1,607 false positives, 2,228 false negatives**, and **4,489 true positives**.

Overall, the logistic regression model performed marginally better in terms of accuracy and F1 score, while precision and recall remained the same for both models. The SGD classifier exhibited a slightly higher rate of false positives and false negatives, which impacted its overall accuracy and F1 score. This comparison suggests that logistic regression may offer a slight edge in predictive accuracy for this dataset, though both models perform similarly in terms of precision and recall, providing consistent results across different metrics.

Decision Tree

The results indicate that the decision tree model is moderately effective for predicting cardiovascular disease. The high precision (**75.15%**) is valuable in minimizing false positives, which is crucial in medical applications to avoid unnecessary testing or treatments for individuals not at risk. However, the recall (**66.50%**) suggests that the model missed approximately one-third of the true cases, which could lead to underdiagnosis. This trade-off is reflected in the F1-score (**70.56%**), which balances the model's precision and recall.

The confusion matrix highlights a considerable number of false negatives, which may warrant further optimization to improve recall. The Decision Tree model identified **systolic blood pressure**, **diastolic blood pressure**, and **cholesterol levels** as the most significant features influencing predictions. These findings align closely with clinical knowledge, as these metrics are well-known indicators of cardiovascular health. Additionally, lifestyle factors such as **smoking** and **physical activity** played a smaller role, indicating that the model prioritized physiological measures over behavioral data when making decisions.

Increasing the tree's depth or employing ensemble methods, such as **random forests**, could address this issue by capturing more complex patterns in the data. The decision tree visualization emphasizes the interpretability of the model, allowing stakeholders to understand the criteria driving predictions. Despite its limitations, the model's performance is a good starting point, providing actionable insights into the factors influencing cardiovascular disease risk. Future iterations could focus on enhancing recall to ensure more comprehensive detection of at-risk individuals.

Random Forest

The results demonstrate that the random forest model is effective in predicting cardiovascular disease with an accuracy of **72%-73%**, which is above the threshold accuracy for this project. The precision of around **75-76%** for all models suggests that the model was able to minimize false positives, which is crucial in medical contexts to reduce unnecessary interventions. However, the recall, ranging from **64% to 68%**, suggests that the model still missed a notable proportion of actual CVD cases, especially at lower depths. This trade-off between precision and recall is balanced with an F1-score of around **70-71%**, indicating an overall stable performance across all settings. The confusion matrices for each depth show that false positives were generally reduced as depth increased, and more true positives were captured.

Furthermore, it is important to consider the impact of using PCA on the model, as it did not significantly impact its performance. The accuracy, precision, recall, and F1-score were very similar between PCA and non-PCA models. However, PCA led to a slight improvement in recall, particularly at higher depths. This suggests that PCA may help reduce noise in the data, allowing the random forest to focus on the most important features.

The use of PCA also resulted in fewer false positives and more true positives, as shown by the confusion matrices. Despite the similar overall performance, the use of PCA could be useful in reducing computational load and enhancing model interpretability without sacrificing accuracy.

🌟 Overall

The overall results from this project indicate that machine learning techniques such as **Logistic Regression**, **Decision Trees**, and **Random Forests** are viable options for predicting cardiovascular disease (CVD) based on health and lifestyle features. Each model demonstrated strengths in terms of accuracy and precision, but also highlighted some weaknesses, particularly in recall, which is crucial for minimizing false negatives in a medical context. The use of **PCA** for dimensionality reduction had mixed effects across the different models, with slight improvements in recall but no significant changes in overall accuracy or precision.

Logistic Regression

Logistic Regression was effective in providing interpretability and producing probabilities for each prediction, making it a useful model for risk assessment in clinical settings. However, it suffered from slightly lower recall, particularly when class imbalance was not addressed. The advanced model with **SMOTE** and **hyperparameter tuning** yielded better-balanced results, reducing false negatives and improving overall robustness. Logistic Regression's **feature importance analysis** provided valuable insights, emphasizing **blood pressure** and **cholesterol** as key indicators of CVD risk.

Decision Trees

Decision Trees offered interpretability, allowing stakeholders to understand the decision-making process clearly. The model's precision was relatively high, minimizing false positives, but the recall was moderate, indicating missed CVD cases. This trade-off suggests that while Decision Trees are informative and useful for preliminary assessments, they might require enhancements—such as increased depth or ensemble methods—to fully capture the complexity of CVD risk factors and reduce underdiagnosis. The visualization of decision paths helped in identifying how individual health metrics contributed to the final predictions, underscoring the transparency of this approach.

The Decision Tree model's interpretability makes it particularly suited for applications in clinical settings. By providing clear and understandable decision paths, it can assist healthcare professionals in explaining CVD risk assessments to patients. This transparency not only aids decision-making but also fosters trust in machine learning tools as part of the diagnostic process. Furthermore, its ability to highlight critical risk factors could support preventative strategies by enabling targeted interventions.

Random Forests

Random Forests proved to be the most effective model in terms of balanced accuracy, precision, and recall. The ensemble nature of Random Forests provided robustness against noise and variability in patient data, capturing non-linear relationships among features. The model performed consistently across different depths, and while increasing depth led to minor improvements in recall, it also risked overfitting. **PCA** had a slight positive impact on recall, reducing noise and improving sensitivity without significantly affecting accuracy or precision. The **feature importance analysis** from Random Forests provided insights into the factors most critical for predicting CVD, reaffirming the importance of **blood pressure, cholesterol, and age**.

Conclusion

In conclusion, the results indicate that **Random Forests** are the most promising model for predicting cardiovascular disease, offering a good balance between interpretability, accuracy, and sensitivity. However, the recall values across all models suggest a need for further improvement to ensure comprehensive detection of at-risk individuals. Future work could include **hyperparameter tuning for Random Forests**, incorporating additional data sources, and exploring more advanced ensemble techniques to improve recall and overall predictive performance. Additionally, more attention to reducing false negatives will enhance early intervention efforts, ultimately improving patient outcomes.

← Previous Methods
Next References →

References & Additional Information

📚 References

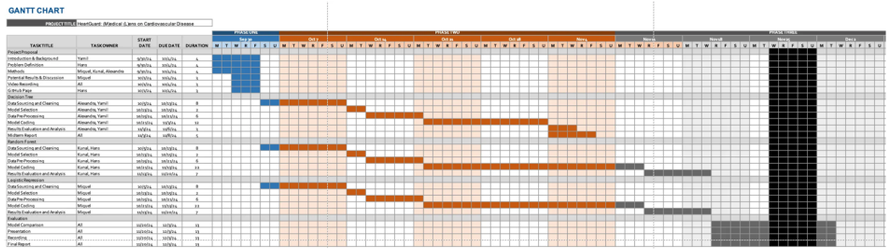
[1] Ahmad, I., & Kalimullah, M. (2021). *Cardiovascular disease prediction using data mining techniques: A review*. International Journal of Advanced Computer Science and Applications, 12(1), 180-186. https://doi.org/10.14569/IJACSA.2021.0120124

[2] A. Ahmad and H. Polat, *Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm*, Diagnostics, vol. 13, no. 14, pp. 2392–2392, Jul. 2023, doi: https://doi.org/10.3390/diagnostics13142392

[3] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, *Effective Heart Disease Prediction Using Machine Learning Techniques*, Algorithms, vol. 16, no. 2, p. 88, Feb. 2023, doi: https://doi.org/10.3390/a16020088

📝 Additional Information

Gantt Chart



👥 Contributions

- **Alexandra** - Decision Tree implementation, Results and Discussion
- **Hans** - GitHub Web Page, Data Preprocessing
- **Kunal** - Random Forest implementation, Results and Discussion
- **Miguel** - Logistic Regression implementation, website
- **Yamil** - Data Preprocessing, GitHub Web Page

Previous
Results and Discussion