Team 20: Heart Disease Prediction

Team Members: Pranjali Agarwal, Varsha Reddy Anugu, Nithish Goud Podeti

Motivation:

Heart disease has been one of the leading health concerns for a long time and has been steadily rising in recent years. In 2020, there was an increase in cases of heart disease by 4.1%, and 80% of these cases were preventable by early intervention [Figure 1]. Change in lifestyle, the adoption of habits like smoking, drinking, and eating junk food has led to this increase which ultimately results in a high mortality rate due to heart disease. In terms of cost incurred as well, it is predicted that by 2035, an estimated $ 1.1 trillion will be spent on treating patients with cardiovascular diseases[Figure 2]. The productivity loss is also huge and will continue to increase if necessary steps are not taken. Hence, this problem needs to be addressed with the help of Data Science. Our project focuses on predicting the chances of Heart Disease based on various factors like lifestyle, health, etc, which have been taken from the survey data by the Centers for Disease Control and Prevention for the year 2020 and can be used in different machine learning algorithms and solved by data science to assist early identification of individuals at risk for preventing the condition. Data science can play a critical role in heart disease prediction by identifying relevant features, building accurate predictive models, evaluating and interpreting the models, and providing tools for risk stratification, monitoring, and tracking. This will ultimately help in early analysis, reduce the amount to be spent in intensive care, and increase the lifespan.

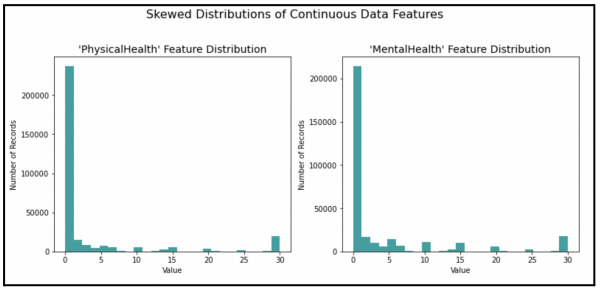
Dataset:

Heart disease prediction dataset is a collection of data that contains information about a patient's health and medical history, which is used to predict the likelihood of developing heart disease. The dataset contains a range of variables such as age, gender, blood pressure, cholesterol level, smoking habits, family history, exercise habits, and other health-related factors. These variables are used to build predictive models that can identify patterns and correlations that may indicate a higher risk of heart disease. The dataset consists of 319,795 rows and 18 columns, and the columns are as follows:

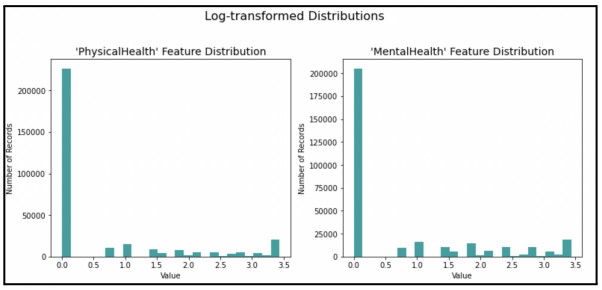
	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity
0	No	16.60	Yes	No	No	3.0	30.0	No	Female	55-59	White	Yes	Yes
1	No	20.34	No	No	Yes	0.0	0.0	No	Female	80 or older	White	No	Yes
2	No	26.58	Yes	No	No	20.0	30.0	No	Male	65-69	White	Yes	Yes
3	No	24.21	No	No	No	0.0	0.0	No	Female	75-79	White	No	No
4	No	23.71	No	No	No	28.0	0.0	Yes	Female	40-44	White	No	Yes

Methodology:

Low risks:
It comprises cleaning and preprocessing of the raw data, followed by EDA and visualizations for interesting insights. We removed the null values, one-hot encoded the categorical variables and normalized the skewed features like Physical and Mental Health (see Fig. below). The EDA and visualizations are done using count plots, box plots, scatter plots, and correlation matrices to gain insights and uncover relationships within features such as HeartDisease and Alcohol consumption, Smoking, and Sleeping habits. Refer to Figures 4, 5, & 6 for interesting visualizations.



Actual distribution having skewness



Distribution after Feature Engineering

Medium risks:
The medium-risk goal is feature engineering and application of different classification algorithms. As part of Feature engineering, we performed oversampling, changing the BMI column from numerical to categorical variable and feature selection using forward stepwise selection to get the 30 most useful features. With the feature set obtained, we trained different classification algorithms such as Logistic Regression, Random Forest, XGBoost, and Adaboost.

Feature Engineering:

- Smote sampling was used to perform oversampling as the data was unbalanced. This increased our target class of people with heart disease from 27,373 to 292,422, removing the skewness from the data.
- The BMI columns were changed from numerical to the following categories Underweight (BMI < 18.5), Normal weight (18.5 <= BMI < 25.0), Overweight (25.0 <= BMI < 30.0), Obese (30.0 <= BMI < 35.0), Extremely Obese (BMI >= 35.0), which resulted in a higher accuracy from all models.
- Feature selection was done using forward stepwise selection, understanding the correlation among all features and information gain, and as a result, the 30 most important features were used for data modeling.

Algorithms:

- **Logistic regression:** It finds the relationship between various risk factors and the presence of heart disease with input features such as demographic data, lifestyle factors, and medical history. The output of the model is a probability of having heart disease, which can then be thresholded to make a binary prediction (positive or negative).
- **Random Forest:** The algorithm uses multiple decision trees to make predictions,

and the final prediction is made based on the average of the individual trees. This helps reduce overfitting and improves the accuracy of the model.

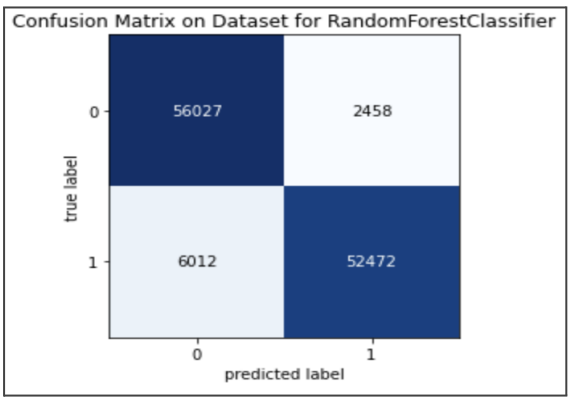
- **Xgboost:** This is an ensemble learning algorithm that uses the averaged results of multiple decision trees to make predictions. It is better in terms of high performance and ability to handle large datasets efficiently.
- **Ada-boost:** AdaBoost is useful for heart disease prediction as it can handle a mix of categorical and continuous predictors and handle imbalanced data where one class is more prevalent than the other. Additionally, it is robust to noisy data and can handle non-linearly separable data by transforming it into a higher-dimensional space.

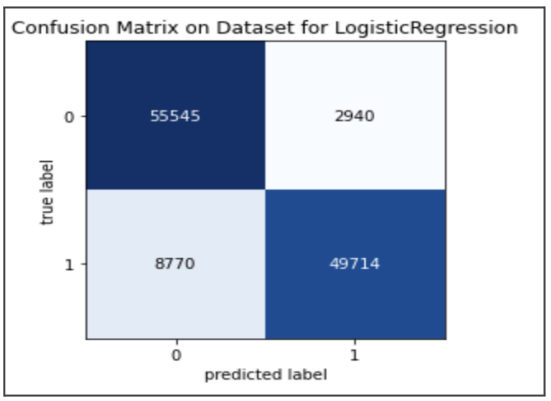
High risks:
The high-risk goal is the evaluation of the applied algorithms using different metrics and use it on test data to determine the accuracy. Evaluation of results from all the different algorithms used is done with metrics such as Accuracy, Recall, F1 score, AUC-ROC score, and Confusion Matrix. Recall is chosen instead of Precision because here Sensitivity measurement metric is of higher significance.

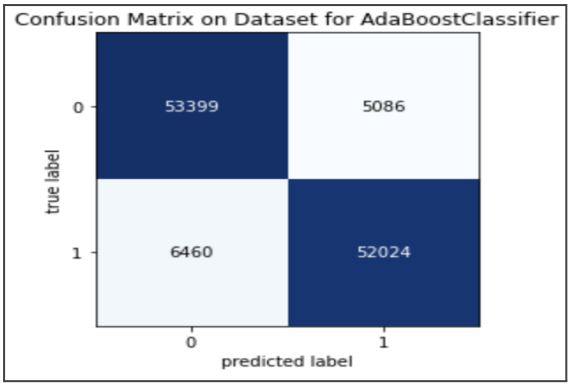
Results:

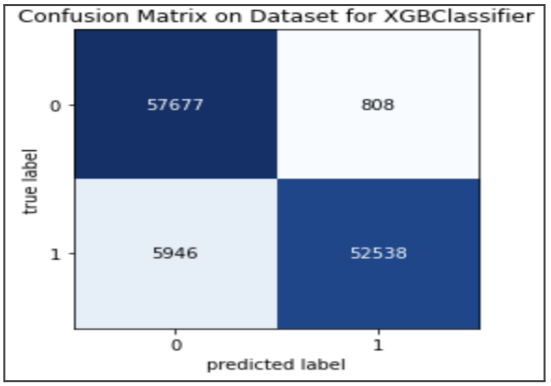
Metrics	Logistic Regression	Random Forest	XGBoost	AdaBoost
Accuracy	0.899	0.927	0.942	0.901
Recall	0.85	0.896	0.898	0.889
F1 Score	0.894	0.924	0.939	0.901
AUC ROC score	0.899	0.928	0.943	0.902

Below is the Confusion Matrix for each model:









Conclusion:

We observe that XGBoost performs the best out of all the chosen algorithms, with the highest accuracy. It also has the highest recall signifying that the less represented class of people with heart disease is being identified accurately compared to other algorithms.

Other algorithms like SVM could not be implemented, which were mentioned in our project proposal. It was computationally expensive and the accuracy was very low for any classification model due to a large number of encoded columns. Implementing more techniques for feature selection and feature engineering, SVM could have given a better result but that cost the accuracy of the other classification models. Hence, this algorithm was not a good choice due to the nature of the dataset.

Future Work:
In the future, we would like to work on artificial neural networks, including deep learning models which can handle large and complex datasets and can learn hierarchical representations that will be useful for heart disease prediction. Apart from these, a user interface can also be done from this project, which can help reach everybody and be easily accessible and easy to calculate by just providing some inputs regarding lifestyle and health factors.

References:

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501
https://towardsdatascience.com/understanding-adaboost-2f94f22d5bfe

Link to Notebook: Heart Disease Prediction

Appendix:

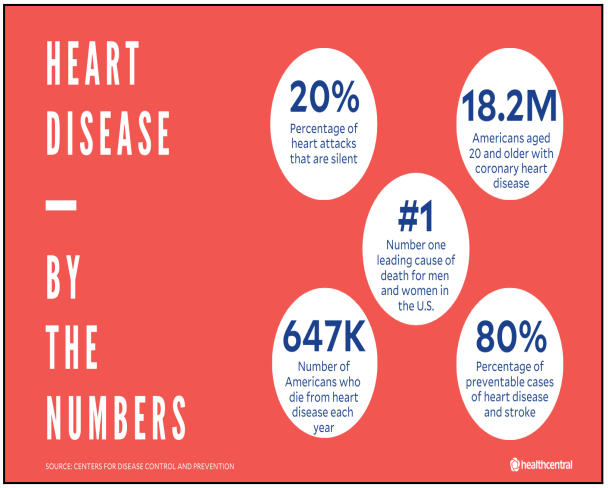


Figure 1

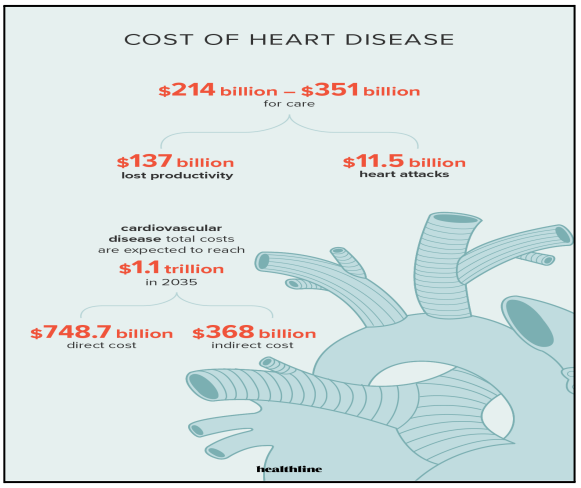


Figure 2

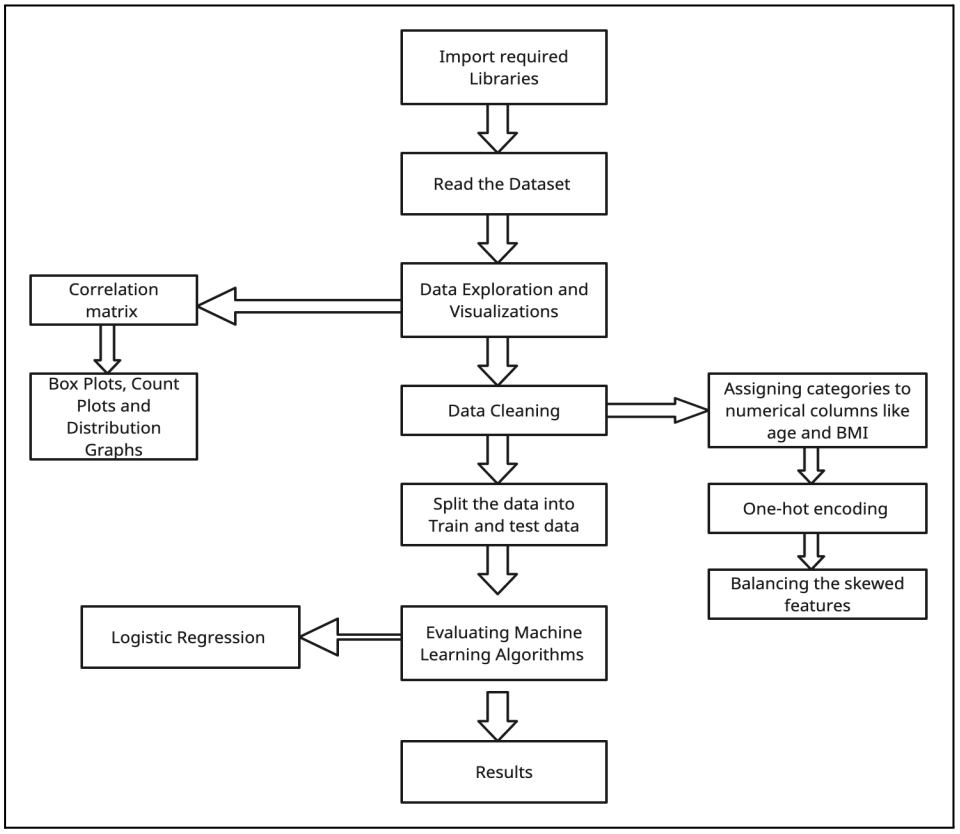


Figure 3: Flow diagram

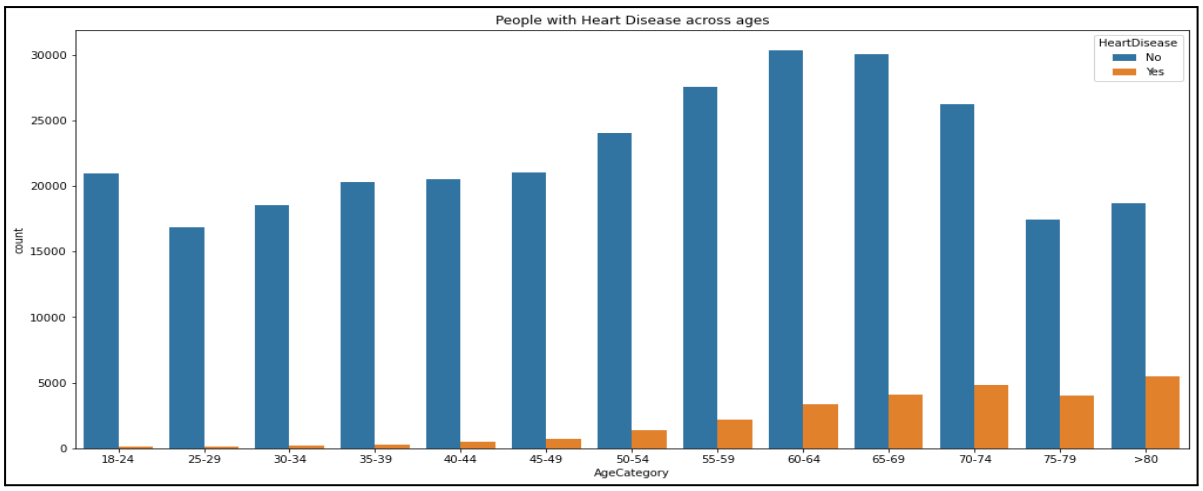


Figure 4(Age Category vs Heart Disease Probability)

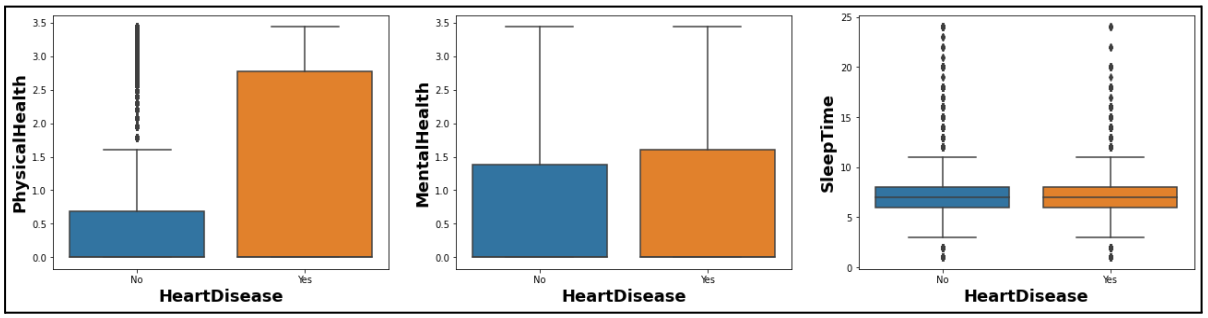


Figure 5 (distribution of the numerical columns)

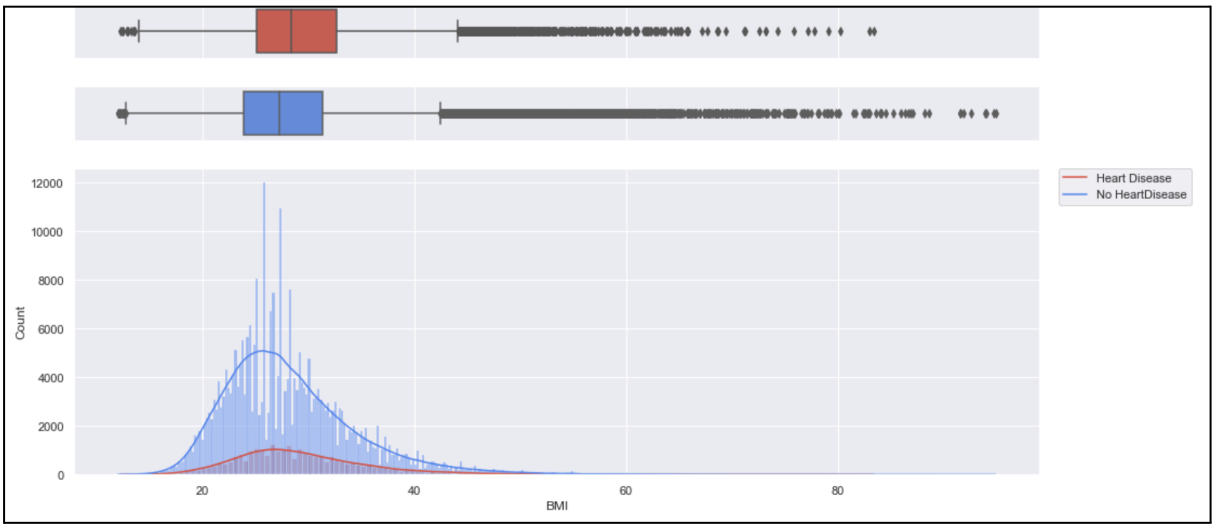


Figure 6 (distribution of BMI for each target class)