Heart Risk Disease

Namie Nakamura
Wentworth Institute of Technology

***Abstract*—This project aims to predict the risk of heart disease by using machine learning models. By observing a dataset of 70,000 patient records, the study demonstrates how factors such as age, smoking, and diabetes contribute to heart risk. Different machine learning models' accuracy is compared and it identifies the most important risk factors infuencling predictions.**

***Keywords*—Heart Disease, Machine Learning, Risk Prediction, Health Data, Classification**

I. Introduction

For my Individual Project, I decided to work on Heart Disease Risk Prediction. This study aims to analyze patient symptoms and risk factors to predict heart disease risk using machine learning. The dataset contains 70,000 samples, making it suitable for training machine learning models for classification tasks. The four key questions I aim to answer with the data are: (1) How do age, smoking history, and diabetes influence heart disease risk? (2) Which machine learning model provides the most accurate prediction of heart disease risk? (3) Which symptoms and risk factors are the main cause of heart disease? (4) Which gender has a higher risk of developing heart disease?

II. Datasets

A. Source of dataset

The dataset I used was obtained from Kaggle, titled “Heart Disease Risk Prediction Dataset.” It contains synthetic patient data, including capturing symptoms (e.g., chest pain, shortness of breath), lifestyle factors (e.g., smoking, obesity, sedentary habits), and medical history (e.g., high blood pressure, high cholesterol). Although it’s not based on real hospital records, it is structured to facilitate machine learning research and health risk prediction. The dataset was created and published by Mahatirat Usher and has been publicly available on Kaggle since 2023.

B. Character of the datasets

The dataset is in CSV format, and it consists of 70,000 rows and 19 columns, where each row represents an individual patient’s data. The dataset includes a mix of binary (0/1), categorical, and numerical attributes related to medical history, symptoms, and lifestyle.

Feature	Description	Type
Age	Patient age	Numeric
Gender	Biological sex (0 = Female, 1 = Male)	Binary
Chest_Pain	Presence of chest pain	Binary
Shortness_of_Breath	Reports of breathing difficulty	Binary
Fatigue	Experience of fatigue	Binary
Palpitations	Irregular hearbeat episodes	Binary
Dizziness	Dizziness or fainting	Binary
Swelling	Limb or facial swelling	Binary
Pain_Arms_Jaw_Back	Discomfort in arms, jaw, or back	Binary
Cold_Sweats_Nausea	Cold sweats or nausea	Binary
High_BP	Diagnosed high blood pressure	Binary
High_Cholesterol	Diagnosed high cholesterol	Binary
Diabetes	Diagnosed diabetes	Binary
Smoking	Smoking status	Binary
Obesity	Obesity indicator	Binary
Sedentary_Lifestyle	Lack of physical activity	Binary
Family_History	Family history of heart disease	Binary
Chronic_Stress	Experience of chronic stress	Binary
Heart_Risk	Target label (0 = Low Risk, 1 = High Risk)	Binary

For data cleaning, I removed duplicates and created an “Age_Group” column for categorical analysis. There were no missing values in the dataset.

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

III. Methodology

For this individual project, I used three different machine learning models: Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. These three models were used to compare their performance to determine the best fit.

A. Logistic Regression

Logistic Regression is a linear classification algorithm to predict binary outcomes. This assumes a straight-line relationship between the inputs and the result, so it might not capture complex patterns. I used "LogisticRegression()" from scikit-learn with a "max_iter=1000" to ensure combination. A disadvantage is that it may not perform well on complex or non-linear datasets and is sensitive to feature scaling.

B. Decision Tree Classifier

A Decision Tree works by splitting data based on yes/no conditions. It is an easy model to understand and doesn't need any data scaling. I decided to use this model because it works well with the binary symptom data. A disadvantage is that it is prone to overfitting, especially with deep trees or noisy data. The Python module/function is "sklearn.tree.DecisionTreeClassifier()"

C. Random Forest Classifier

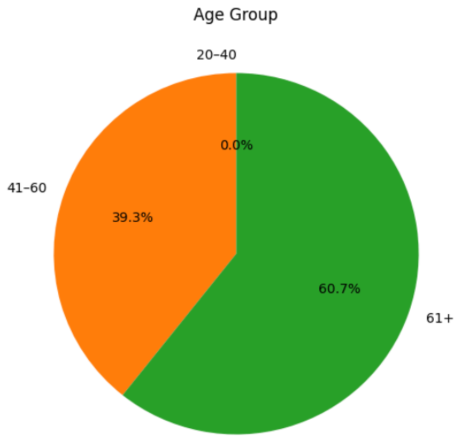
Random Forest is a model combining many decision trees for better performance. This helps reduce overfitting and usually improves accuracy. I decided to use this model because it performed well across multiple classification problems and can identify which features are most important. The Python module/function is "sklearn.ensemble.RandomForestClassifier()"

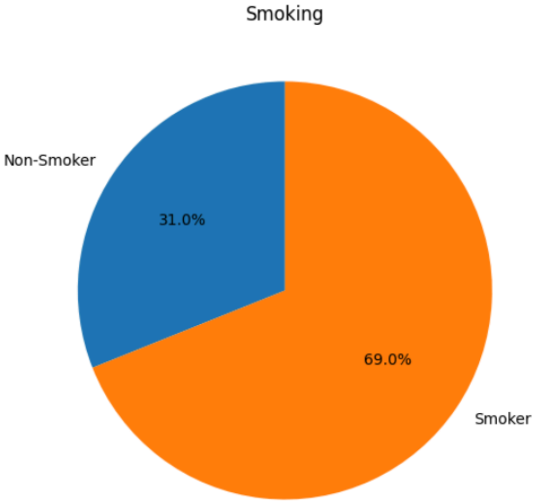
IV. Results

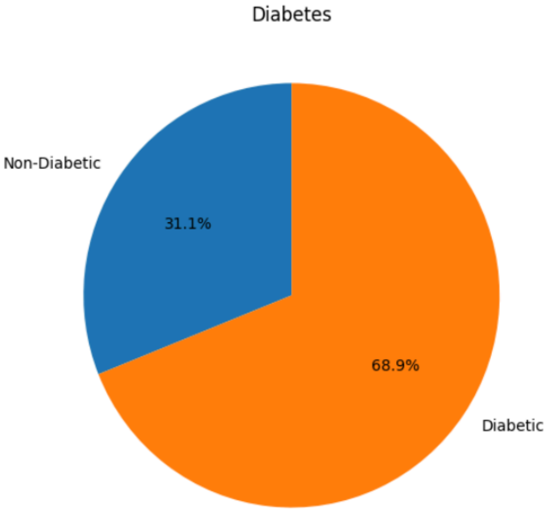
For this section, I will provide the results from the data analysis and model evaluations. I used pie charts, bar charts, and numerical summaries to demonstrate how different risk factors affect heart disease. I also compared the performance of three machine learning models.

A. Age, Smoking, and Diabetes Risk

I grouped the data by Age Group, Smoking, and Diabetes to get an understanding of how certain factors affect heart disease risk by using pie charts. The factor with the highest percentage of heart disease risk is "Smoking" with 69%, followed by "Diabetes" with 68.9%, and lastly "Age Group" for ages "20-40" with 0%, ages "41-60" with 39.3%, and ages "61+" with 60.7%.

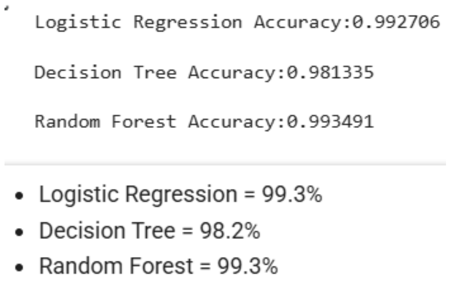






B. Model Accuracy Comparison

I trained and tested three machine learning models: Logistic Regression with an accuracy of 99.3%, Decision Tree with an accuracy of 98.2%, and Random Forest with an accuracy of 99.3%. This demonstrates that Logistic Regression and Random Forest are both performed the best.

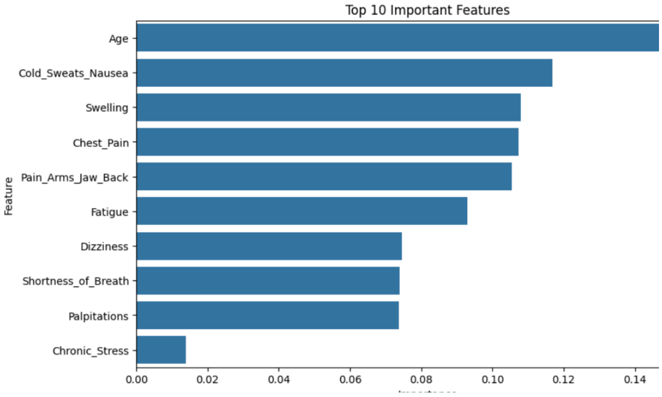


```
Logistic Regression Accuracy:0.992706

Decision Tree Accuracy:0.981335

Random Forest Accuracy:0.993491
```

- Logistic Regression = 99.3%
- Decision Tree = 98.2%
- Random Forest = 99.3%

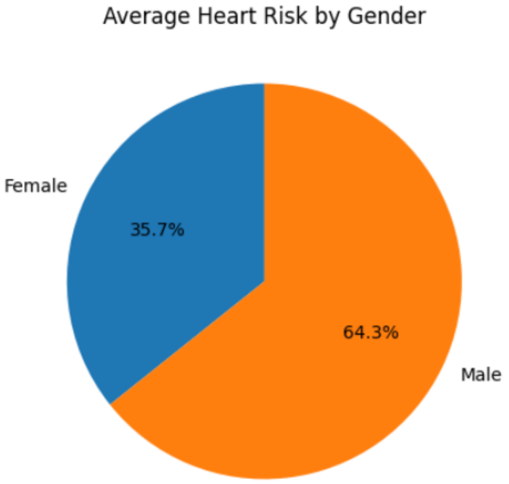
C. Most Important Features

Using the Random Forest model with "feature_importances_", I was able to identify the most influential features contributing to heart disease predictions. To make it shorter, the top 5 were (1) Age, (2) Cold_Sweats_Nausea, (3) Swelling, (4) Chest_Pain, and (5) Pain_Arms_Jaw_Back.



D. Gender-Based Risk

I analyzed heart disease risk based on gender with a pie chart visualization. For males, it had a percentage of 64.3%, making it higher than females with a percentage of 35.7%. This demonstrates that males are often prone to earlier heart-related issues.



V. Discussion

Although the project had very successful results, there were a few limitations. Since this dataset is synthetic, it would have been better to have these data collected from real patients or clinical studies. This means that while the patterns and model performance look promising, they may not represent real-world outcomes. Another limitation is that I only used three models without applying any advanced tuning or optimization, grid search, or hyperparameter tuning. I believe that the models could have performed better with proper tuning. For work suggestions, using a real-world medical dataset from a hospital or research institution could help a lot.

VI. Conclusion

For this project, I wanted to demonstrate how machine learning can be used to predict heart disease risk using a synthetic patient dataset. I was able to get two models with the same accuracy, 99.3%, meaning they both performed the best. While this dataset was not real-world clinical data, the results show the potential for using machine learning to help with early detection and health risk assessments.

Acknowledgment

I would like to thank Doctor Weijie Pang for being an incredible Data Science instructor and for her guidance throughout the completion of this project. Also, a special thanks to the creators of the Kaggle dataset, Heart Disease Risk Prediction Dataset, as well as the developers of Python libraries that helped me complete this project.

References

[1] M. Usher, "Heart disease risk prediction dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/mahatiratusher/heart-disease-risk-prediction-dataset/data