BIRZEIT UNIVERSITY

Faculty of Engineering and Technology
Electrical and Computer Engineering Department

ENCS5341 Machine Learning and Data Science

Cardiovascular Disease Prediction

Student name: Motaz Faraj

Student ID: 1190553

Instructor: Dr. Yazan Abu Farha

Date: January 26th, 2024

Abstract

This project aims to find the best fit model for a Cardiovascular Disease Prediction problem as three models were studied which are the K-Nearest Neighbor (KNN) with k equals to 1 and 3 as this model was used as the baseline to compare between the other two models, the second model to study was support vector machine (SVM) that showed a better performance compared to KNN, the last model to study was random forest which the best results between all three models.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

1.1. Context

Globally, cardiovascular diseases (CVDs) account for 17.9 million deaths annually, making them the leading cause of death. Preventive measures must be put into place as soon as possible in order to maximize healthcare resources and identify individuals who are at risk. By combining a variety of patient-specific data to improve accuracy and clinical applicability, this machine learning project aims to create a strong and reliable predictive model for estimating the risk of cardiovascular diseases.

This project's main goal is to develop a machine learning model that can accurately predict a person's chance of developing cardiovascular diseases. Through the utilization of various datasets that include clinical measurements, lifestyle factors, medical history, and genetic data, the model seeks to offer a thorough and individualized risk assessment.

1.2. Studied Models

In this project a total of 3 models will be studied and compared among each other and these models are the k-nearest neighbor which will be the baseline model for comparison then the decision tree model and support vector machine (SVM) model will be trained and compared with the other models. The support vector machine (SVM) model was chosen due to its effectiveness in high-dimensional spaces which is often present when working with medical scenarios and complex data, SVM is also robust to overfitting which is a serious problem especially when working with medical scenarios as it can lead to misleading predictions. As for choosing random forest because of its capabilities in working with imbalanced data that could be the case in most medical scenarios.

1.3. Evaluation Metrics

Since this problem is a serious health problem and any mistake will potentially lead to one’s death which means that the precision of the model must be accounted for and a high precision is required. Precision will be used alongside other evaluation techniques such as recall which will be considered as it represents the true positive rate and accuracy the FNR will also be considered and a low FNR is required. The receiver operating characteristic (ROC) curve with the area under it will be considered as well to compare between the tested models with the best model having a higher AUC. Grid Search was used to find the best hyper-parameters for the studied models with cross-validation of five folds provided by the grid search cv parameter.

2. Dataset

The dataset used is the Cardiovascular_Disease_Dataset, this dataset contains a total of 14 columns or features with total sample number of 1000. This dataset describes the patient across a wide range of features related to cardiovascular health. For a better understanding of the data both quantitative measures and visualizations of the data were done and some of which is the distribution of the numerical data which figure 1 shows.

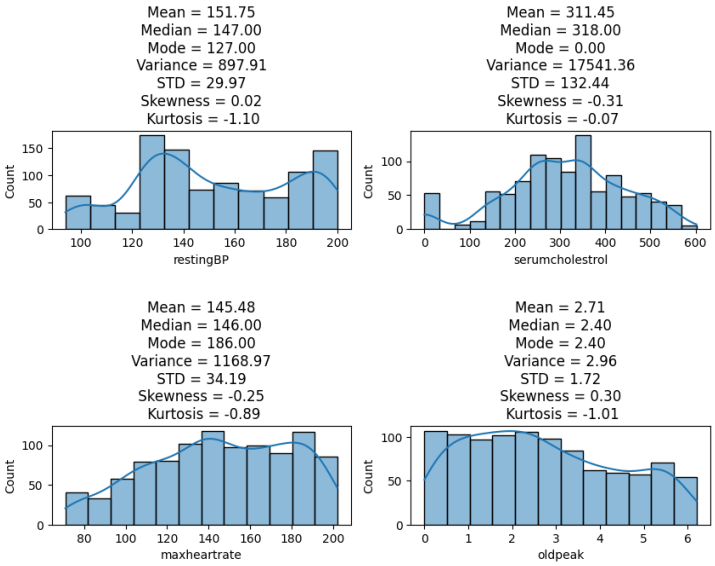

Figure 1: Statistics for some Numeric features

A few statistics about the data are displayed in the above figure, including the mean, which represents the data's average value, and the median, which represents the data's middle value after sorting. It also displays the data's skewness, which indicates whether the data is asymmetrical or not. For a normal distribution (bell curve), it can display zero skewness as well as right (positive) or left (negative) skewness to varying degrees. The following table shows some statistics.

Feature	Mean	Median	Mode	Variance	STD	Skewness	Kurtosis
restingBP	151.75	147	127	897.91	29.97	0.02	-1.1
serumcholestrol	311.45	318	0	17541.36	132.44	-0.31	-0.07
maxheartrate	145.48	146	186	1168.97	34.19	-0.25	-0.89
oldpeak	2.71	2.4	2.4	2.96	1.72	0.3	-1.01

Table 1: Some features with their statistics

The following figure shows the distribution of categorical features to cardiovascular disease.

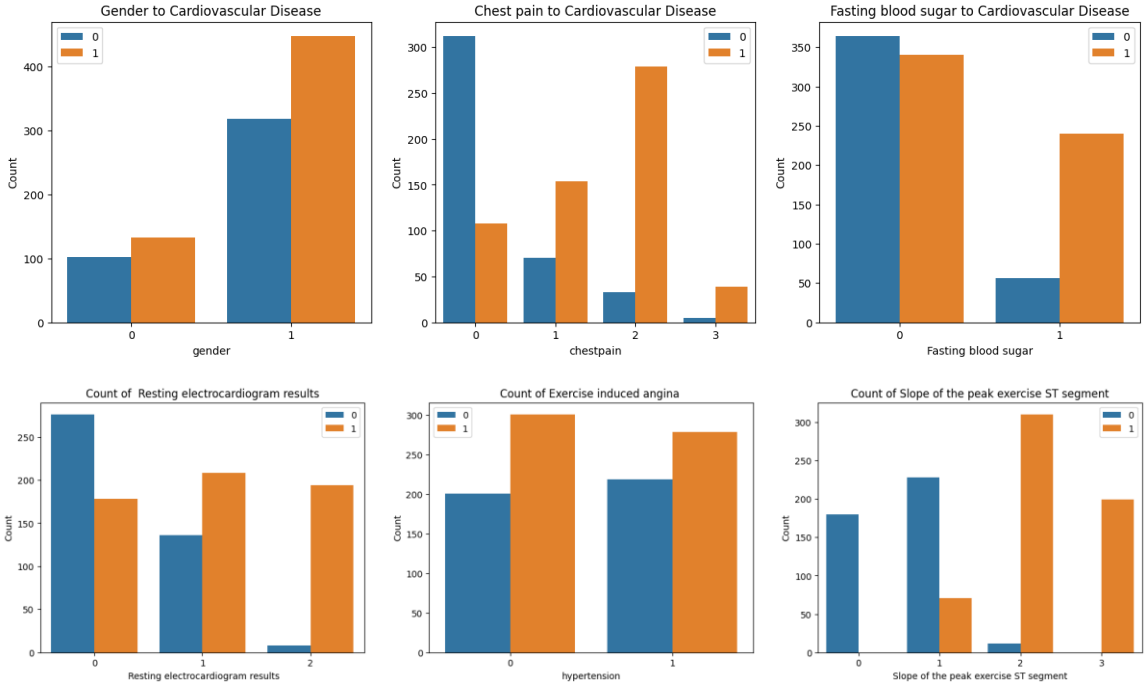

Figure 2: Categorical features to Cardiovascular Disease

The data was preprocessed in order to determine whether there were any missing values, and it was found to be clean and free of any missing values. Additionally, the interquartile range (IQR) test and box plot were used to examine the presence of outliers, and the results indicated that there were none in the data. The last step before start training the models was to scale the features with high values so that all features contribute equally at the training of the models. The following figure shows the percentage of each class using pie plot.

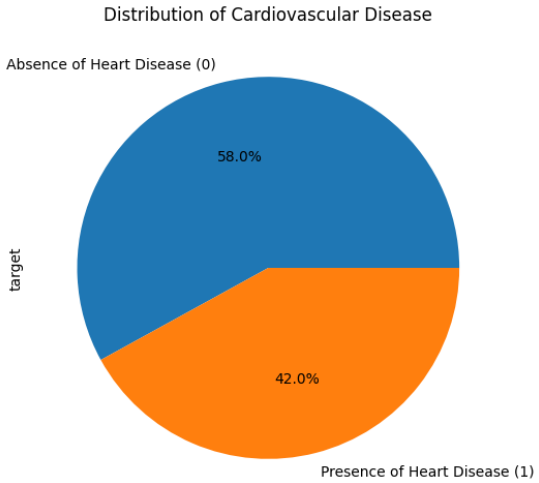

Figure 3: Distribution of Cardiovascular Disease

3. Experiments and Results

3.1. K-Nearest Neighbor

As a baseline model a K-nearest neighbor was trained using two k values which are 1 and 3 and the performance of each model is summarized in the following table.

Metric \ K value	K = 1	K = 3
Accuracy	0.93	0.95
Precision	0.963963963963964	0.9572649572649573
Recall	0.9145299145299145	0.9572649572649573
F1-score	0.9385964912280702	0.9572649572649573
FNR	0.08547008547008547	0.042735042735042736
ROC-AUC	0.933	0.968
Precision-Recall Curve-AUC	0.96	0.98

Table 2: K-nearest neighbor results

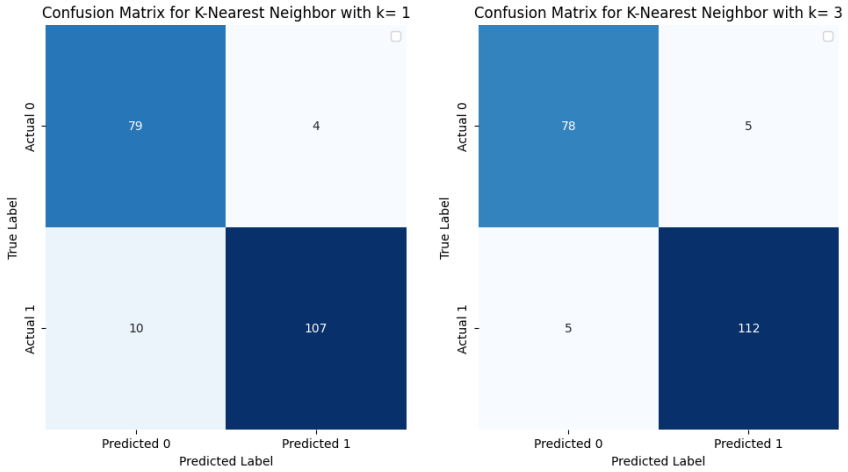

Figure 4: Confusion Matrix for KNN

With k=1, the model performs well, obtaining high recall, F1-score, and precision. Good overall classification performance is given by the ROC-AUC and Precision-Recall Curve AUC. Only 4 false positives and 10 false negatives, according to the confusion matrix, are present. Moreover, a little increase in accuracy and a balanced precision and recall are obtained when k is increased to 3. Better model performance is given by the ROC-AUC and Precision-Recall Curve AUC. Only 5 false positives and 5 false negatives are shown in the confusion matrix. It can be seen that the second model achieved better performance indicated by the reduction of false negatives from 10 to 5.

3.2. Support Vector Machine (SVM) model

The first model to study after KNN was SVM and it gave slight improvement compared to the results obtained in the last part. The following table shows the results for training the model with its hyper-parameters tuned using grid search and cross validation of 5 folds.

Metric	Results
Accuracy	0.96
Precision	0.9658119658119658
Recall	0.9658119658119658
F1-score	0.9658119658119658
FNR	0.0341880341880341
ROC-AUC	0.994
Precision-Recall Curve-AUC	1.0

Table 3: SVM results

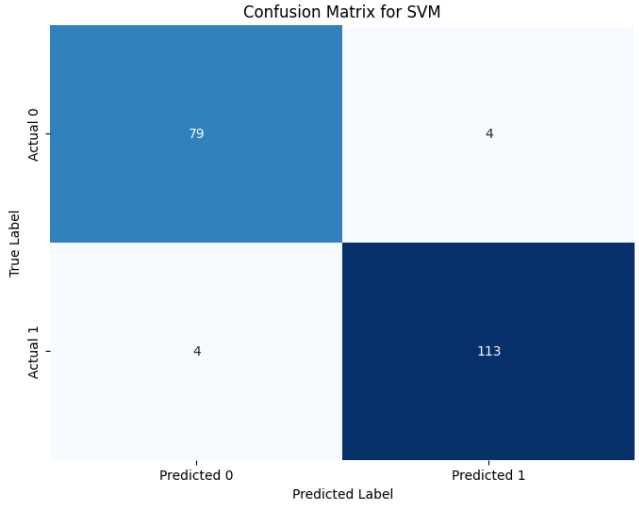

Figure 5: Confusion Matrix for SVM

With an accuracy of 96%, SVM's overall performance shows a strong ability to correctly classify instances. Furthermore, the high precision indicates a high probability of correctness when the model predicts a positive case. This is especially important when false positives can be expensive. In medical applications, where missing a positive case (false negative) can have serious repercussions, the model's recall of 0.965 suggests that it is effective at capturing a sizable portion of the actual positive cases. Also, the confusion matrix shows only 4 false positives and 4 false negatives, indicating a low number of misclassifications. Compared to KNN with k = 1 and a slight improvement over KNN with k = 3.

3.3. Random Forest model

The final modal to study is random forest which gave the results shown by the following table. With its hyper-parameters being tuned using grid search with cross validation of 5 folds.

Metric	Results
Accuracy	0.985
Precision	0.9913793103448276
Recall	0.9829059829059829
F1-score	0.9871244635193134
FNR	0.01709401709401709
ROC-AUC	0.99969
Precision-Recall Curve-AUC	1.0

Table 4: Results for Random Forest

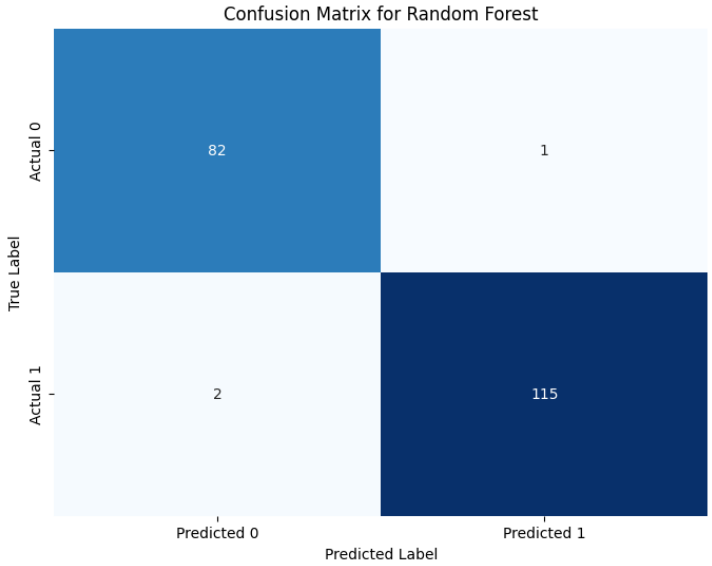

Figure 6: Confusion Matrix for Random Forest

With an impressive overall accuracy of 0.985, the model demonstrates a strong capacity for accurate instance classification. Furthermore, at 0.982 and 0.991, recall and precision are both very good. This implies that practically all real positive cases are effectively captured by the model, which has a high degree of accuracy in identifying positive cases. The AUC values of 0.99969, are remarkably high and suggest exceptional discriminative power across a range of evaluation thresholds. Perfect precision-recall trade-off is indicated by an AUC of 100% for the Precision-Recall Curve, which makes it especially remarkable. The confusion matrix, which only displays 1 false positives and 2 false negatives, makes all of this visible. This suggests an extremely low amount of misclassifications, highlighting the accuracy of the model compared to the other studied models.

4. Performance analysis

The random forest models performed the best out of all of them, according to a comparison of the results from each model. Additionally, to examine the random forest model's performance using evaluation metrics like the confusion matrix, ROC-AUC, accuracy, precision, recall, F1-score, and precision. Table 4 provides an accurate description of the model's overall performance. Its high accuracy indicates a high level of overall correctness, while its high precision and recall indicate a low rate of false positives and false negatives, respectively. Furthermore, an AUC of 1 on the precision-recall curve denotes a perfect precision-recall trade-off. Furthermore, the ROC-AUC of 0.99969 indicates that the model distinguishes between positive and negative instances extremely well.

The model was also able to correctly predict the majority of the data points, with a few errors, as evidenced by the confusion matrix, which showed that 1 cases were predicted to be false positives when in fact they were negative cases of cardiovascular disease. Additionally, 2 false negative cases—which are actually positive cases of cardiovascular disease—were predicted. This behavior indicates that although the overall error rate is low, the model is showing a slight tendency toward false negative, or predicting 0 when the actual label is 1.

The misclassified instances were also studied and one of these misclassified instances was instance 513 of the dataset as it had a male patient with typical angina for chest pain, normal fasting blood sugar and resting ECG, no angina during exercise, an upsloping ST-segment slope, with above average maximum heart rate and slightly elevated Oldpeak (ST segment), and no major vessel blockage and it was labeled as high risk of heart disease (class 1) but was predicted as low risk (class 0) which could be due to some limitations in the model as it may not capture complex interactions between features, overlooking aspects like being young for this specific profile or normal resting blood pressure and cholesterol might have masked underlying issues or due to the models underemphasized the significance of slightly elevated oldpeak and upsloping ST segment in this specific combination of other features. Another case that was studied is instance 439 that was an older patient with almost the same features and classification. The last misclassified instance is 901 which was one of the oldest female patients with no chest pain, normal resting blood pressure, high serum cholesterol, exercise-induced angina and it was classified as positive but predicted as negative and this may be due to some features like low heart rate and downsloping ST segment might have led the model to underestimate the risk.

5. Conclusions and Discussion

After conduction this project it was clear that KNN performed the worst of the three tested models and that's because all the features contribute equally to the distance calculation which makes KNN more sensitive to irrelevant or redundant features like the patient ID in this case which lead to the need to use a feature selection or dimensionality reduction techniques. As for SVM it showed a slight improvement compared to KNN with some errors in this prediction due to its parameter sensitivity as its performance depends greatly on the choice of kernel and other hyper-parameters and selecting the right combination can be time consuming as well as resources but with the right hyper-parameters combination it can perform will on most tasks. But when it comes to random forest which gave the best results out of the studied models it can be prone to overfitting especially with high numbers of trees in the forest which be solved by hyper-parameters tuning like the maximum depth of trees. Some of the limitations that might affect all models is the quality of the used data and if the dataset is biased or incomplete the models may not generalize well to new unseen data.