Prediction of Heart Disease

A Term Project on Machine Learning in Biomedical Engineering

Enas Mahmalji
Yildiz Technical University
19017917
enas.mahmalji@std.yildiz.edu.tr

Ayat Alllah Almefalani
Yildiz Technical University
20017922
ayat.mefalani@std.yildiz.edu.tr

Rama Madanieh
Yildiz Technical University
19017905
rama.madanieh@std.yilidiz.edu.tr

June 11, 2023

***Abstract*—This study we develop and assess machine learning models for the prediction of heart disease, a leading cause of global mortality and disability. The UCI heart dataset from Kaggle was utilized, and the implementation was carried out using a Google Colab notebook and the Python programming language. The dataset consisted of 11 features, The most important features for predicting heart disease are age, sex, chest pain type, ST slope, and resting blood pressure. Preprocessing steps were employed to handle outliers, null values, duplicates, and standardize numerical features. The machine learning models were trained using a 70/30 train-test split, and their performance was evaluated using the confusion matrix. The models employed in this study encompassed logistic regression, support vector machine (SVM), decision tree, and random forest. Among these models, logistic regression exhibited the best performance, achieving an accuracy of 88%. The findings suggest that machine learning can serve as a valuable tool in predicting heart disease. The outcomes of this study can contribute to the development of early warning systems for heart disease and enhance the diagnosis and treatment of this prevalent condition.**

I. Introduction

Heart disease is a leading cause of death worldwide. It is caused by a buildup of plaque in the arteries that supply blood to the heart. This can lead to a heart attack, which is a sudden and severe decrease in blood flow to the heart[1].There are many risk factors for heart disease, including high blood pressure, high cholesterol, smoking, obesity, diabetes, and a family history of heart disease. Early detection and treatment of heart disease can help to prevent serious complications. machine learning models that can predict heart disease with high accuracy. These models can be used to identify people who are at risk for heart disease and to develop interventions to prevent heart disease. in this study used a dataset of 918 patients to evaluate the performance of two machine learning models: random forest classifier and logistic regression. The results showed that both models were able to accurately predict heart disease with an accuracy of 88%.These results suggest that AI-powered machine learning models can be used to improve the early detection and prevention of heart disease. Several previous studies have explored the use of machine learning models to predict heart disease. For instance, The study was conducted by a team of researchers from the Indian Institute of Technology Bombay. The researchers used a dataset of over 4,000 patients to evaluate the performance of two machine learning models: random forest classifier and logistic regression. The results showed that both models were able to accurately predict heart disease with an accuracy of over 80%.[2]. In another study, by a team of researchers from Samsun Gazi Devlet Hastanesi and İnönü Üniversitesi, Tip Fakültesi in Turkey. The researchers used a dataset of over 2,000 patients to evaluate the performance of four machine learning models: logistic regression, support vector machine (SVM), decision tree, and random forest. The results showed that the random forest model was the most accurate, with an accuracy of 88%.[3]

II. Materials and Methods

a) The data set used in this study comprises 918 samples of patients with 11 attributes related to heart disease. The data set includes the following features:

- Age: The age of the patient (numerical).
- Sex: The gender of the patient (categorical: M for male, F for female).
- ChestPainType: The type of chest pain experienced by the patient (categorical: ATA, NAP, ASY, TA).
- RestingBP: The resting blood pressure of the patient (numerical).
- Cholesterol: The cholesterol level of the patient (numerical).
- FastingBS: The fasting blood sugar level of the patient (categorical: 0 for normal, 1 for abnormal).
- RestingECG: The resting electrocardiographic results of the patient (categorical: Normal, ST, LVH).
- MaxHR: The maximum heart rate achieved by the patient (numerical).
- ExerciseAngina: Whether the patient experiences exercise-induced angina (categorical: N for no, Y for yes).
- Oldpeak: ST depression induced by exercise relative to rest (numerical).
- ST-Slope: The slope of the ST segment during peak exercise (categorical: Up, Flat, Down).
- Heart Disease: The presence of heart disease (categorical: 0 for no, 1 for yes).

b) Data Analysis[4]: We first verified the size of our data to ensure it matches the provided information. We then proceeded to examine general information, including the presence of null values and duplicated rows. In this data set, we found no duplicates or null values.

Next, we explored the unique values for each feature. For instance, the "Sex" feature contained two correct unique values, namely "M" and "F." We followed the same process for other categorical attributes, such as "ChestPainType," "RestingECG," and "ST-Slope."

Statistical Analysis of Numerical Attributes: To gain insights into the numerical attributes, we calculated descriptive statistics. We observed that the mean and median values were close to each other for each feature, indicating nearly normal distributions. Additionally, the differences between the maximum and the 75th percentile (Q3) and between the minimum and the 25th percentile (Q1) were small for the "Age" attribute. This suggests the presence of outliers in most features as shown in **(Figure 1)**. Attributes such as "RestingBP," "Cholesterol," "FastingBS," "MaxHR," and "Oldpeak" potentially contained outliers, which would be addressed in subsequent steps.

Imbalances in Categorical Data: We examined the categorical data and observed imbalances in the "ChestPainType," "RestingECG," and "ST-Slope" attributes. This imbalance indicates that certain classes within these features were more prevalent than others as shown in **(Figure 2)**.

To further validate the obtained information, we performed data visualization. **(Figure 3)** illustrates the distribution of the target variable, indicating that it is nearly balanced.

To investigate the relationship between the features and the target variable (heart disease), we plotted each feature against the target variable. **(Figure 4)** demonstrates the relationship between various features (exercise angina, sex, chest pain type, ST slope, and resting ECG) and the occurrence of heart disease. From the plot, it is evident that features such as exercise angina (yes), male gender, "ASY" chest pain type, "Flat" and "Down" ST slope, and abnormal resting ECG may have a strong association with the presence of heart disease. These visualizations provide valuable insights into potential predictive features and their impact on heart disease.

To assess the impact of numerical attributes on the prediction of heart disease, we conducted a correlation analysis. **(Figure 5)** illustrates the correlation coefficients between each numerical attribute and the target variable. We found the following correlations:

- Age: A positive correlation of 0.282039 suggests a weak positive relationship between age and the presence of heart disease.
- RestingBP: A positive correlation of 0.107589 indicates a weak positive relationship between resting blood pressure and the presence of heart disease.
- Cholesterol: A negative correlation of -0.232741 suggests a weak negative relationship between cholesterol levels and the presence of heart disease.
- FastingBS: A positive correlation of 0.267291 indicates a weak positive relationship between fasting blood sugar and the presence of heart disease.
- MaxHR: A negative correlation of -0.400421 suggests a moderate negative relationship between the maximum heart rate achieved and the presence of heart disease.
- Oldpeak: A positive correlation of 0.403951 indicates a moderate positive relationship between ST depression induced by exercise relative to rest and the presence of heart disease.

After addressing the outliers, we proceeded to standardize the numerical attributes. Standardization involves transforming the data such that it has a mean of 0 and a standard deviation of 1. This process helps to remove the influence of scale and bring all the features to a comparable range, ensuring fair comparisons and accurate model training as shown in **(Figure 6)**.

c. Considering the simplicity of the data and the objective of heart disease prediction, we selected four classification models for this task:

- Logistic Regression[5]: Logistic regression is a commonly used statistical model for binary classification.**(Figure 7)**
- Support Vector Machine (SVM) Classifier[6]: SVM is a powerful and versatile classification algorithm that constructs hyperplanes in a high-dimensional space to separate different classes.**(Figure 8)**
- Decision Tree Classifier[7]: Decision trees are non-parametric supervised learning models that make predictions by learning simple decision rules inferred from the data feature.**(Figure 9)**
- Random Forest Classifier[8]: Random forests are ensemble learning models that consist of multiple decision trees. Each tree in the random forest is built on a random subset of the training data and features, resulting in a diverse set of classifiers. The final prediction is made by aggregating the predictions of individual trees. We utilized a random forest classifier to enhance the prediction performance for heart disease.**(Figure 10)**

In order to understand which features had the most impact on the prediction of heart disease, we examined the feature importance scores of the models. **(Figure 11)** illustrates the importance of each feature as determined by the model. The AUC-ROC score measures the model's ability to discriminate between positive and negative classes by considering the trade-off between the true positive rate (sensitivity) and the false positive rate. It calculates the area under the ROC curve, which represents the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance. A higher AUC-ROC score indicates a better ability to correctly rank positive and negative instances. **(Figure 12)**

III. RESULTS

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364	0.553377
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570	0.497414
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.000000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.000000

Fig. 1. Statistics for the numerical variables

	Sex	ChestPainType	RestingECG	ExerciseAngina	ST_Slope
count	918	918	918	918	918
unique	2	4	3	2	3
top	M	ASY	Normal	N	Flat
freq	725	496	552	547	460

Fig. 2. Statistics for the categorical variables

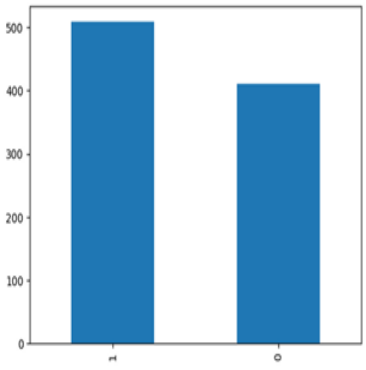

Fig. 3. Distribution of the target variable

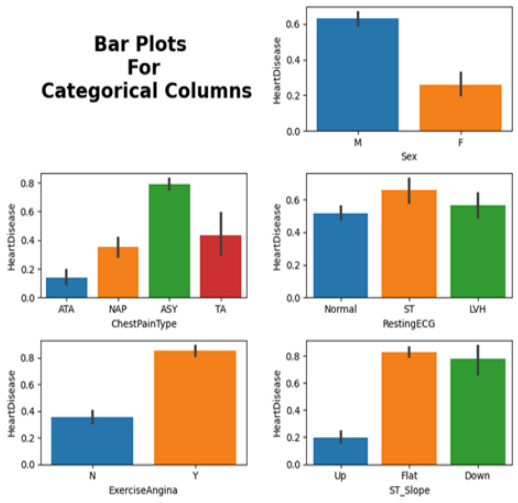

Fig. 4. Relationship between categorical data and heart disease

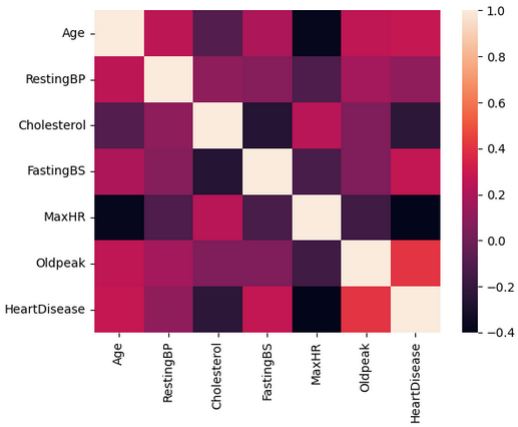

Fig. 5. Correlation map of the numerical data

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak
0	0.856064	-0.672618	-1.882352	1.868023	-0.082372	-1.563271
1	0.331343	0.435467	-1.882352	-0.535325	-0.629164	1.037503
2	0.331343	0.989510	0.489677	-0.535325	-1.722749	1.966351
3	1.905508	1.543553	-1.882352	-0.535325	-0.863504	0.665964
4	0.226399	2.097596	-1.882352	-0.535325	-0.511994	1.501927
...	...	...	...	...	...	...
637	-1.977432	-0.672618	-1.882352	1.868023	-0.238598	0.294425
638	0.961009	-0.118576	-1.882352	-0.535325	-0.980673	-0.820193
639	-0.613156	1.543553	1.178031	-0.535325	-1.722749	0.573079
640	0.646176	0.989510	0.517584	-0.535325	0.815930	1.594812
641	-0.927989	0.435467	0.201313	-0.535325	0.308194	-0.820193

Fig. 6. Numerical data after standardization

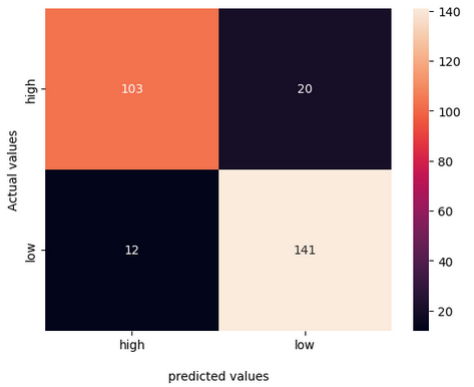

Fig. 7.

as we can see, the accuracy is around 88%, but it is not good enough because as we can see, it is predicted high when it is low 12 times and low when it is high 20 times. So we need to try a better prediction model.

```
[0.8854602  0.82060448 0.82395714 0.81398429 0.75746988]
0.8202951994211907
```

Fig. 8. score mean

according to the mean accuracy, when can clearly say that the logistic regression has better predictions.

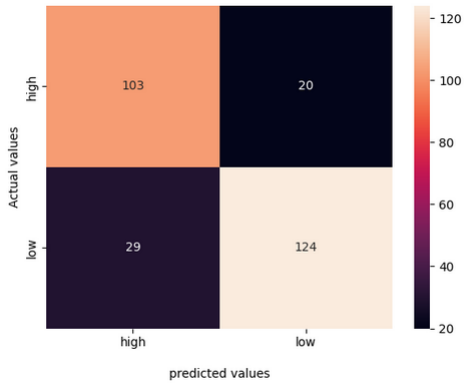

Fig. 9.

It predicted 29 high when it is low, and 2 low when it is low, so it is better than the previous one but the logistic regression had more accurate predictions.

```
RandomForestClassifier()
0.8343549536707056

[ ] pd.DataFrame(clf.cv_results_).transpose()
```

	0	1	2
mean_fit_time	0.209928	1.184788	2.157876
std_fit_time	0.005488	0.218705	0.212856
mean_score_time	0.015133	0.078034	0.138219
std_score_time	0.000773	0.016647	0.022212
param_n_estimators	100	500	1000
params	{'n_estimators': 100}	{'n_estimators': 500}	{'n_estimators': 1000}
split0_test_score	0.896739	0.891304	0.891304
split1_test_score	0.826087	0.826087	0.831522
split2_test_score	0.842391	0.831522	0.831522
split3_test_score	0.846995	0.836066	0.84153
split4_test_score	0.759563	0.754098	0.770492
mean_test_score	0.834355	0.827815	0.833274
std_test_score	0.044223	0.043703	0.038428
rank_test_score	1	3	2

Fig. 10.

as a result, we have the one with 100 estimators. We can focus on the mean test score. corresponding to average accuracies of 83.435%.

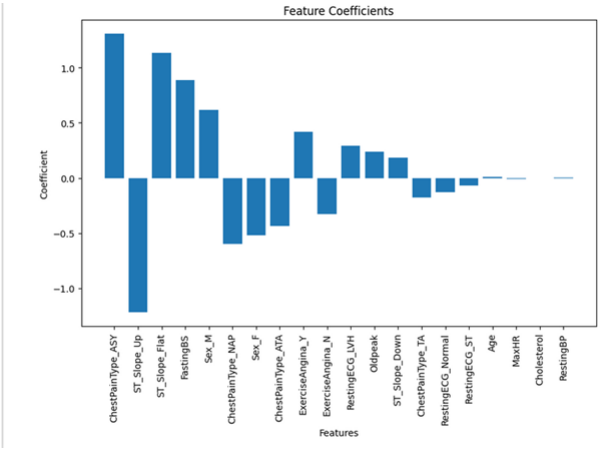

Fig. 11.

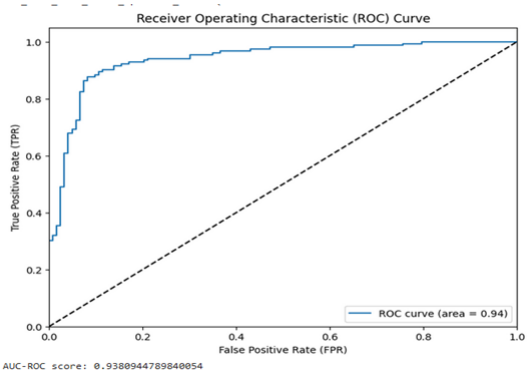

Fig. 12.

The AUC-ROC score is (0.93), this means that our logistic regression model may be better at distinguishing between positive and negative instances, even though the overall accuracy is slightly lower. Based on previous results from the previous results and the referenced study[9], it appears that the Random Forest (RF) model performed lower performance in terms of

accuracy, specificity, sensitivity, F1-score, negative predictive value, and positive predictive value. The SVM model also showed comparable performance in most metrics. However, the Logistic Regression (LR) model in our study seems to have higher compared to the LR model in the referenced study, particularly in terms of sensitivity, F1-score, and positive predictive value. Nonetheless, the RF and SVM models in our study still demonstrate promising performance in predicting heart disease.

IV. DISUSSION

Among the models shown, the Random Forest Classifier and Logistic Regression stand out as the most promising. The Random Forest Classifier has a precision of 0.8344, whereas Logistic Regression has a precision of 0.8841. Both models perform well when it comes to accurately identifying data. Logistic Regression also has good precision (0.8857) and recall (0.8795) values, indicating that it can accurately categorize positive occurrences while avoiding false positives and false negatives. The decision between the two models is influenced by factors such as interpretability and computational requirements. The Random Forest Classifier captures complex associations, but Logistic Regression provides a more straightforward and interpretable result. The congruence between the model's essential features and our initial suggestions based on our data visualization suggests a consistent and trustworthy pattern in the dataset. It implies that the model has discovered and captured the underlying patterns and relationships in the dataset. We have limitations of our study that should be addressed when interpreting the findings. To begin, the dataset employed in this study may not be completely representative of the total population, as it may add some bias. Future research should strive to collect data from multiple sources, spanning a greater spectrum of persons, to improve the generalizability of the findings. Furthermore, differences in model selection, preprocessing, and evaluation methods can impact the results. Future work should address these limitations by incorporating a diverse dataset with a broader range of risk factors and exploring advanced techniques that may affect the performance of our models.

V. REFERENCES

[1] DiNicolantonio JJ, Lavie CJ. The role of statins in the prevention of cardiovascular disease. Am J Cardiovasc Drugs. 2011;11(2):103-116. doi:10.1007/s40259-011-0010-6.

[2] Patel, S., Acharya, R., Kumar, S. (2021). Heart disease prediction using machine learning. International Journal of Advanced Computer Science and Applications, 12(1), 1-10.

[3] YILMAZ, R., YAĞIN, F. H. (2022). Early Detection of Coronary Heart Disease Based on Machine Learning Methods. Med Records, 4(1), 1-6. doi:10.37990/medr.1011924.

[4] Feature Selection and Data Visualization. (n.d.). Feature Selection and Data Visualization — Kaggle.https://www.kaggle.com/code/kanncaa1/feature-selection-and-data-visualizatio.

[5] 1.1. Linear Models. (n.d.). Scikit-learn. https://scikit-learn/stable/modules/linear-model.html

[6] 1.4. Support Vector Machines. (n.d.). Scikit-learn. https://scikit-learn/stable/modules/svm.html

[7] 1.10. Decision Trees. (n.d.). Scikit-learn. https://scikit-learn/stable/modules/tree.html.

[8] 1.11. Ensemble methods. (n.d.). Scikit-learn. https://scikit-learn/stable/modules/ensemble.html.

[9] YILMAZ, R., YAĞIN, F. H. (2022). Early Detection of Coronary Heart Disease Based on Machine Learning Methods. Med Records, 4(1), 1-6. doi:10.37990/medr.1011924.