CSE422  
Artificial Intelligence Project

Introduction

Diabetes is a long-term metabolic problem that affects a lot of people all over the world. It is marked by high blood glucose levels caused by either not enough insulin being made or the body not being able to use insulin well. Diabetics need to be diagnosed and treated as soon as possible to avoid problems and improve their quality of life.

In addition to the invention of machine learning and AI, there have been big steps forward in making models that can help doctors diagnose diabetes. Using the power of machine learning algorithms, we suggest a new way to figure out if someone is likely to get diabetes based on their activities and physical traits.

The principal objective of this project is to make a reliable and accurate predictive model that can figure out a person's chance of getting diabetes by looking at their lifestyle and physical traits. By using this model, people and health care workers can find people who might be at risk for diabetes. This can lead to early intervention and more effective preventive measures.

Related Work

The use of machine learning has been increasing day by day in the field of diagnosis and health care. There are various approaches of machine learning which have proved to be successful in predicting various diseases. The algorithms such as Deep neural network (DNN), Decision Tree, Support vector, Random Forest are mostly used in diagnosis related works [4]. We have used linear and logistic regression, Decision tree regression and classifier, Random forest regression and classifier and neural network models in our diabetes prediction model. The dataset that we are using consists of multiple features where one can set targets on different parameters. One of the related works includes the prediction of stroke as the primary target and hypertension as well as diabetes [8]. The models were trained on various classifiers including k-nearest neighbor, Logistic Regression, Decision Tree, Random Forest classifiers, Naive Bayes classifiers. The average accuracy for the aforementioned techniques is around 72% which varies with our work, 75%, due to different approaches in data preprocessing.

Methodology

## Dataset Description

The information we used has been retrieved from the Kaggle repository. In this dataset, there are 70692 instances as well as 17 feature variables and 1 target variable.

Using structured questionnaires, clinical assessments, and laboratory tests, information regarding the activities, physical attributes, and diabetes status of participants is collected. To ensure confidentiality and anonymity, all data collected stick to applicable privacy regulations and moral guidelines.This Diabetes Determination Through Activities and Physical Traits dataset is a useful tool to examine how well living activities and physical traits can predict the risk of diabetes. Using this dataset, the machine learning model developed for this project hopes to provide accurate and timely diabetes risk evaluation, aiding in early detection and proactive disease management.

## Data Pre-Processing

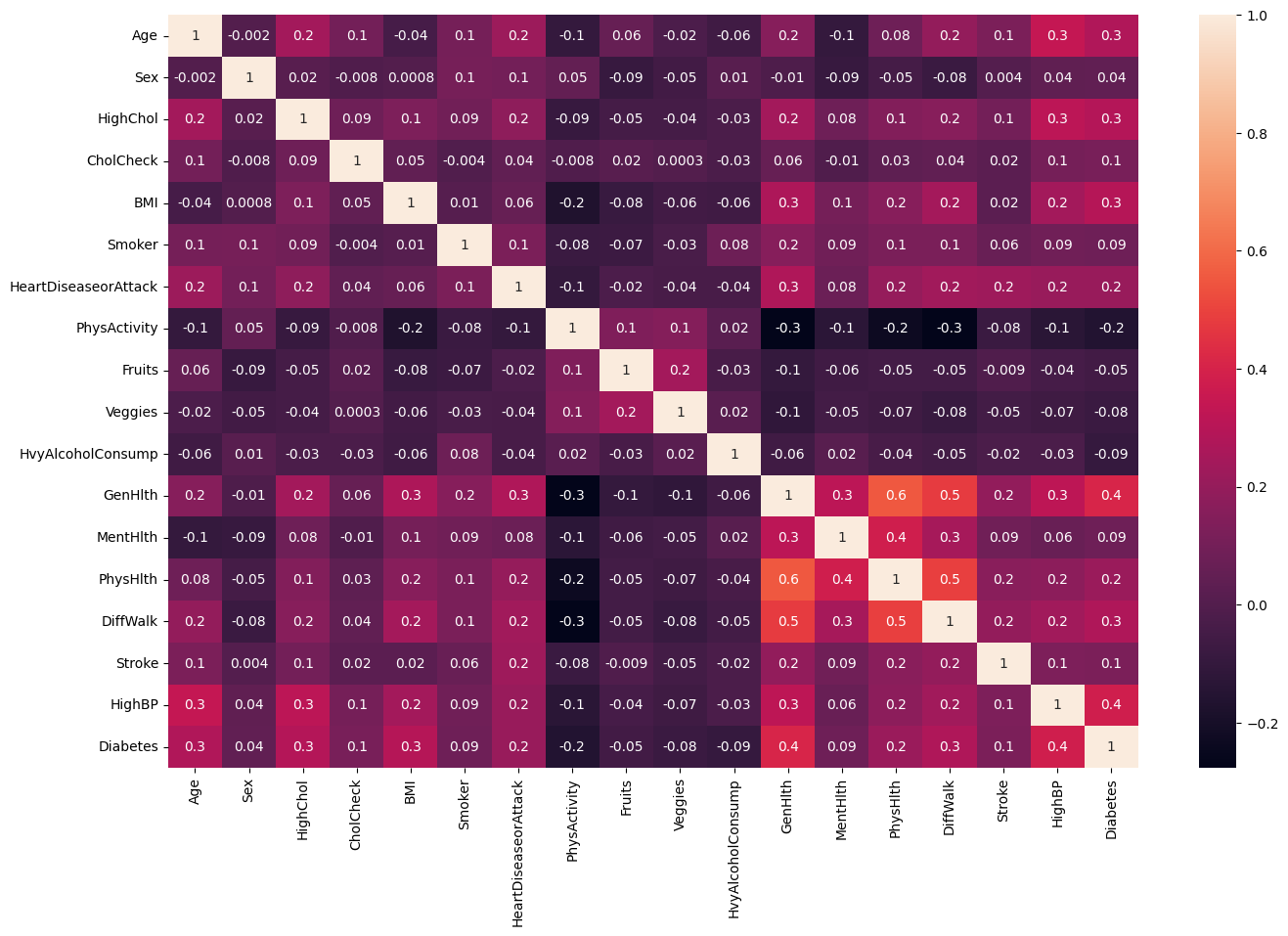
### Dropping unnecessary values

Null Verification**:** Our analysis has encompassed all instances of null values within the dataset. The presence of null values in the dataset delays the ability to effectively train the data. The presence of a null value in our data can significantly impact the accuracy of our predictions, potentially resulting in a substantial deviation from the intended outcome. Thus, we must either eliminate the values, generate random values, or replace the values using mean, median, etc. procedures. We have used the "mean" procedure to replace those values in our dataset.

### Feature Selection:

In feature selection, machine learning models try to find and pick the most important features from a dataset while getting rid of the less important or redundant ones. This step is important because adding extra factors can hurt the performance and accuracy of the model as a whole.

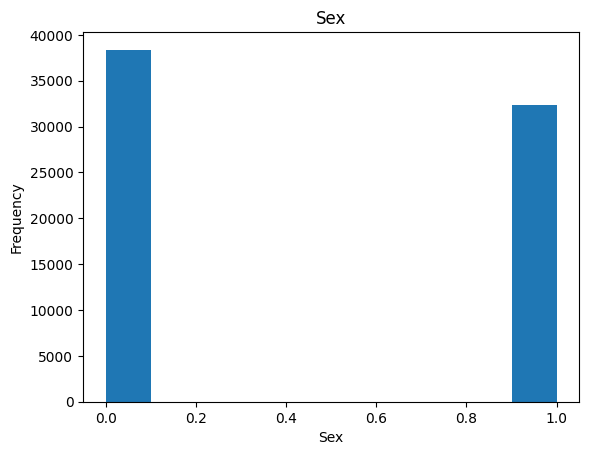
Correlation coefficients are often found for pairs of traits to learn more about how they relate to each other and how they affect each other. The correlation coefficient ranges from -1 to 1, where a value of -1 shows a strong negative correlation (i.e., as one feature decreases, the other feature increases), a value of 1 shows a strong positive correlation (i.e., as one feature increases, the other feature also increases), and a value of 0 shows that there is no correlation between the features.When two features are found to be linked, it means that we can guess or estimate the value of one feature based on the other. During feature selection, this duplication can be taken out to make the model easier to understand and more efficient.From our correlation matrix, we can see that “Sex” has comparatively insignificant relation with any of the other features. So, we dropped the “SEX” from our feature list.



### Categorical Features

In machine learning, data often includes words that describe the data so that people can understand it better. But it can be hard to work such information into the study. There are many machine learning methods that can work with category values as they are, but there are also a few that can only work with numbers. Because of this, the categorical data needs to be changed into a numeric version so that machine learning algorithms can process and analyze the names.

Let's look at the "Sex" category feature in our dataset as an example. This property shows if a person in the information is marked as "Male" or "Female." We used a method called Labelencoder to turn these gender labels into a shape that machine learning systems could use. With this method, we turned "Male" and "Female" into numbers, giving "1" to "Male" and "0" to "Female."



## Applied Models

### **Linear Regression**

Linear regression is a statistical method for figuring out how two factors are related by fitting a linear equation to the data that is given. One variable in this method is called the "independent variable," and the other is called the "dependent variable." A scatter plot can be used to figure out how strong the link is between these factors.

If there seems to be no link between the proposed explanatory variable and the dependent variable, fitting a linear regression model to the data is unlikely to give a useful model. In these situations, a linear equation might not be the best way to show how the factors are related.

The correlation coefficient is used to measure how strongly the factors in the recorded data are related to each other. It is a number between -1 and 1 that tells us a lot about the size and direction of the link between the two factors.

In the equation Y = b0 + b1X for linear regression, the variable X is called the independent variable, and the variable Y is called the dependent variable. The equation shows a line, where b0 is the intercept, which is the value of Y when X is equal to zero, and b1 is the slope, which shows how quickly Y changes in response to changes in X.

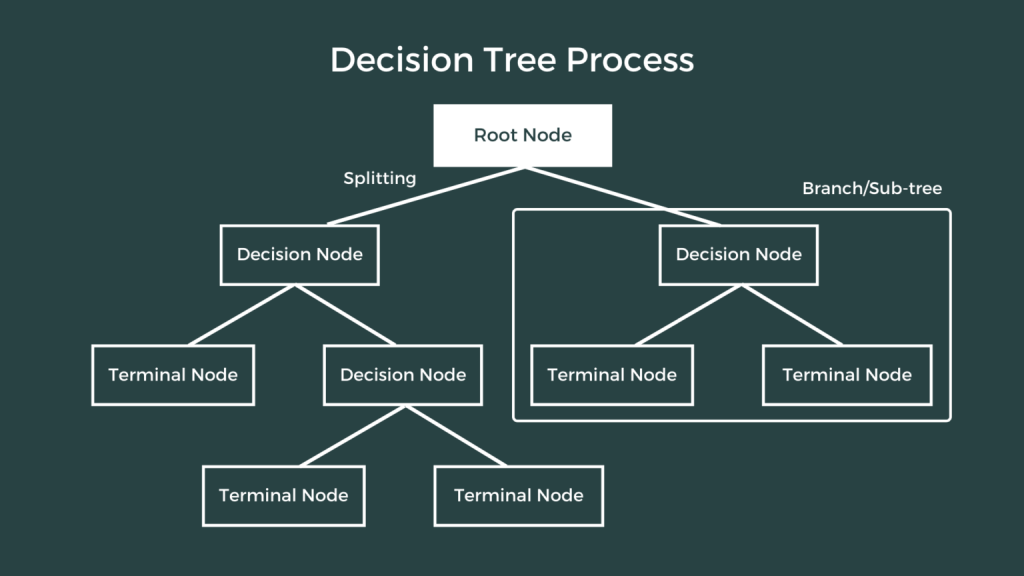
**Y = b0 + b1x1 +b2x2 + b3x3 … . + bnxn**

### **Logistic regression**

Logistic regression is a type of statistical modeling that is used to classify things into two groups. It models the link between the input factors and the probability of a binary result, such as "yes" or "no." The model calculates values for each input variable and uses a logistic function to turn a linear combination of the inputs into a chance between 0 and 1. By figuring out how to best use the coefficients, the model predicts the likelihood of the result and uses a decision level to make either yes or no predictions. Logistic regression is easy to understand and quick, but it assumes that the link between the inputs and the log-odds of the result is linear. It is often used to classify things into two groups. It shows how the input factors affect the probability of the result.

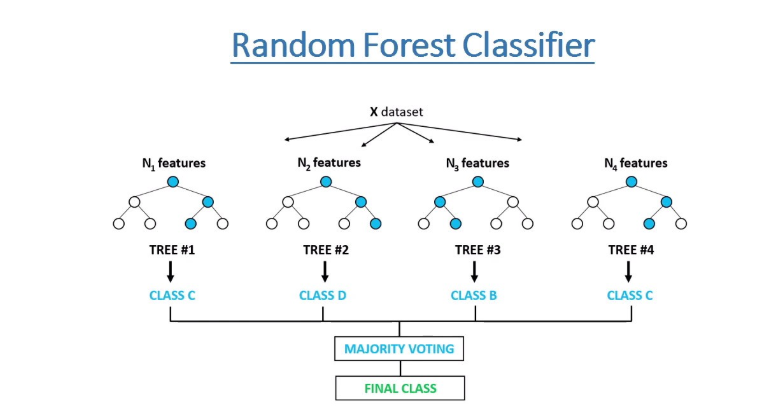
### **Decision Tree Regressor**

Machine learning uses decision trees to do both classification and regression jobs. They guess goal values or classes using a structure that looks like a tree. The goal variable in a regression tree can be a constant number. The tree is made from the top down, with each branch indicating a binary split based on certain conditions. When the forecast is a continuous number, regression trees can handle it.



### **Random Forest Regressor**

Random forest regression is a widely utilized machine learning method that can be used for both classification and regression problems. It includes building multiple decision trees with different samples from the information. In classification tasks, the estimate is based on the majority vote of the trees, while in regression tasks, it is based on the average number. Random forest can handle datasets with continuous factors and works well even when there are missing or partial data. By adding the guesses of all the trees, random forest gives more accurate results than other algorithms and learns in less time. This method works best for regression problems with fixed factors. Overall, random forest regression is a guided method of machine learning that uses the information from various decision trees to make accurate predictions.



### **Neural networks**

Neural networks are strong algorithms that are based on the way the human brain works. They are used for jobs like recognising patterns, making predictions, and sorting data. They are made up of fake neurons that are linked together and set up in layers, with input, secret, and output layers. Neurons do calculations by adding weights to inputs and using activation functions to create outputs. Training uses both forward and backward transmission to change weights and reduce mistakes. Deep neural networks with many hidden layers have changed the way deep learning works. They are especially good at computer vision and natural language processing. Neural networks automatically learn traits from data and can handle complex data sets, but training can be hard on computers and can lead to overfitting.

## Dataset and Performance:

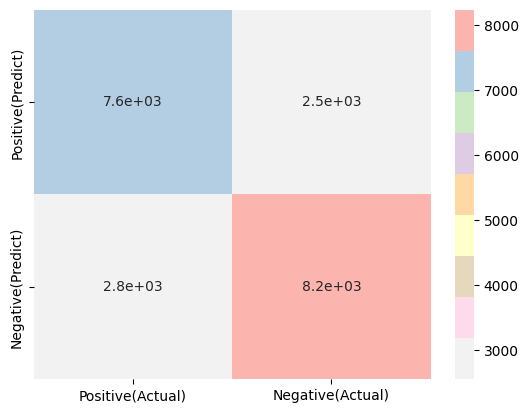
## After preprocessing our dataset, we have used some metrics to compare our models. We know, Linear Regression model works better when dependent variables are continuous in nature. However, in our dataset, most of the values are discrete values (0 or 1), thus, the Linear Regression model could not perform well in our dataset. Therefore, we have applied these metrics on the Logistic Regression, Decision Tree Regressor, Random Forest Regressor and Neural Network (Multi-layer Perceptron) Classifier model.

The **accuracy metric** signifies the accuracy of the model. The **precision** indication gives us the number of true positive values divided by the total number of positive predictions. The **recall** **matric** signifies the total number of positive values predicted successfully by the model divided by the total number of positive samples. Also, the **f1 score** identifies the performance of the model using the precision and recall value.

* precision =
* recall =
* f1 score =

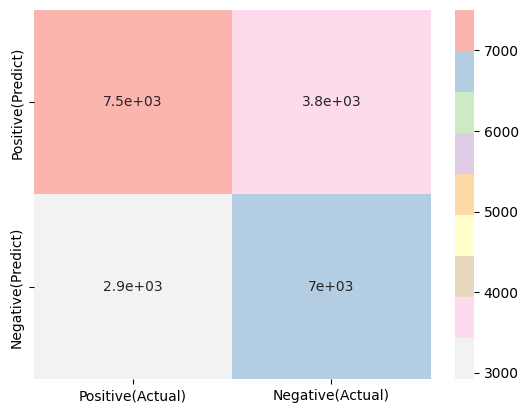
We have also applied a confusion **matrix** in our dataset as confusion matrix helps in visualizing the results of a classification problem by providing a table arrangement of the various outcomes of the prediction. It displays a table of each classifier's predicted and actual values. By using the heatmap method, we have generated the visual representation of the confusion matrix of our models.

**Logistic Regression:**



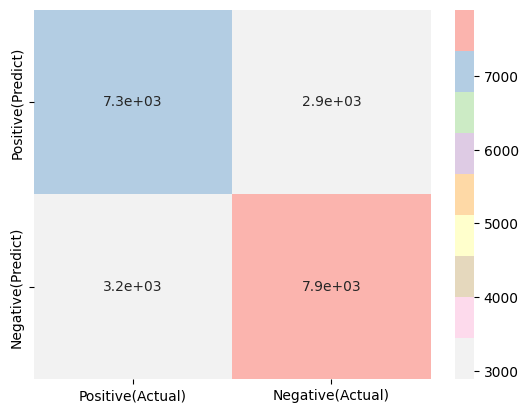
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy (%) | F1\_score (%) | Precision (%) | Recall (%) |
| Logistic | **74.67** | **75.41** | **74.45** | **76.39** |

**Decision Tree Classifier:**

****

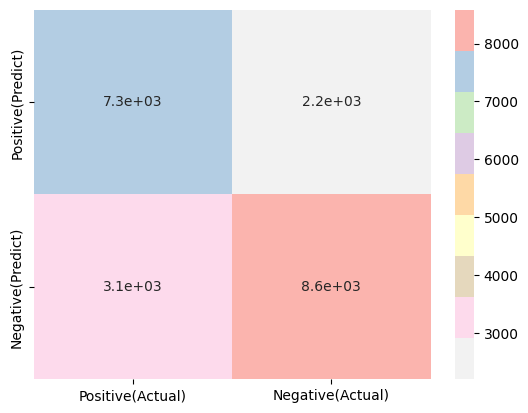
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy (%) | F1\_score (%) | Precision (%) | Recall (%) |
| Decision Tree | **68.29** | **67.44** | **70.5** | **64.64** |

**Random Forest Classifier:**



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy (%) | F1\_score (%) | Precision (%) | Recall (%) |
| Random Forest | **71.53** | **72.30** | **71.39** | **73.24** |

**Neural Network (Multi-layer Perceptron) Classifier:**

****

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy (%) | F1\_score (%) | Precision (%) | Recall (%) |
| Neural Network | **75.06** | **76.45** | **73.53** | **79.62** |

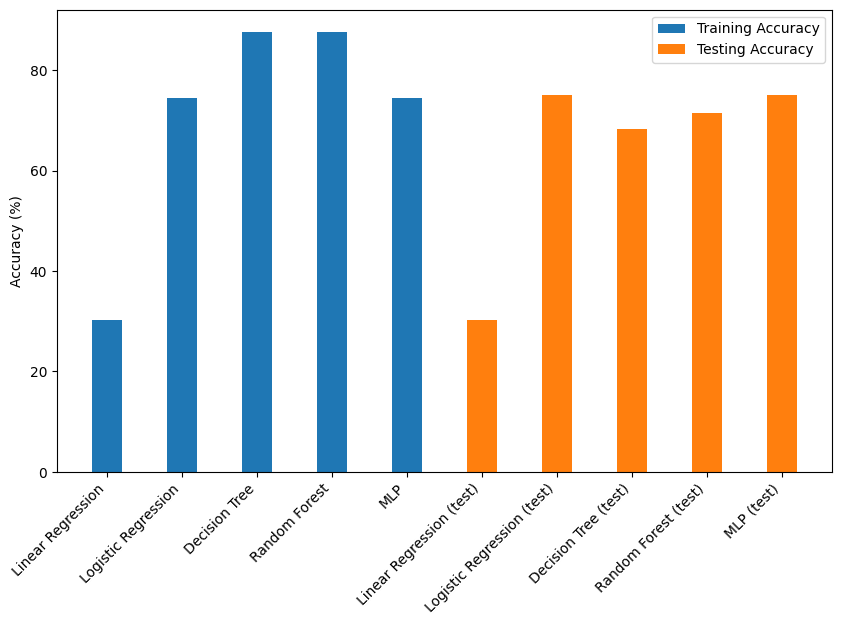
Results:

After Applying Linear Regression, Logistic Regression, Decision Tree Regressor, Random Forest Regressor and Neural Network (Multi-layer Perceptron) Classifier,

we have achieved predicted results of the training dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Without pre-processing** | | **With pre-processing** | |
| **Model** | Training accuracy | Testing accuracy | Training accuracy | Testing accuracy |
| **Linear** | 30.63 | 30.85 | 30.24 | 30.26 |
| **Logistic** | 74.69 | 74.83 | 74.56 | 74.67 |
| **Decision Tree** | 97.37 | 65.98 | 87.70 | 68.29 |
| **Random Forest** | 97.37 | 73.30 | 87.70 | 71.53 |
| **Neural Network** | 74.61 | 74.36 | 75.03 | 75.06 |

The training and testing accuracy percentages of our dataset following preprocessing are shown in the bar chart in the image below:



Conclusion and Future work:

In this Diabetes Determination Through Activities and Physical Traits machine learning project, we came up with a new way to figure out how likely someone is to have diabetes by looking at what they do and how they look. Through careful study of the data, we have shown that these factors may be good ways to tell if someone has diabetes.

Based on the traits we gave it, the machine learning model we made seems to be able to correctly predict a person's risk of getting diabetes. By putting this model into an easy-to-use interface, both healthcare workers and people can make decisions ahead of time and take targeted measures to avoid illness.

This project helps the field of diabetes detection and control by using machine learning techniques to improve early diagnosis and personalized healthcare strategies. When living factors and physical traits are taken into account, diabetes risk assessment becomes more accurate and efficient. This lets people make better decisions and improves patient results.

Future work for the Diabetes Determination Through Activities and Physical Traits machine learning project includes expanding the dataset to make it larger and more diverse, ranging adding time series data to capture temporal aspects, adding more biomarkers to improve predictive accuracy, looking into real-time monitoring using wearable devices, and evaluating the effectiveness of intervention strategies based on the model's predictions. These improvements could make diabetes risk screening and control even more accurate and effective, which could lead to better health care around the world.

In all, this project shows how important it is to use machine learning techniques to use actions and physical traits to predict diabetes. This will lead to better ways to avoid and treat diabetes in the long run.

References

1. *Linear Regression in Machine learning*. (n.d.). Javatpoint. <https://www.javatpoint.com/linear-regression-in-machine-learning>
2. *Decision Tree Regression*. (n.d.). The Click Reader. <https://www.theclickreader.com/decision-tree-regression/>
3. *Random Forest Regression in Python*. (2023, April 18). GeeksforGeeks. <https://www.geeksforgeeks.org/random-forest-regression-in-python/>
4. Fregoso-Aparicio, L., Noguez, J., Montesinos, L. *et al.* Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetol Metab Syndr* 13, 148 (2021). https://doi.org/10.1186/s13098-021-00767-9

<https://dmsjournal.biomedcentral.com/articles/10.1186/s13098-021-00767-9>

1. *What is a Neural Network? - Artificial Neural Network Explained - AWS*. (n.d.). Amazon AWS. <https://aws.amazon.com/what-is/neural-network/>
2. *Logistic Regression in Machine Learning*. (n.d.). Javatpoint. <https://www.javatpoint.com/logistic-regression-in-machine-learning>
3. Guide, S. (2022, January 4). *Diabetes Prediction Using Machine Learning*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/>
4. Dhrubo, T. H., Haque, F. B., Taposhi, U. J., and Najifa (2022), [Diabetes, Hypertension and Stroke Prediction](https://www.kaggle.com/datasets/prosperchuks/health-dataset), *Kaggle*. Available at: <https://www.kaggle.com/code/tahsinulhaquedhrubo/cse422-project#Diabetes>