Coronary Heart Disease prediction

Abdul Afrid Mohammed
Data Science and Analytics (Computer Science)
Georgia State University
Atlanta, United States
amohammed34@student.gsu.edu

***Abstract*— Cardiovascular disease is one of the leading causes of death worldwide. According to the World Health Organization, 12 million people die from cardiovascular disease each year. Predicting cardiovascular disease is a major challenge in the field of medical data analysis. Artificial intelligence and machine learning (ML) are poised to have a major impact on the healthcare field. ML has been shown to be effective in helping to make decisions and predictions from the vast amount of data generated by the healthcare industry. Heart disease can be predicted based on a variety of factors, including heart rate, cholesterol levels, blood pressure, and other common attributes. This project aims to identify the most relevant risk factors for heart disease and predict the overall risk using various machine learning techniques.**

Keywords—Logistic Regression, KNN, Model Selection, SVM, k-fold cross validation, DTC, Data Preprocessing.

I. Introduction

Heart disease is the leading cause of death for both men and women in the world, and it is important to be able to identify people who are at risk so that they can take steps to prevent or manage the disease. This issue is related to making changes to the lifestyle to be able to live better. I am interested in lifestyle changes, which motivates me to take over this project. I believe that by making small changes to our lifestyles, we can reduce our risk for heart disease and live longer, and healthier lives.

The problem statement of this project is to build a machine learning model that can predict whether a person is at risk of developing heart disease based on their medical history and demographics.

The goal of this project is to provide a tool that can help healthcare professionals identify patients who may need further testing or treatment to prevent or manage heart disease. Additionally, the project aims to raise awareness about the risk factors for heart disease [1] and encourage individuals to make lifestyle changes to reduce their risk.

The project will use a machine learning model to predict the risk of heart disease based on a patient's medical history and demographics. The model will be trained on a dataset and will be tested on separate data which the model has not seen before.

I will use multiple machine learning models [2] and evaluate their performance of them to choose the best-performing model to help us predict the risk of developing heart disease.

II. Materials and Methods

A. Data Explanation

Data used in the project is a publicly available dataset obtained from Kaggle. The dataset contains over 4,240 records, each of which represents a patient. It includes 16 columns, of which 15 are attributes that describe the patient and their medical history. Data types of continuous features are integer and float. Categorical features are the binary features describing if a feature is present or not. The shape of the data is (4240, 16).

B. Data Preprocessing

Data quality checks were performed to ensure the reliability of the data. Duplicate instances were checked for and none were found. Missing values [3] were then checked for and mean imputation was performed on them. The data was also reviewed using 'data.describe()' to understand the basic statistics of the data.

Data Exploration was conducted on the obtained data from the previous steps to understand the content and structure of the data. This was carried out by visualizing [5] the input features.

Histograms were used to visualize the distribution of each feature. This helped in understanding the range and peak distribution of the input variables. For example, the histogram of the age feature shows that most of the patients are between the ages of 40 and 60. The histogram of the blood pressure feature shows that most of the patients have blood pressure that is high. As you can see, the age feature is slightly right-skewed, meaning that there are more younger people in the dataset than older people. The blood pressure feature is slightly left-skewed, meaning that there are more people with lower blood pressure in the dataset than people with higher blood pressure. The cholesterol feature is normally distributed.

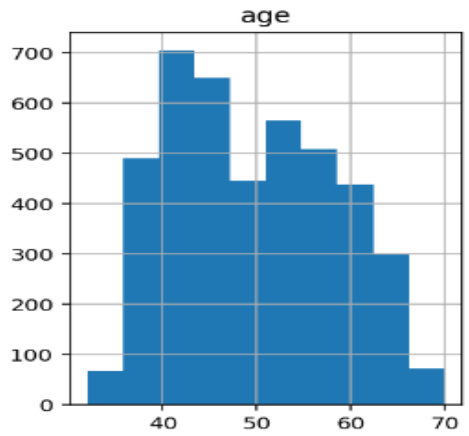

Fig 1: Histogram of AGE feature.

A heatmap is a graphical representation of the correlation between two or more variables. It is created by creating a matrix of cells, where each cell represents the correlation between two variables. As you can see, the age and blood pressure features are the most correlated with the target

variable. This suggests that these two features are the most important for predicting heart disease.

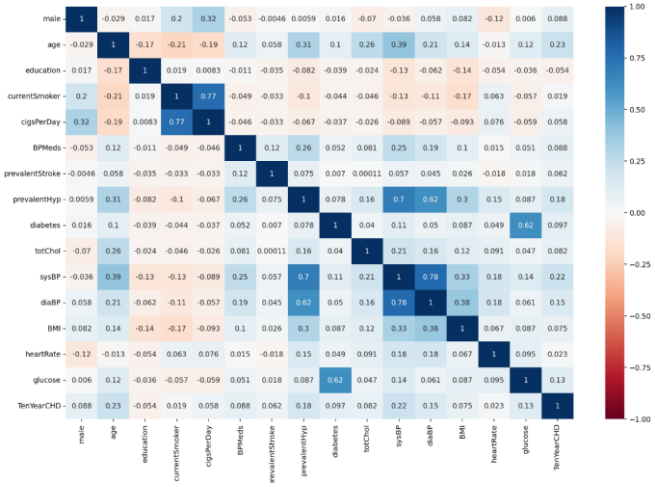

Fig 2: Heatmap of the features representing correlation.

Pairplot was used to visualize the distribution of the features and the relation among them. This helped in identifying any potential relationships between the features. For example, the pair plot of the age and blood pressure features shows that there is a positive correlation between these two features. This means that as age increases, blood pressure also tends to increase.

Overall, data exploration helped in gaining a better understanding of the data. This information was used to select the features that were most relevant to the target variable and to develop a model that could predict the likelihood of a patient developing heart disease.

Feature selection [7] has been performed to select the k best features. The method used was a filter method with a chi-square test which is a statistical measure to calculate the relation and the score between the input features and the target variable.

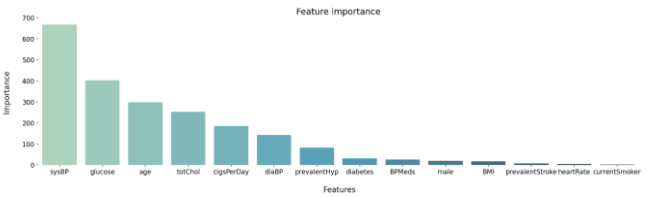

Fig 3: Feature Importance.

I used the IQR method to identify outliers [4] in the dataset. The IQR is a robust method for identifying outliers that is not affected by outliers. The IQR is calculated by finding the interquartile range, which is the difference between the third and first quartiles. The IQR is then used to identify values that are more than 1.5 times the IQR away from the first or third quartile. I found that there were several outliers in the dataset. The most common outliers were in the sysBP, glucose, and totChol features. These outliers were real values and I clamped them to the upper and lower bounds. This means that I replaced any values that were outside of the interquartile range with the upper or lower bound of the interquartile range. Box plots are used to visualize the outliers before and after the clamping method is performed.

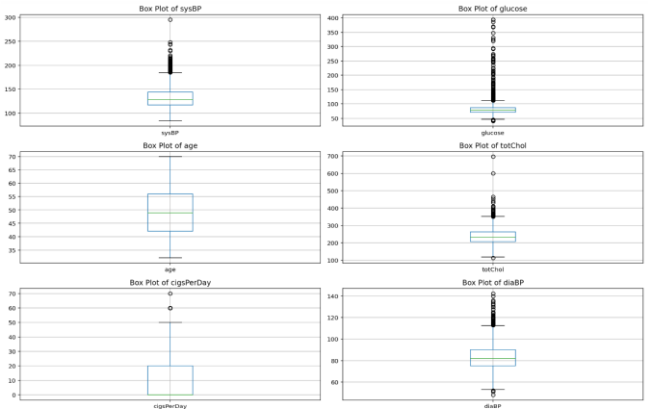

Fig 4: Box plot of features before clamping outliers.

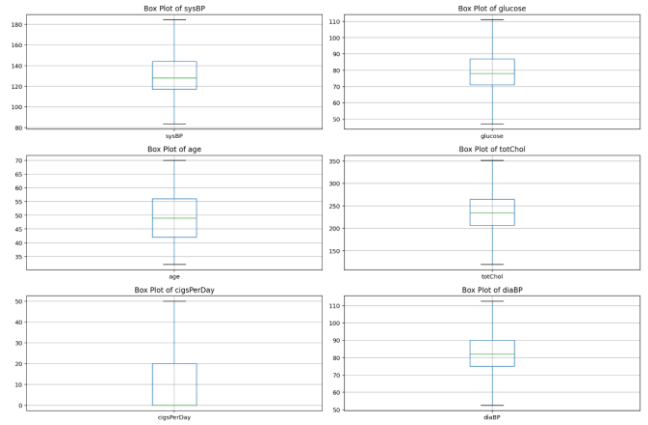

Fig 5: Box plot of features after clamping outliers.

The data were transformed using the standard scaler. This is a method of normalization that scales each feature to have a mean of 0 and a standard deviation of 1. This makes the features comparable to each other and helps to improve the performance of machine learning algorithms. The standard scaler is a more reliable method of normalization than other methods, such as min-max scaling, because it is not affected by outliers. Outliers are data points that are very different from the rest of the data. They can skew the results of machine learning algorithms, so it is important to remove them before normalizing the data. Once the outliers were removed, we used the standard scaler to normalize the data.

C. Data Mining

The data was split into a training set and a test set. The training set was used to train the machine learning model, and the test set was used to evaluate the model's performance. The training set consisted of 80% of the data, and the test set consisted of 20% of the data. The random_state parameter was set to 57 to ensure that the data was split randomly. The data was split in this way to ensure that the model was not overfitting to the training data. Overfitting occurs when the model learns the training data too well and is unable to generalize to new data. By splitting the data into a training set and a test set, we can assess the model's performance on data that it has not seen before. This helps us to ensure that the model is not overfitting.

The data was imbalanced, with the majority class (no heart disease) having 75% of the data and the minority class (heart disease) having 25% of the data. This imbalance can lead to models that are biased toward the majority class. To address

this imbalance, the minority class was up-sampled. Upsampling is a technique that increases the number of samples in the minority class by creating new samples from existing samples. This was done by randomly selecting samples from the minority class and duplicating them. This resulted in a balanced dataset with both classes having 50% of the data. This balanced data was used to train the machine learning model.

The data obtained from the previous upsampling step was used to prepare the training set for the model. The training set consisted of 50% of the data, with both classes having an equal number of samples. The data was split into features and labels. The features were the independent variables that were used to predict the label. The label was the dependent variable that was to be predicted. The data were normalized to ensure that all the features were on the same scale. The training set was then shuffled to ensure that the data was presented to the model in a random order. This helps to prevent the model from learning patterns that are not representative of the data.

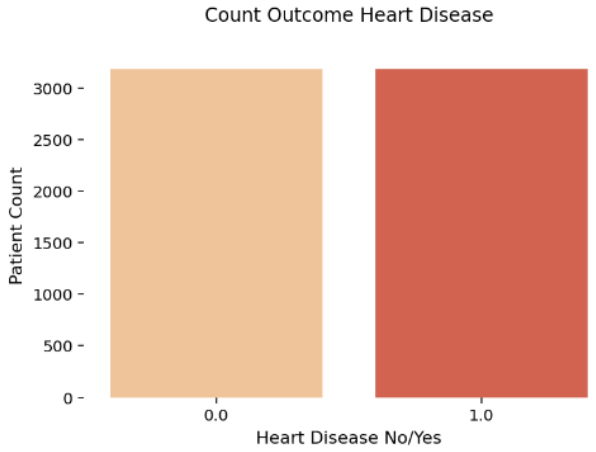

Fig 6: Bar plot of the dataset partitioned by class labels.

D. *Models and Algorithms*

Four different classification algorithms were evaluated to predict the likelihood of heart disease: logistic regression, support vector machine (SVM), decision tree, and k-nearest neighbors (KNN).

Logistic regression is a linear model that is used to predict the probability of a binary outcome. It is a simple and effective model that is often used as a baseline for other models.

Support vector machine is a non-linear model that is used to find a hyperplane that separates the two classes. It is a powerful model that can be used to solve a variety of classification problems.

The decision tree is a non-linear model that is used to create a tree-like structure of decisions. It is a simple and intuitive model that can be used to understand the relationships between the features and the outcome.

K-nearest neighbors is a non-parametric model that is used to find the k most similar instances in the training set to a new instance. It is a simple and robust model that is often used for tasks such as recommender systems.

Hyperparameter tuning [8] is the process of finding the optimal values for the hyperparameters of a model. Hyperparameters are parameters that are not learned from the data but rather are set by the user. The goal of hyperparameter tuning is to find the values for the hyperparameters that result in the best performance of the model. GridCV is a technique used to find the best combination of hyperparameters for a model. It works by creating a grid of possible hyperparameter values and then evaluating the model on each combination of hyperparameters. The combination of hyperparameters that results in the best performance of the model is then chosen.

E. *Evaluation and interpretations*

The models were evaluated using a variety of metrics [9], including accuracy, precision, recall, and f1 score. The model was also evaluated using k-fold cross-validation [6].

K-fold cross-validation is a technique used to evaluate the performance of a machine-learning model. It is a more robust evaluation method than using a single training and test split because it helps to mitigate the effects of overfitting. To perform k-fold cross-validation, the data is first split into k-folds. Then, the model is fit on k-1 folds and evaluated on the remaining fold. This process is repeated k times, and the results are averaged. In this case, we used k=5 folds. This means that the data was split into 5 folds, and the model was fit on 4 folds and evaluated on the remaining fold. This process was repeated 5 times, and the results were averaged.

Accuracy is a measure of how well the model predicts the target variable. It is calculated by dividing the number of correct predictions by the total number of predictions.

Precision is a measure of how precise the model is. It is calculated by dividing the number of true positives by the number of predicted positives.

Recall is a measure of how sensitive the model is. It is calculated by dividing the number of true positives by the number of actual positives.

F1 score is a measure of the balance between precision and recall. It is calculated by taking the harmonic mean of precision and recall.

We can also use the confusion matrix to carry out the above metrics.

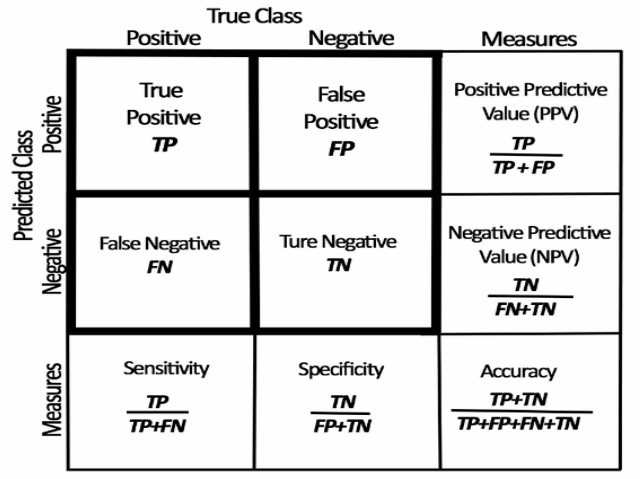

Fig 7: Confusion matrix and Evaluation measures.

III. RESULTS

The results here show the Accuracy per number of folds in k-fold cross-validation.

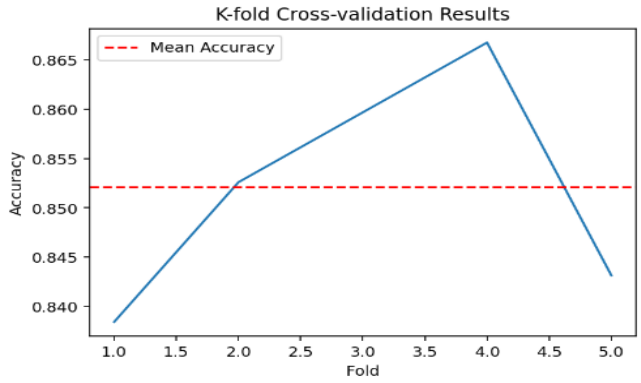

Fig 8: Logistic regression k-fold vs Accuracy

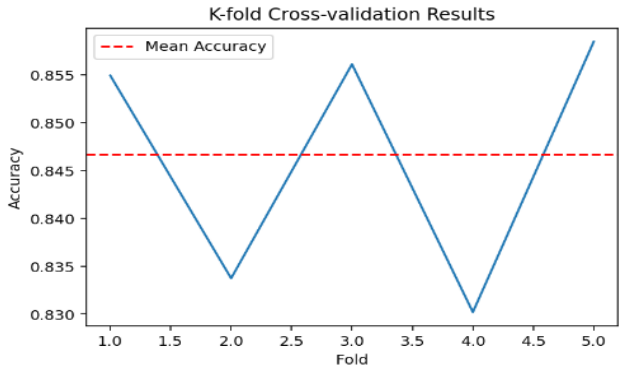

Fig 9: Support Vector Classifier k-fold vs Accuracy

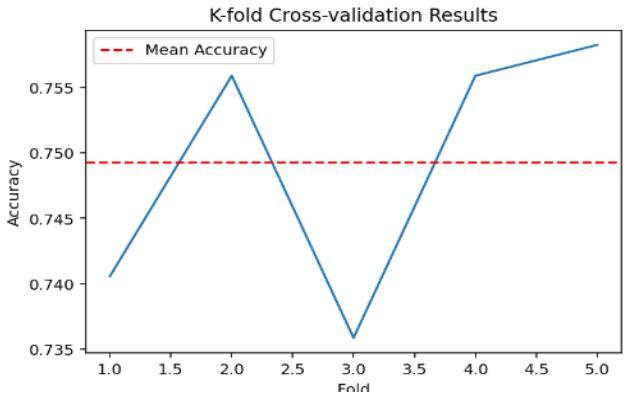

Fig 10: Decision Tree Classifier k-fold vs Accuracy

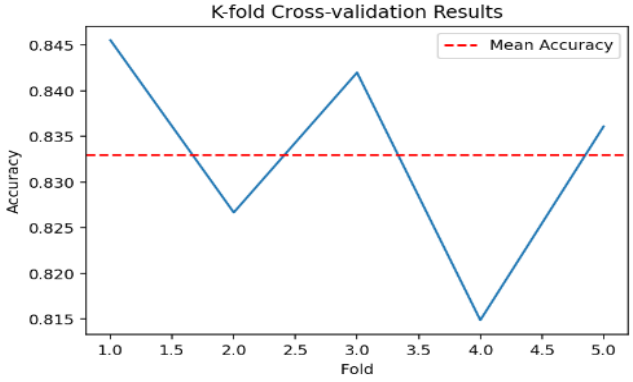

Fig 11: KNN Classifier k-fold vs Accuracy

Based on the results I chose Logistic Regression as my model for this problem. The evaluation metrics and confusion matrix of it are as follows.

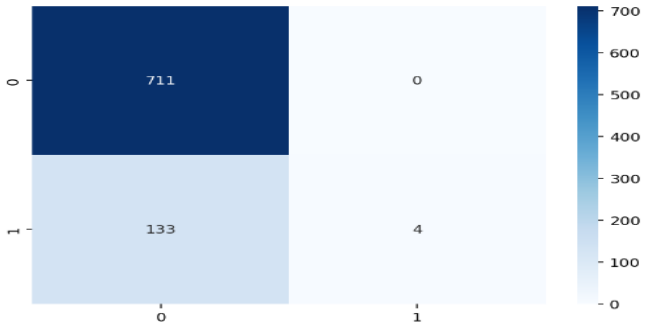

Fig 12: Confusion Matrix of Logistic Regression model

```
LogisticRegression(C=0.3)

Mean accuracy score for LRM with k-fold cross validation: 85.0
Mean f1 score for LRM with k-fold cross validation: 8.0
Mean precision score for LRM with k-fold cross validation: 80.0
Mean recall score for LRM with k-fold cross validation: 4.0
```

Fig 13: Evaluation Metrics of Logistic Regression Model

```
GridSearchCV(cv=5, estimator=LogisticRegression(),
             param_grid={'C': [0.1, 0.3, 0.5, 0.6, 1, 10, 100]})

Best hyperparameters: {'C': 0.3}
Best mean cross-validation score: 0.8517727474478474
```

Fig 14: GridCV results of the Logistic Regression model.

A. *User Interface*

A web user interface was created using the Stream-lit framework [10]. The user interface allows a user to input their medical history and demographics. Based on the inputs provided by the user, a trained logistic regression model will be applied to the data and the risk of developing heart disease for that person will be predicted.

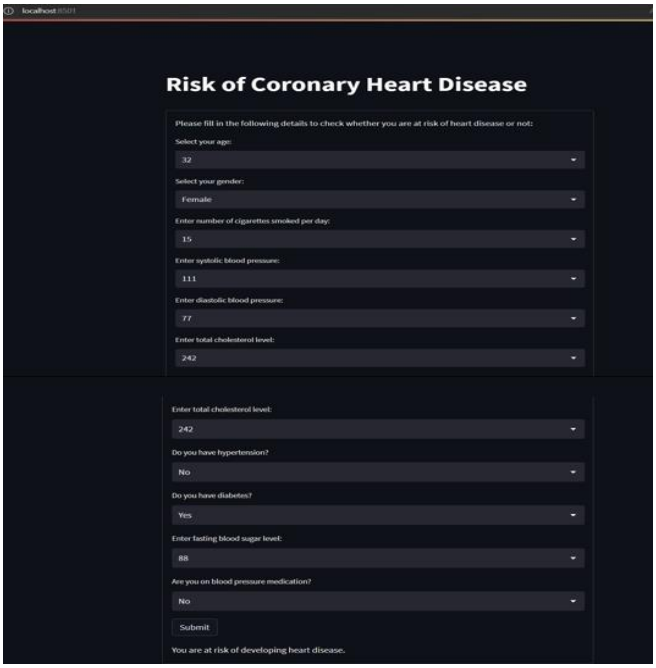

Fig 15: User Interface for predicting the output.

IV. DISCUSSION AND CONCLUSION

I evaluated the performance of four different machine learning models for predicting heart disease. It has been observed that the features of age and blood pressure are highly impacting the risk of developing heart disease over a period meaning as they increase so does the risk.

The results of the evaluation showed that the logistic regression model was the best-performing model. The models were evaluated using a variety of metrics, including accuracy, precision, recall, and f1 score. The models were also evaluated using k-fold cross-validation. The logistic regression model achieved an accuracy of 85% on the test set. The SVM model also performed well, achieving an accuracy of 85%. The decision tree classifier and the KNN classifier performed less well, achieving accuracies of 76% and 83%, respectively.

It is important to note that these results are based on a small dataset. Further research is needed to validate these results on a larger dataset. The web user interface is still under development, but it has the potential to be a valuable tool for both individuals and healthcare professionals.

Overall, the results of this project are promising. The logistic regression model is a good predictor of heart disease, and a web user interface is a valuable tool for individuals and healthcare professionals. Further research is needed to validate these results on a larger dataset, but the results of this project suggest that these tools have the potential to improve the prevention and treatment of heart disease

V. REFERENCES

[1] *JACC: Cardiovascular Disease Burden, Deaths Are Rising Around the World*. (2020, December 9). Institute for Health Metrics and Evaluation. https://www.healthdata.org/news-release/jacc-cardiovascular-disease-burden-deaths-are-rising-around-world

[2] Gong, D. (2022, July 12). *Top 6 Machine Learning Algorithms for Classification*. Medium. https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

[3] Badr, W. (2019, January 12). *6 Different Ways to Compensate for Missing Data (Data Imputation with examples)*. Medium. https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779

[4] *Data Analytics Explained: What Is an Outlier?* (2021, October 5). CareerFoundry. https://careerfoundry.com/en/blog/data-analytics/what-is-an-outlier/

[5] Chakravarthy, S. (2020, July 8). *Data Visualization for Machine Learning*. Medium. https://towardsdatascience.com/data-visualization-for-machine-learning-d1f906664f56

[6] Z., & posts by Zach, V. A. (2020, November 4). *An Easy Guide to K-Fold Cross-Validation - Statology*. Statology. https://www.statology.org/k-fold-cross-validation/

[7] M. (2022, December 9). *Feature Selection (Data Mining)*. Feature Selection (Data Mining) | Microsoft Learn. https://learn.microsoft.com/en-us/analysis-services/data-mining/feature-selection-data-mining

[8] Pandian, S. (2022, February 22). *A Comprehensive Guide on Hyperparameter Tuning and its Techniques*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/02/a-comprehensive-guide-on-hyperparameter-tuning-and-its-techniques/

[9] Mankad, S. (2020, November 24). *A Tour of Evaluation Metrics for Machine Learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/11/a-tour-of-evaluation-metrics-for-machine-learning/

[10] *A Beginners Guide To Streamlit - GeeksforGeeks*. (2020, November 24). GeeksforGeeks. https://www.geeksforgeeks.org/a-beginners-guide-to-streamlit/