Using Personal Key Indicators and Machine Learning-based Classifiers for the Prediction of Heart Disease

Raghad Jahed
Computer Engineering Department
Princess Sumaya University for technology
Amman, Jordan
rag20180194@std.psut.edu.jo

Omar Aseer
Computer Engineering Department
Princess Sumaya University for technology
Amman, Jordan
oma20180582@std.psut.edu.jo

Amjed Al-Mousa
Computer Engineering Department
Princess Sumaya University for technology
Amman, Jordan
a.almousa@psut.edu.jo

2023 International Conference on Smart Computing and Application (ICSCA) | 979-8-3503-4705-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSCA57840.2023.10087430

***Abstract*— This paper explores different machine learning algorithms and data splits to measure each model's accuracy, precision, and recall. The models use personal key indicators to predict heart disease. Heart disease, according to the center for disease control, is the leading cause of death in the United States, and about 32% of all global deaths are due to heart diseases such as heart attacks and strokes. Therefore, it is of the utmost importance to integrate machine learning into heart disease prediction and hopefully alert people of the dangers that lie ahead using personal key indicators. Multiple models are recommended in this paper and produced different but relatively high accuracies, the highest accuracy being 99% using Random Forest Classification and a data split based on race.**

Keywords—Machine learning, Heart disease, Coronary heart disease, myocardial infarction, Stochastic gradient descent, Decision tree classifier, Random forest classifier

I. Introduction

According to the Centers for Disease Control and Prevention (CDC), heart disease is the leading cause of death in the United States. About one in every four deaths (about 660 000 deaths annually) is caused by heart disease [1]. Coronary Heart Disease (CHD) is a disease where the coronary arteries are narrowed or blocked. CHD can often result in a heart attack, which happens when one or more areas of the heart do not get enough oxygen due to the flow of oxygenated blood to the heart muscles being cut off [2].

Many factors can increase the risk of heart diseases, such as diabetes, obesity, and the lack of physical activity [1]. The high mortality rate caused by heart disease, along with the fact that many of the factors that can cause heart disease are well-known and well-documented, makes creating a system that can act as a preliminary test to inform individuals if they are likely to have heart disease very helpful for early detection.

A machine learning powered system that can act as a preliminary test to inform individuals if they are susceptible to heart disease can be very helpful for early detection. The primary advantage of this system is the fact that many of the factors that are used are generally known by the patient. These factors can be provided by the patient without the need for extensive medical testing.

This work will focus on creating a machine-learning model that can detect heart disease based on several key indicators. The dataset used is based on an annual survey conducted by the CDC in 2020 [3]. The dataset contains over 300 000 instances and 17 numerical and categorical features (excluding the label). The dataset is a particularly good fit for this system due to the diversity it offers in terms of features.

II. Related Work

Machine Learning has been successfully deployed in many fields, like safety and education [4, 5]. However, the use of machine learning in the medical field has rapidly increased [6]. And one of the important and relevant cases is machine learning used in heart disease detection. Different preprocessing of data, different algorithms used, or even completely different datasets can significantly transform the machine learning outcome. The work presented in [7] used a relatively small dataset (303 instances), with algorithms such as SGD Classifier, KNN Classifier, Random Forest Classifier, and more, and resulted in the highest accuracy of 90% using hard-voting ensemble.

Using the same dataset, [8] proposed an "efficient and accurate" system to diagnose heart disease, using different algorithms along with standard feature selection algorithms such as Relief, Minimal redundancy, and maximal relevance. The algorithms on the full features data resulted in the highest accuracy of 85% with the Support Vector Machine Algorithm (SVM). And with different feature selection algorithms such as Relief, LASSO, and LLBFS, they obtained an accuracy of 86%, 86%, and 87% for the same algorithm.

The work in [9] used the same Cleveland UCI repository, with R studio rattle, to perform the heart disease classification in a method called HRFLM. This method produced the highest accuracy of 88.4% and the lowest classification error among all algorithms at 11.6%. Then it compared the results with other algorithms, such as Decision Trees, which obtained an accuracy of 85%, Deep learning, which resulted in an accuracy of 87.4%; and the second highest accuracy for the "Vote" algorithm of 87.41%.

To test different machine learning environments and their effects on different SVM techniques and ML algorithms such as Linear, Quadratic and Cubic SVM along with Decision trees and Ensemble Subspace Discriminant, the work in [10] compared the accuracy results of MATLAB and WEKA for heart disease detection. For Decision Trees, MATLAB resulted in an accuracy of 60.9% and an accuracy of 67.7% for WEKA. However, for SVM algorithms, the difference in accuracies between MATLAB and WEKA reached 9.3% for Cubic SVM, proving that different environments could result in a huge accuracy difference for some algorithms.

III. Experimental Setup

This paper will use different classification algorithms to predict heart disease using personal key indicators. Those key indicators are likely to be already known by the patient, such as whether or not the patient has ever had a stroke, and their general physical and mental health, without the need for medical examination.

A. Dataset attribute information

Each of the 320,000 instances in the dataset has 18 attributes, including the target variable. A description of each attribute is presented in Table 1.

Table 1 Personal Key indicators dataset attribute description

Attribute	***Range/Categories***	**Description**
Heart Disease	Yes/No	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI).
BMI	12.02-94.85	Body Mass Index (BMI).
Smoking	Yes/No	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes].
Alcohol Drinking	Yes/No	Heavy drinkers (Defined as adult men who have more than 14 drinks per week and adult women having more than seven drinks per week).
Stroke	Yes/No	Have you ever had a stroke?
Physical Health	0-30	Thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 was your physical health not good? (0-30 days)
Mental Health	0-30	Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days)
Diff Walking	Yes/No	Do you have serious difficulty walking or climbing stairs?
Sex	Male/Female	Are you male or female?
Age Category	18-80+	Fourteen-level age category.
Race	White / Black /Asian / Hispanic (American Indian,Alaskan Native) Other	Imputed race/ethnicity value.
Diabetic	No / No, borderline diabetes Yes (during pregnancy) / Yes	Have you ever had diabetes?
Physical Activity	Yes/No	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job.
Gen Health	Poor/Fair / Good / Very Good / Excellent	Describe your general health.
Sleep time	1-24	On average, how many hours of sleep do you get in 24 hours?
Asthma	Yes/No	Have you ever had asthma?
Kidney Disease	Yes/No	Were you ever told you had kidney disease, not including kidney stones, bladder infection, or incontinence?
Skin Cancer	Yes/No	Have you ever had skin cancer?

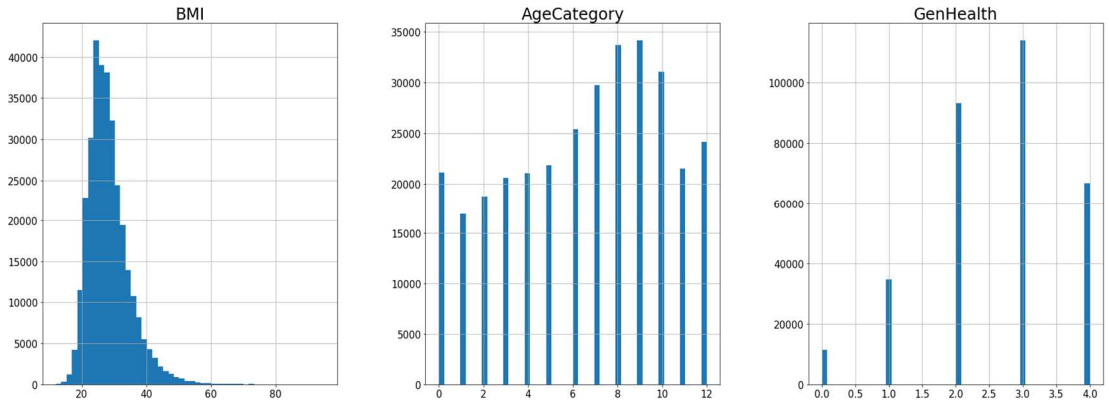

Figure 1 Histograms for categorical data

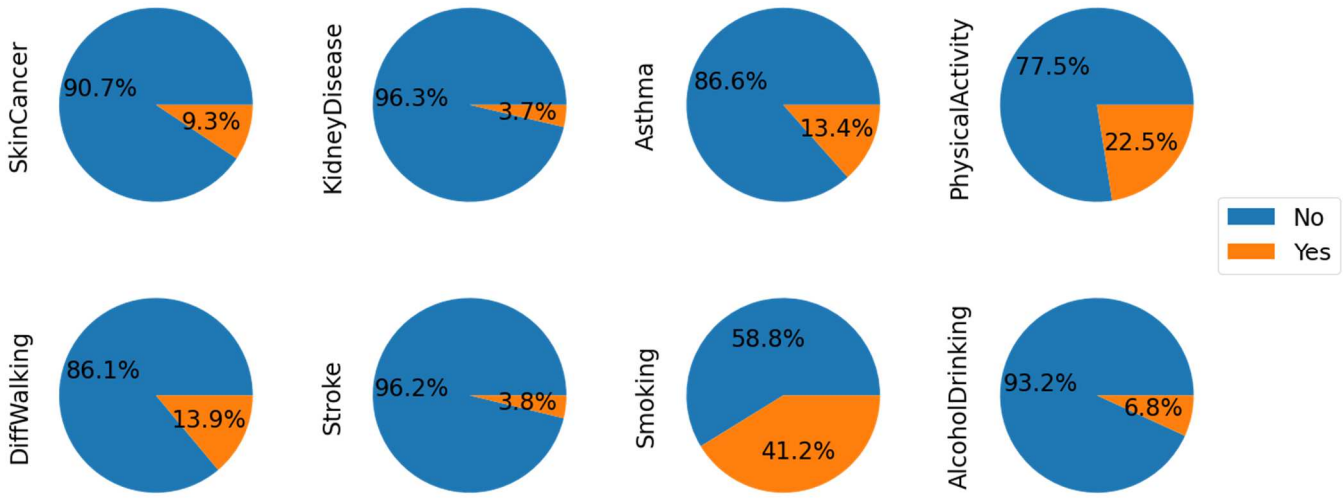

Figure 2 Binary feature distributions

To see the data distribution for BMI, Age category, and General health, as they are non-binary categorical features, we used histograms, as shown in Figure 1. As for the binary features, such as smoking, physical activity, and asthma, their distributions are shown in Figure 2

The Pearson correlation method calculated the correlation factor to gather more information on the relationship between the attributes and the target attribute (Heart Disease). The correlation values are shown in Table 2.

B. Data preprocessing

To further understand the data distribution, Figure 3 and Figure 4 show pie chart graphs of Gender and Race distribution. It should be noted that these attributes were chosen to study as they are used to split the data in section V.

Since most data attributes were categorical, we resorted to encoding the data to use it for training the machine learning models. To do so, One Hot Encoding was used for the "Race" attribute, and Ordinal Encoding was used for the rest of the categorical attributes.

Before starting the training process, Standard Scalar from the scikit-learn library was used to scale the data. The data was then split into 80% training data (467,875 training instances) and 20% test data (116,969 testing instances).

Table 2 Correlation between each attribute and heart disease

Attribute	**Correlation with label**
BMI	0.051803
Smoking	0.107764
Alcohol Drinking	-0.032080
Stroke	0.196835
Physical Health	0.170721
Mental Health	0.028591
Diff Walking	0.201258
Sex	0.070040
Age Category	0.233432
Race	0.034854
Diabetic	0.180826
Physical Activity	-0.100030
Gen Health	-0.243182
SleepTime	0.008327
Asthma	0.041444
Kidney Disease	0.145197
Skin Cancer	0.093317

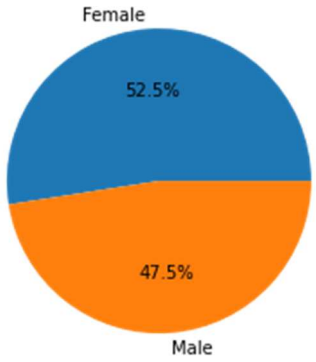

Figure 3 Original data distribution based on gender

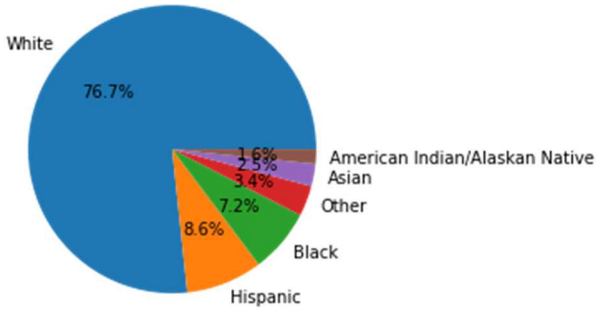

Figure 4 Data distribution based on race

The dataset was unbalanced in terms of Heart disease count. There are two types of data sampling techniques for imbalanced data. These techniques are oversampling; randomly duplicating data from the smaller subset. Under-sampling involves randomly removing data from the larger subset. Oversampling was used to overcome this issue resulting in an equal number of instances in each category of heart disease (292422 instances).

The difference between the original and the over-sampled data is shown below in Figure 5 and Figure 6.

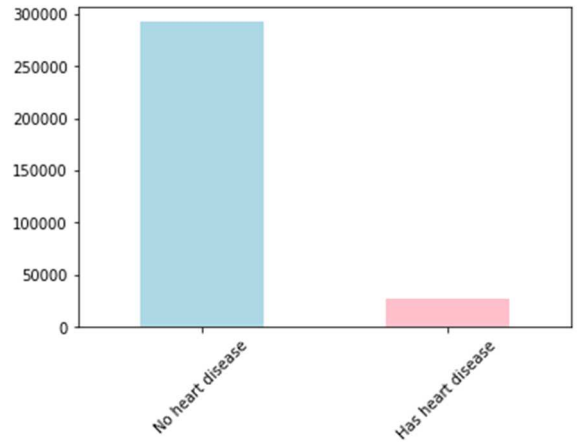

Figure 5: Instance counts before oversampling

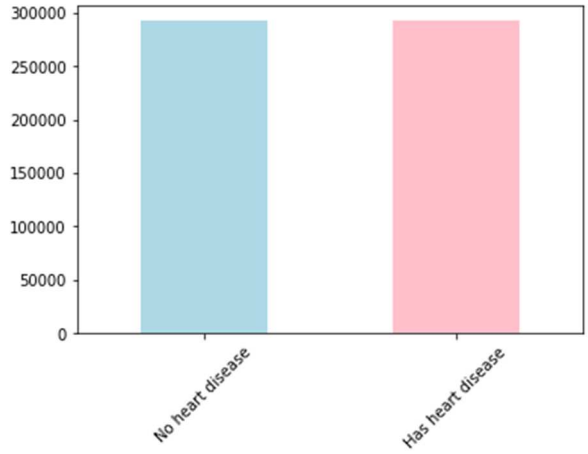

Figure 6 Instance counts after oversampling

IV. Machine Learning Algorithms

The following machine learning models were trained on the entire training dataset, and then the data was split based on the attributes “Sex” and " Race " to compare the different models.

A. Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) was the first algorithm to train the model. Stochastic Gradient Descent uses a certain cost function and the function's gradient to arrive at the minimum value. In each iteration, the SGD uses one instance from the data.

B. Decision Tree Classifier

As the name suggests, Decision Trees Classifier has a tree-like model that makes if-else-like decisions at each node. To train the Decision Trees, a metric called “gini” is used to measure the purity of each class.

C. Random Forest Classifier

Ensemble methods use multiple learning models to achieve better performance. Random Forest Classifier is a type of ensemble method which uses multiple Decision Trees. The output of the Random Forest Classifier is the class selected by the majority of the Decision Trees.

V. Results and Analysis

In addition to analyzing the unsplit data, different data splits were used to experiment with some attributes' effects on the model's performance. The data was first split based on the “Race” attribute. This attribute was chosen since most instances were white, as shown in Figure 4. The second data split was based on the “Sex” attribute. This data split was chosen to see the effect of gender on the model's accuracy. The data split based on different attributes was then split into training and testing data using a random train-test split.

A. Unsplit data

Starting with the stochastic gradient descent classifier, an accuracy of 76% was achieved. It should be noted that the SGD classifier performed poorly across the board. The precision was 75%, and the recall was 79%. Its confusion matrix is shown in Figure 7. To compute the accuracy, the accuracy_score function from the sklearn.metrics library with the following formula:

$$\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

Where $\hat{y}$ is the predicted vector and y is the actual vector [11].

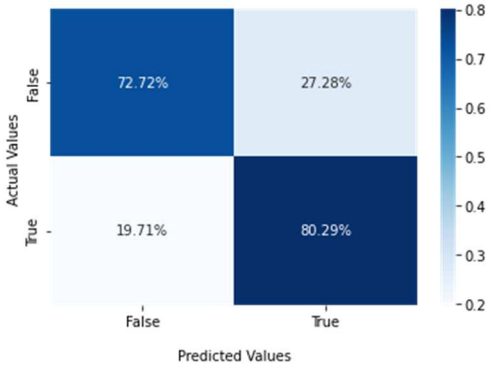

Figure 7 SGD Confusion matrix

Figure 8 shows the confusion matrix for the decision tree classifier, where an accuracy of 95% was achieved. The precision was 91%, and the recall was 100%.

Finally, with the random forest classifier, an accuracy of 96% was achieved, giving us the highest precision for unsplit data at 92% and a recall of 100%. The confusion matrix is shown in Figure 9.

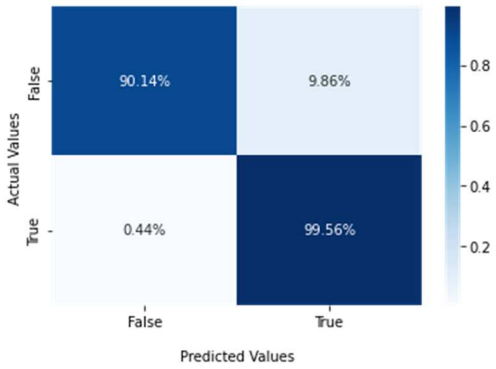

Figure 8 Decision Tree confusion matrix

Figure 9 Random Forest confusion matrix

B. *Splitting by Sex*

The effect of splitting the data by sex was minor> it was merely visible in the improvement in accuracy for the random forest classifier for female patients. Figure 10 and Figure 11 show the accuracy, precision, and recall scores for males and females, respectively.

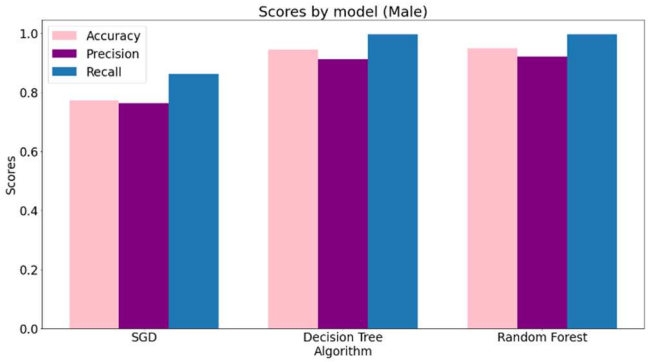

Figure 10 Scores by the model (Male)

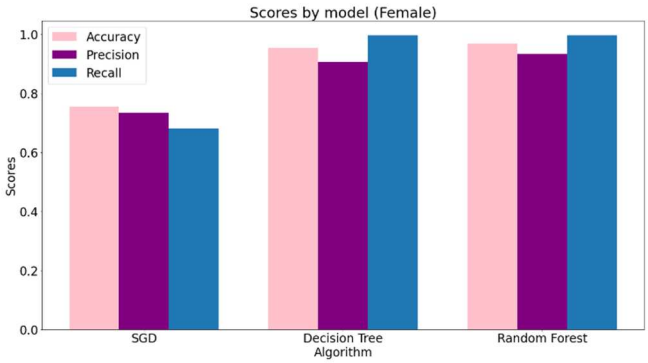

Figure 11 Scores by the model (Female)

C. *Splitting by Race*

Splitting the dataset by race showed a substantial increase in recall for decision trees and random forests (up to 100% in some cases) for all races. Figure 12 - Figure 17 show each race's accuracy, precision, and recall scores.

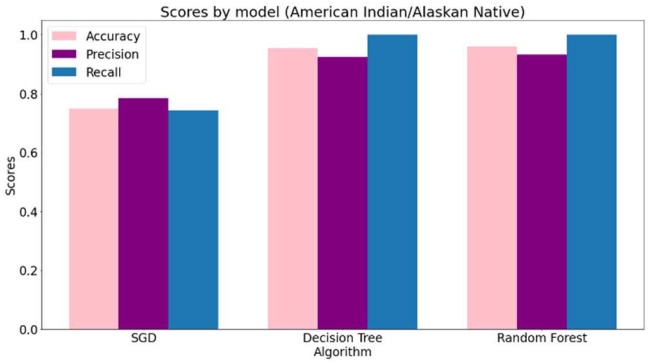

Figure 12 Scores by the model (American Indian/Alaskan Native)

Interestingly, the recall in SGD for the Asian subset was much lower than what it was for the unsplit data.

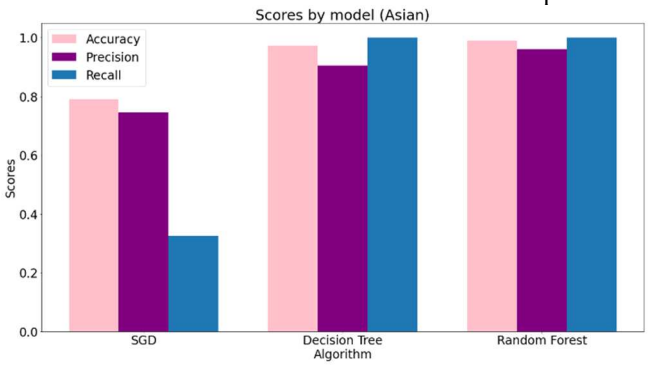

Figure 13 Scores by the model (Asian)

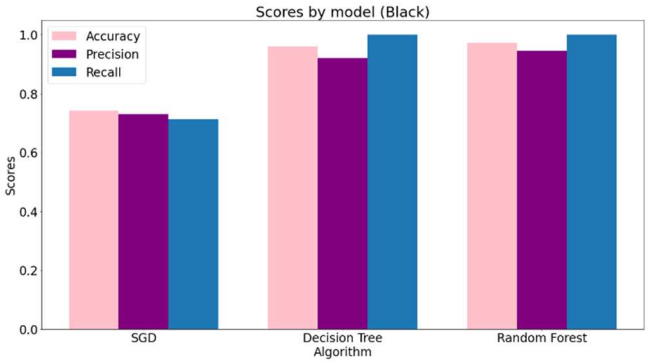

Figure 14 Scores by the model (Black)

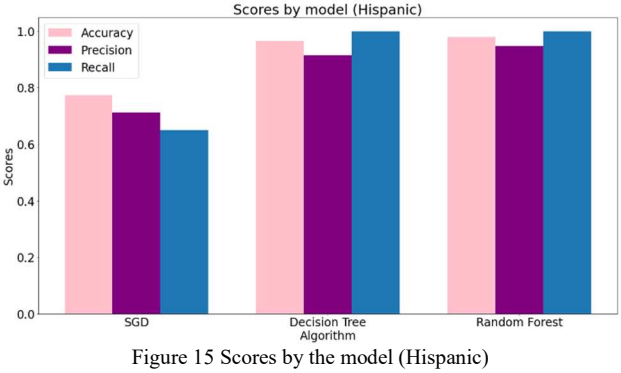

Figure 15 Scores by the model (Hispanic)

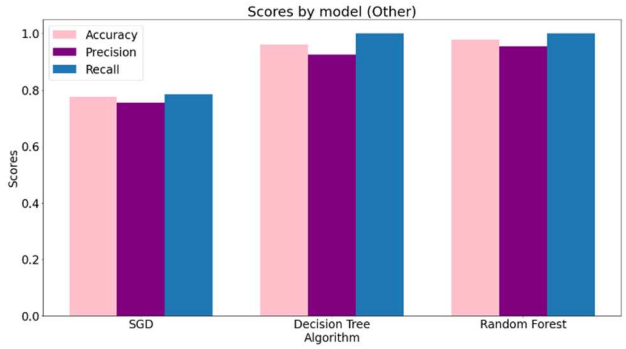

Figure 16 Scores by the model (Other)

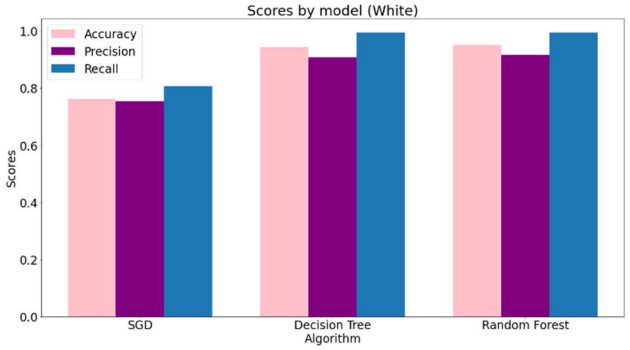

Figure 17 Scores by the model (White)

Table 3 below summarizes the results achieved in this project.

Table 3 Accuracy based on split

Split	Algorithm		
	SGD	Decision Tree	Random Forest
Unsplit	76%	95%	96%
Male	77%	94%	95%
Female	75%	95%	97%
American Indian/ Alaskan Native	75%	96%	97%
Asian	82%	97%	99%
Black	74%	96%	98%
Hispanic	76%	96%	98%
Other	76%	96%	98%
White	76%	95%	95%

VI. Conclusion

In conclusion, this project gave high-accuracy preliminary results to predict heart disease in suspecting individuals using decision trees and random forests. The results also show that it's possible to increase accuracy substantially by splitting the data according to certain attributes, such as race. The most accurate models were decision trees and random forests, achieving an accuracy as high as 99% when the data was split by race and a minimum accuracy of 94%. Stochastic gradient descent was the least accurate model, which could not achieve an accuracy higher than 82%, even when the data was split.

VII. References

[1] "Centers for Disease Control and Prevention," 7 2 2022. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm.

[2] "National Heart, Lung, and Blood Institute," 24 3 2022. [Online]. Available: https://www.nhlbi.nih.gov/health/heart-attack. [Accessed 27 5 2022].

[3] K. Pytlak, "Kaggle," Kaggle, 15 2 2022. [Online]. Available: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease. [Accessed 14 4 2022].

[4] A. Atwah and A. Al-Mousa, "Car Accident Severity Classification Using Machine Learning," in *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Zallaq, Bahrain, 2021.

[5] Z. Bitar and A. Al-Mousa, "Prediction of Graduate Admission using Multiple Supervised Machine Learning Models," in *IEEE SoutheastCon*, Raleigh, 2020.

[6] N. Abdulhadi and A. Al-Mousa, "Diabetes Detection Using Machine Learning Classification Methods," in *2021 International Conference on Information Technology (ICIT)*, Amman, 2021.

[7] R. Atallah and A. Al-Mousa, "Heart disease detection using machine learning majority voting ensemble method," *2019 2nd international conference on new trends in computing sciences (ictcs),* 2019.

[8] S. Ekız and P. Erdoğmuş, "Comparative study of heart disease classification," *2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT),* 2017.

[9] S. Mohan, C. Thirumalai and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE access 7,* 2019.

[10] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare.," *IEEE Access 8,* 2020.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.