Heart Disease Detection Using Machine Learning Models

Rafael González Cartagena - 90352 , Yadriel A. Ocasio Adorno – 135914, Mentor: Edna Chaar

Polytechnic University of Puerto Rico- Undergraduate Research Program 2022-2023

HSI STEM TITLE III
POLYTECHNIC UNIVERSITY OF PUERTO RICO
A MULTIFACETED APPROACH TO STUDENT CENTERED STEM EDUCATION
P031S210139

ABSTRACT

Heart disease, that is, the set of various health complications that negatively affect the heart, is currently one of the main causes of worldwide deaths in human beings. For instance, in the United States (US), it has the highest mortality rate for both men and women alike amounting to 545,000 deaths in 2021 alone. For this very reason and because of the current advancements in computing technology, this research project studies the accuracy of machine learning algorithms; these being: K – Nearest Neighbor, Gradient Boost and Light GBM, in the detection of heart disease using already compiled data, namely, datasets. Of the three (3) models, it was found that the Light GBM model presented the best results with a 98.5% of accuracy score between the two (2) datasets, followed by the Gradient Boost (95%) and the K – Nearest Neighbor (90.5%). With that being said, the datasets used for this project are Rashik Rahman's Heart Attack Analysis and Prediction Dataset and David Lapp's Heart Disease Dataset - Public Health Dataset; both acquired from the Kaggle website. Moreover, regarding the methods and technologies used, these include the Python 3 programming language with its SK-Learn library, Google's Collaboratory service and various topics associated with Machine Learning, such as: Feature Scaling, Data Imputation and Data Endcoding, to name a few.

OBJECTIVES

The objectives of this research project include:

1. The usage of the following machine learning algorithms:
   - K-Nearest Neighbor,
   - Gradient Boost, and
   - Light GBM

   for the evaluation of their respective accuracies in the detection of heart disease using already compiled datasets, namely: Rashik Rahman's Heart Attack Analysis and Prediction Dataset, and David Lapp's Heart Disease Dataset - Public Health Dataset.
2. Evaluate the accuracy results differences between data data that was transformed and feature-scaled and data which was only feature-scaled.
3. Identify the possibility of testing already trained machine learning models with new data from other datasets.

Materials

The materials and tools used during this investigation are as follows:

- **Programming Languages:** Python 3 with its NumPy, Matplotlib, Pandas, SciPy, and SK-Learn libraries,
- **Compilers:** Google's Collaboratory service and Microsoft's Visual Studio Code,
- **Datasets:** Rashik Rahman's Heart Attack Analysis and Prediction Dataset and David Lapp's Heart Disease Dataset - Public Health Dataset found in *Kaggle.com*.

Methodology

The overall methodology of the research project is as follows:

Stage 1 – Examination of the Datasets

Both datasets include the recommended attributes that should be used whenever using machine learning models for the detection of heart disease, as denoted in the investigative report of S. Challammal and R. Sharmila, namely Recommendation of Attributes for Heart Disease Prediction using Correlation Measure *(S.Chellammal & Sharmila, 2019)*. These attributes are shown in **Figure 1: S.Chellammal & Sharmila, 2019**.

Attributes	Description	Type	Value
Age	Age	Integer	[29 77]
Sex	Sex	Integer	1 = male; 0 = female
Pc	Chest pain type	Integer	1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic
Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	Integer	[94,200]
Chol	Serum cholestoral in (mg/dl)	Integer	[126,564]
Fbs	Fasting blood sugar (>120 mg/dl)	Integer	1 = true; 0 = false
Resteeg	Resting electrocardiographic results (values 0, 1, 2)	Integer	[0, 2]
Thalach	Maximum heart rate achieved	Integer	[71,202]
Exang	Exercise induced angina	Integer	1 = 4 yes; 0 = no
Oldpeak	ST depression induced by exercise relative to rest	Real	[0.00, 62.00]
Slope	The slope of the peak exercise ST segment	Integer	1 = upsloping; 2 = flat; 3 = downsloping
≠ of MajorVessels	Number of major vessels (0-3) colored by flourosopy	Integer	[0, 3]
Thal	Thal	Integer	3 = normal; 6 = fixed defect;

Figure 1: S.Chellammal & Sharmila, 2019

Furthermore, regarding their respective sizes and female-to-male ratios, the Heart Attack Analysis and Prediction Dataset included 303 entries and a gender ratio of 1:0.464, and the Heart Disease Dataset - Public Health Dataset attained to 1025 and 1:0.438, respectively. Subsequently, an analysis of the relationship between the attributes of both datasets was also developed, one of which is shown below:

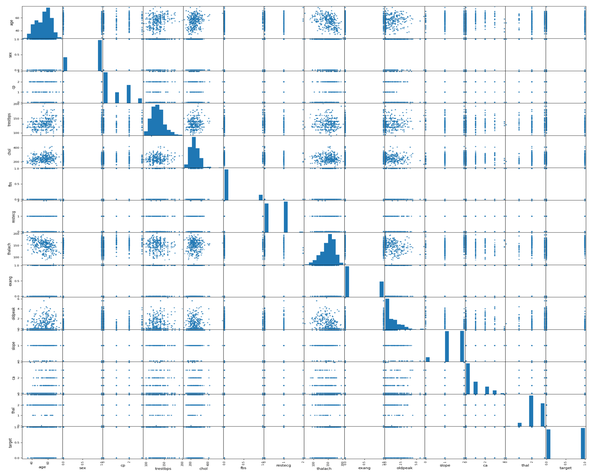

Figure 2: Correlation Matrix of Heart Disease Dataset - Public Health Dataset

Stage 2 – Preparation of the Datasets

It should be noted that the data found in both datasets included no missing values nor categorical data; henceforth, the data was not imputed or encoded. Instead, both datasets were divided into *training* sets and *testing* sets. The data, like that which was related to the cholesterol (chol) of patients, showed *left-heavy-tail distributions*, that of which was later converted into its most closely related *gaussian distributions* with the usage of logarithmic transformations. Lastly, the features of both the already mentioned training sets were scaled with the *Standard Deviation*.

Stage 3 – Model Evaluation

The performances of each of the models (K – Nearest Neighbor, Gradient Boost and Light GBM) were evaluated with arbitrary hyperparameter inputs against David Lapp's Heart Disease Dataset - Public Health Dataset using *cross validation*. Some of the best hyperparameters that were found are as follows:

Best Hyperparameters with Original Data	
K – Nearest Neighbor	N_NEIGHBORS: 4
Gradient Boost	LEARNING_RATE: 0.31, MAX_DEPTH: 27, N_ESTIMATORS: 566
Light GBM	LEARNING_RATE: 0.11, MAX_DEPTH: 6, N_ESTIMATORS: 918

Best Hyperparameters with Transformed Data	
K – Nearest Neighbor	N_NEIGHBORS: 4
Gradient Boost	LEARNING_RATE: 0.76, MAX_DEPTH: 16, N_ESTIMATORS: 555
Light GBM	LEARNING_RATE: 0.86, MAX_DEPTH: 27, N_ESTIMATORS: 655

RESULTS

The accuracy scores that were obtained of each of the machine learning models after first testing and training them with David Lapp's dataset, and then testing them with Rashak Rahman's one are:

Algorithm Name	David Lapp's DT AC Score	Rashik Rahman's DT AC Score
K – Nearest Neighbor	88%	93%
Gradient Boost	95%	95%
Light GBM	99%	98%

CONCLUSION

Throughout this research investigation, the machine learning models known as K – Nearest Neighbor (KNN), Gradient Boost and Light GBM were studied and used with Python 3 SK-Learn Library to determine the possibility of using the algorithms to predict the existence of heart disease in a patient; moreover, if possible, identify which of the three (3) would work best to accomplish such a task. For a start, it was found that it was indeed possible to predict the existence of heart disease in a human patient with the usage of machine learning models; moreover, that with adequate tuning and given the structure of the data itself, such predictions may be certain. For instance, after evaluating the results of each of the models, such as, their respective recalls, precisions, accuracies and F1 scores, it was possible to identify the Light GBM machine learning model, with an overall accuracy of 98.5%, as the best one of the three. It should be noted that this investigation did not find any differences between the results that were obtained from training and testing the models with data which was only feature-scaled via standardization and those which were acquired with data which was both feature-scaled in the same manner and logarithmically transformed; that is, the results were the same. Moreover, as can be already deduced, machine learning models can be tested with multiple sets of new data; of course, such data must include the attributes that such models were trained with.

REFERENCES

American Heart Association Editorial Staff. (2021, November 8). *Angina (Chest Pain)*. Retrieved from American Heart Association: https://www.heart.org/en/health-topics/heart-attack/angina-chest-pain

Fiori, L. (2020, May 22). *Distance metrics and K-nearest neighbor (KNN)*. Medium. https://medium.com/@luigi.fiori.lf0303/distance-metrics-and-k-nearest-neighbor-knn-1b840969c0f4

GeeksForGeeks. (2019, June 12). *ML | Rainfall Prediction Using Linear Regression*. Retrieved from GeeksForGeeks: https://www.geeksforgeeks.org/ml-rainfall-prediction-using-linear-regression/

GeeksForGeeks. (2022, August 24). *ML | What is Machine Learning?* Retrieved from GeeksForGeeks: https://www.geeksforgeeks.org/ml-machine-learning/

Geron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow.*

Lapp, D. (2019). *Heart Disease Dataset*. Retrieved from Kaggle: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset