**Predicting Diabetes in the United States**

Lorena Estrada

230149445

***Abstract*— Diabetes occurs in 37.7 million people in the United States. This disease occurs when blood glucose is abnormally high. The disease, over time, can cause heart disease, nerve damage, eye problems, and kidney disease. [1] About 28.7 million of the population get diagnosed, but 8.5 million adults above the age of 18 go undiagnosed. [2]**

**Machine learning algorithms that are applied are logistic regression and decision trees. Methods were used to determine type 1, type 2, or no diabetes based on diabetes data containing different gender and age ranges**.

# *Keywords—Diabetes, Logistic Regression, Decision Tree*

**I. INTRODUCTION**

The following test investigations involve using prediction of diabetes patients to determine if they had diabetes and, if so, what type of diabetes. Applied methods were Logistic Regression and Decision Tree algorithms. These two methods would be applied to patients of different genders, ages, hypertension, heart disease, smoking history, BMI, Hba1c levels, and blood glucose levels resulting in diabetes type. This information can help identify diabetic patients using current medical history. The dataset on Kaggle [3] involved nine factors and over a hundred thousand patients including medical history.

Diabetes involves various health issues to determine whether you fall under the category of type 1 or type 2. Nevertheless, both involve the body not producing insulin or not producing enough insulin to endure daily activities. The non-production of insulin affects the body by causing more health issues, such as kidney and heart disease.

Type 1 diabetes is a reaction from the autoimmune system that attacks the body and stops it from producing insulin. Most people who are diagnosed with diabetes have type 1. Type 2 diabetes involves the body not accepting insulin, resulting in high blood glucose levels. The diagnosis of type 2 is around adults. Both types can be controlled with an active lifestyle and healthy eating habits.

In this study, we will use different health questions revolving around diabetic concerns (e.g., blood glucose levels) to determine diabetes. We will do this analysis using machine learning methods to determine the answer, giving us a better understanding of how to diagnose diabetes in a more analytical approach.

**OBJECTIVES**

* Evaluate which factors weigh higher on a person's association with diabetes.
* Apply two data analytics techniques, Logistic Regression and Decision tree.
* Analyze data to understand factors involved and clean data if necessary.
* Demonstrate the data clearly for better understanding and visual representation. - Assess the models made and check for precision

**II. LITERATURE REVIEW**

Diabetes is a significant topic in the world due to the fact that millions of countries have doubled their numbers of diabetic patients. Applying machine learning to public healthcare has led us to learn more about how certain diseases or health conditions correlate with certain things. For example, the application of machine learning to diabetes can help us understand how a person's glucose levels can affect diabetic prediction. In the following journals, the demonstration of how machine learning has been applied and how diabetes has become an increasing topic due to the rising number in all sectors of the world will be analyzed.

The first journal thoroughly reviews hospital statistics from the Medical Service Department of Bangkok, Thailand. Stating that from 2011 to 2019, there was a drastic increase in patients with diabetes, showing that many patients did not know they had diabetes at all. It is known that there is no cure for diabetes, yet there is information involving type 1 and type 2 that can lead us to see if they have things in common. There is little information on type 1 other than type 1, which is known to be found in children and young adults. One known risk is having a parent or siblings with type 1 diabetes. On the other hand, type 2 diabetes has more known risks, such as being overweight, having prediabetes, being older than forty-five, and presenting a family history of diabetes. Overall, the diagnosis of diabetes is pivotal for a patient's life. [4]

The application of machine learning algorithms could help the medical diagnosis of diabetes. The data presented goes over several machine learning models that could improve diagnosis. Decision trees, random forests, support vector machines, and K-nearest neighbors are presented as machine-learning techniques that could be applied. The proposal of using hyperparameter tuning and interaction teams is evaluated due to providing more efficient results than most machine learning techniques. Concluding that the random forest algorithm is shown to be the best.[4]

The second journal discusses diabetes mellitus. A type of diabetes that is not type 1 or type 2 but a life threatening disease that can affect the heart, kidney, and liver. Biotechnology continues to help diagnose these conditions at a remarkable rate. Approaches taken using machine learning techniques were Random Forest, Multilayer Perceptron, and Logistic Regression. These approaches have been applied to the PIMA Indian dataset for diabetes classification to be identified in the early stages of diabetic diagnosis. The predictive analysis applied were Long Short Term Memory, Moving Averages, and Linear Regression.

Demonstrating that in the experimental test, Multilayer Perceptron achieves 86.06% accuracy in diabetes classification, and Long Short Term Memory improves accuracy to 87.26 in the prediction of diabetes. [5]

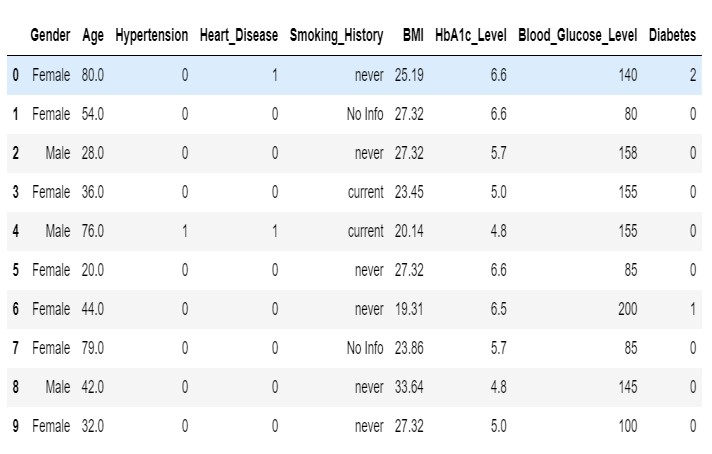
The last journal goes into the condition of diabetes and how blood glucose levels can elevate it. Elevation of diabetes can lead to stroke, kidney failure, heart failure, amputation, and blindness. Demonstrating that diabetes affects the world and the future of world health. Making it a significant concern for the future of healthcare and the need for efficient diagnosis. Type 1, Type 2, and gestational diabetes (occurs during pregnancy) are the most common known diabetic conditions. Testing methods using machine learning could lead to better prediction and diagnosis of diabetes. Machine learning techniques applied K-nearest neighbor, logistic regression, random forest, support vector machine, and decision tree algorithms. Revealing that the methods with higher accuracy are the best for predicting diabetes. The overall objective is to examine the best machine learning techniques to enhance the prediction of the onset of diabetes. [6]

**III. DATA MANAGEMENT**

# Data Source and Description

The diabetes prediction dataset can be found on Kaggle [3]. It includes patients' medical and demographic history to test for the positive or negative results of diabetes. The data consists of nine columns and one hundred thousand rows. Healthcare professionals can use the dataset to predict diabetes in patients based on their history, helping medical professionals and patients target diabetic diagnosis early on. Researchers can explore the relationship between factors to better understand the development of diabetes.

The dataset can be downloaded and shown as an Excel document. The document consists of gender including female, male, and other, age ranging from one year old to eighty years old, hypertension demonstrated as one if hypertension is presented and zero if hypertension is not present, heart disease given as one for presence of heart disease and zero as non-presence of heart disease, smoking history consisting of current, ever, former, never, no info, not current, body mass index (BMI) presented as integer format, hemoglobin A1C (HbA1c) Levels presented in integer format, and ending with one, two or zero to identify the positivity or negativity of diabetes presented.

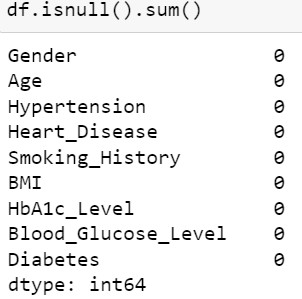


***Table 1: Original dataset sample***

# Data pre-processing / Data cleaning

Cleaning of data is a fundamental part of the test. This is because raw data is inconsistent, and cleaning the data leads to organized data analysis. Null data can lead to errors and misinformation. Consistent data leads us to read correct information. These are the reasons why data preprocessing is essential for the accuracy of testing. [8]

Data cleaning applied to the Diabetes prediction dataset was verified in the rows and columns, which consisted of 100,000 rows and 9 columns—going ahead and checking for duplicates and dropping any of those duplicates. After applying this method, we verified if any duplicates were dropped, and our columns went down to 96,146, clearly showing that some duplicates were dropped. Further cleaning applied was checking for null values, returning zero of all of our nine columns presenting null values. Lastly, we are checking for zero values in our dataset. Blood Glucose Levels, BMI, and HbA1c Levels returned zero numbers of zero presented. Hypertension returned 8,8685, and Heart Disease returned 92223 of zero values returned. For better analysis, zero values were replaced with the mean of these rows and verified once again if any zero values were presented.



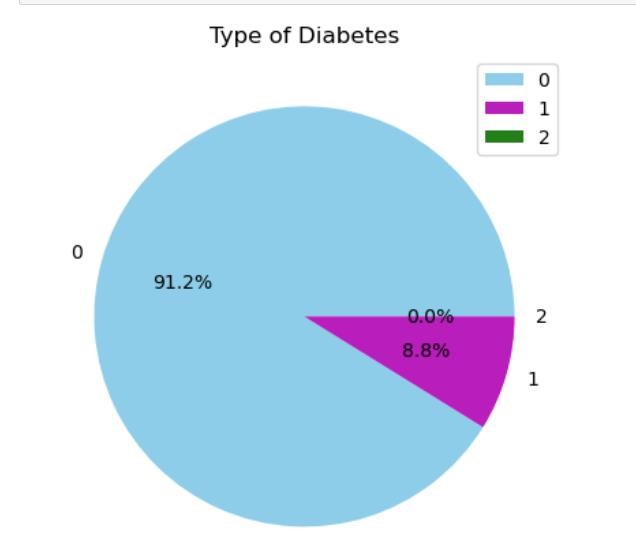
***Table 2 : Checking Null Values***

# Data Exploration

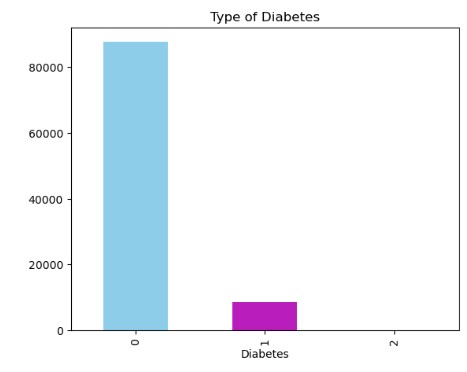
It is important to explore our data to understand better how the testing for Diabetes prediction is going to work. Data exploration reveals information that needs to be seen or understood. This process enables predictive modeling and forecasting [7]. Helping support significant insights for decision-making

As seen in Figure 3, the majority of patients showed no signs of diabetes. (91.2%) percent of people have no diabetes, number one indicates the presence of type 1 diabetes (8.8%), and number two indicates the presence of diabetes but type 2 (0.0%).

Figure 4 uses a different plot to better understand the diabetes results found in the data. It might be simpler to identify outcomes from this perspective. Approximately 80,000 patients had no diabetes, and less than 20,000 individuals reported having diabetes. Finally, there was nearly no presence of type 2 diabetes.



***Figure 3 :***



***Figure 4 :***

# Feature Selection

To improve accuracy, feature selection is done after data cleaning. To generate more accurate predictions, the correct data must be selected. This involves removing nonrelevant data to improve the accuracy.

The correlation heatmap and correlation table suggest that we could remove Smoking History from our data. Figures 4 and 5 give a better understanding of why we should do this.

***Figure 5: Correlation Analysis***

The correlation heatmap shows how to identify the correlation patterns between the variables in our dataset. Correlation can be identified by color scheme. The lighter the color, the more positive the correlation, and the darker the color, correlation is negative.

A correlation table can be used to make a more precise analysis. Variables in the dataset can be correlated by using 1, -1, or 0 to indicate a strong positive correlation, a strong negative correlation, or no correlation. Based on the table, it can be interpreted that Diabetes and Blood Glucose levels have a strong positive correlation, Age and BMI have a slightly positive correlation, and Smoking History and Gender have a weak negative correlation.

***Figure 6 : Correlation Table***

***Figure 7: Ex***

***tra***

***TressCla***

***ssif***

***i***

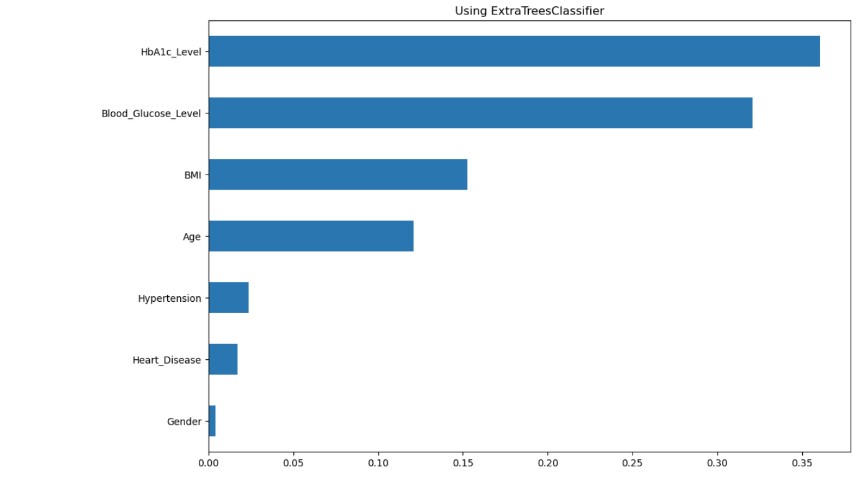
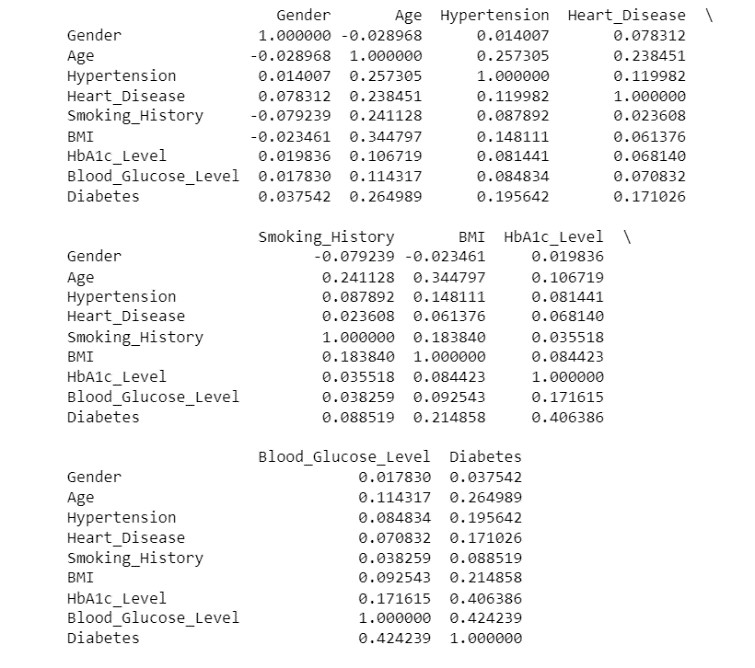
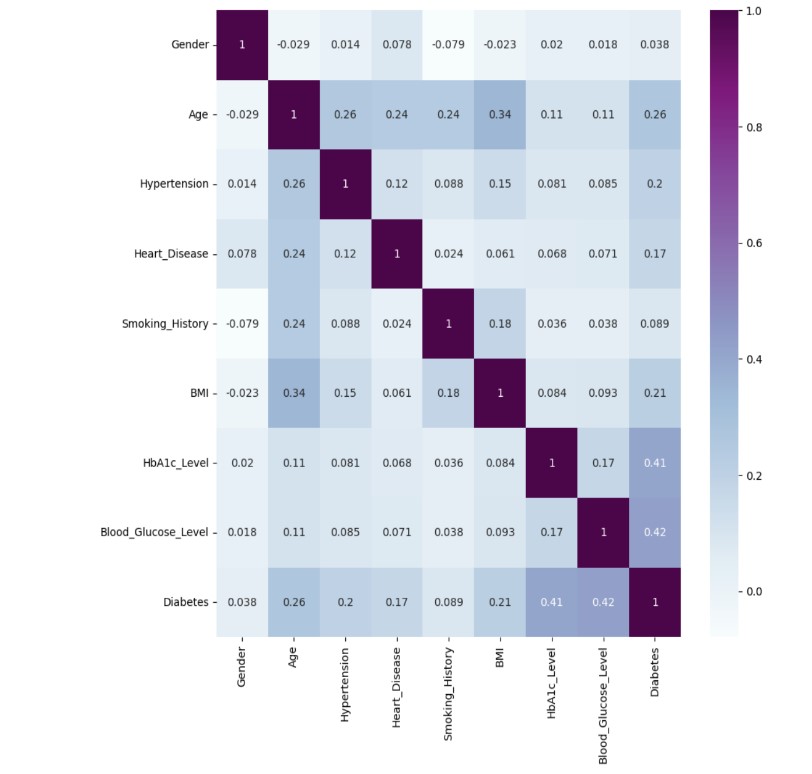
***er***

It appears that all nine columns can be reduced to eight

columns after using these techniques. Gender, Heart

disease, Age, BMI, Blood Glucose Levels, and HbA1c

Levels are the remaining columns.



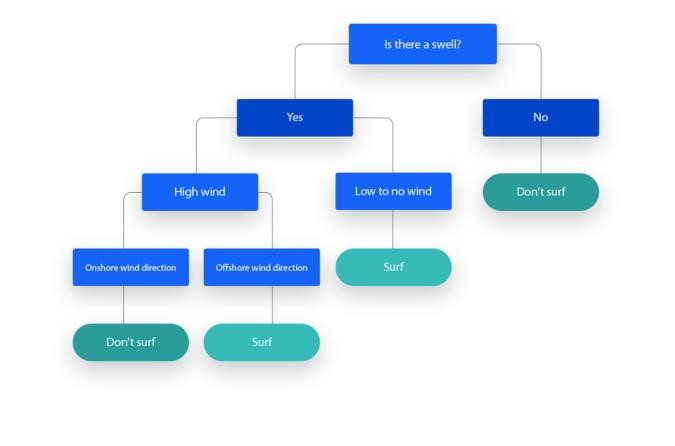
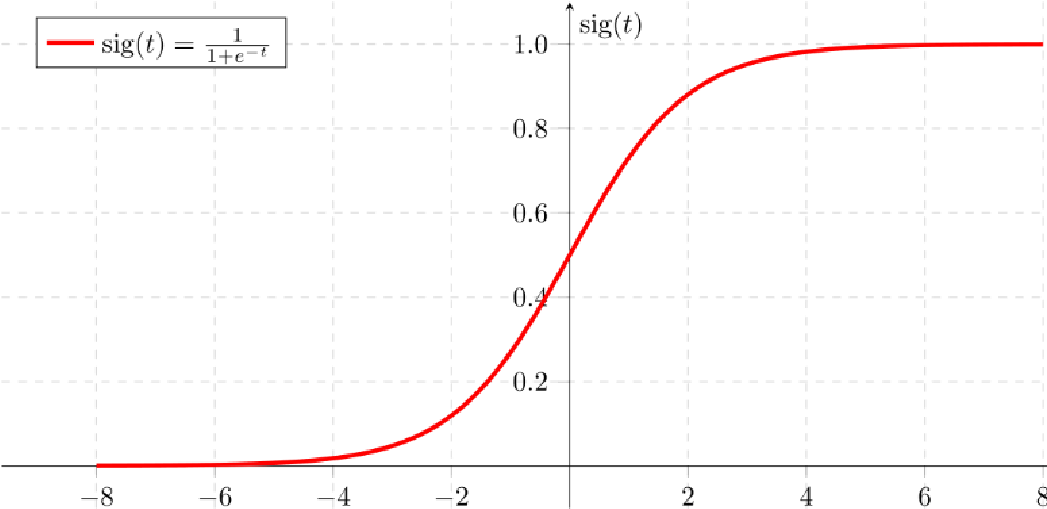
As a result of developing a predictive model, reducing our data through machine learning techniques generates information about the strength of each target variable. Various approaches can be taken; the Extra Trees

Classifier was utilized in this case.

Extra Trees Classifier is a Scikit-Learn method based on decision trees. It functions to reduce overfitting and overlearning from the data by randomly selecting specific options and data subsets. Extra Tress aims to decrease overfitting and increase predictive accuracy by adding randomness. [11]

# IV. DATA ANALYTICS METHODS Logistic Regression

Logistic regression is a classification method that can measure the probability of an event occurring based on a given dataset. It is frequently used for classification and predictive analytics. It runs by predicting the likelihood of specific outcomes occurring. This predictive analysis is one of the most commonly used in data analysis. Most Logistic regression models are based on a categorization response and can be Binary Logistic regression, Multinomial Logistic regression, and Ordinal Logistic regression. It can only be applied between categories; other applications would not work. The Scikit-Learn method can be used in logistic regression machine learning. This classification method can be used in disease prediction, fraud detection, and performance predictions. [10]



## Justification

Logistic regression was applied to the diabetes prediction dataset because it is a statistical method most commonly used in the medical field. The dataset would predict a yes or no answer, making it helpful in applying logistic regression. This method works by using a variable and modeling the possibility of an event happening, in this case, if a patient had or did not have diabetes. Utilizing logistic regression helps determine risk factors that led up to diabetes. It can be applied using variables like BMI, age, and blood glucose levels. Heart disease and HbA1c Levels are used to determine if a person has the condition. By using logistic regression, it can be determined which major risk factors can lead a

patient to have diabetes.

# Decision Tree

Decision Tree is a machine learning algorithm that is used in classification and regression situations. Visualization can be a hierarchical tree structure consisting of a root node, branches, internal nodes, and leaf nodes. The representation of this method is easy to follow for decision-making, and it is simple to comprehend why it is that outcome. The decision tree works by assigning a class label, dividing the process, and repeating this downwards. Smaller decision trees are easier to maintain; the more significant the complexity of the tree, the harder it is to apply this method. The use of decision trees has its limitations. It is prone to overfitting and high variation and can be extensive to work on depending on the data. Advantages include the fact that it is easier to follow the outcome, it can process various data types, handle missing values, and is flexible to use. [9]

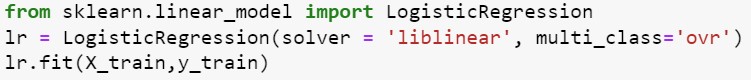
## Justification

When utilizing decision trees on the diabetes prediction datasets, a clear visual representation is offered; each split links to a variable, making it clear to understand which variables affect the prediction. Diabetes prediction involves connections with various health risk factor variables, like BMI, age, blood glucose levels, Heart disease, and HbA1c Levels. A decision tree categorizes variables depending on how significant the role they play in predicting diabetes, making it visually better to identify which variables play a higher role in prediction. It is also known to work with missing data and still produce a good prediction. This is essential in the medical field since data is sometimes up to date or unavailable. Lastly, the decision tree has been applied to

**V. RESULTS**

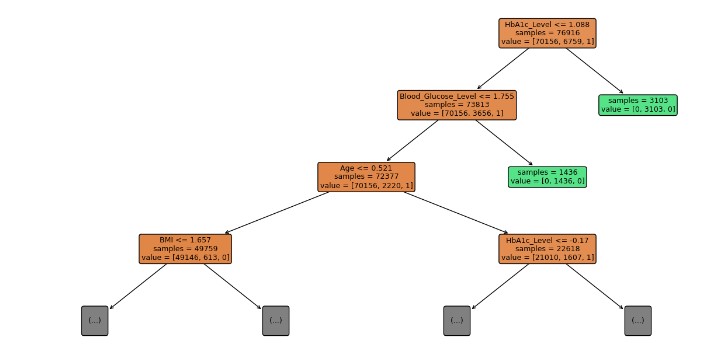
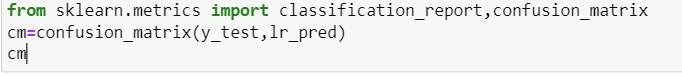
# Logistic Regression

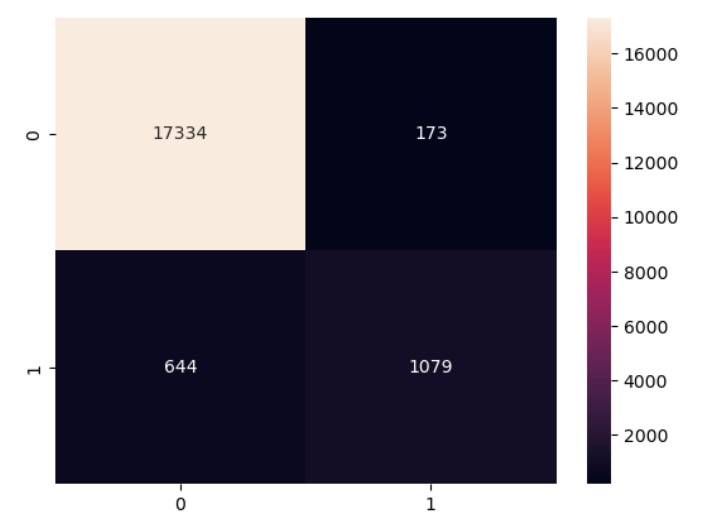
Logistic regression is used to test the data in Python, and Scikit-Learn is used as a method. Applying from sklearn.linear model import LogisticRegression. Regression analysis is used to forecast the probability of diabetes data set. An object for logistic regression is applied, shown as (lr). Liblinear is used for small datasets and is used in this case. Multi\_class=’ovr’ is utlizated in multiclass classification. “Ovr” is used in a binary classification. Lr.fit(X\_train, y\_train) is fitted to follow the two parameters. After running these lines of code, the logistic regression model lr will be trained using the provided training data (X\_train and y\_train). X\_train refers to the independent variable, and y\_train refers to the dependent variable that matches with X\_Train. After running the code, the logistic regression algorithm will be trained.

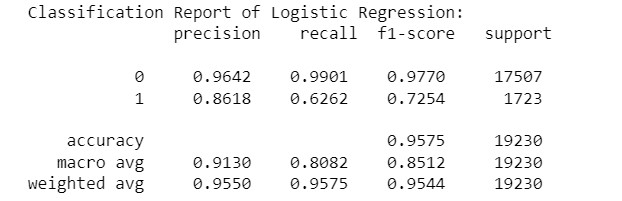


# Confusion Matrix

Applying the dataset, the confusion matrix will calculate the predictions generated by the logistic regression. From sklearn.metrics import classification\_report, confusion\_matrix, which will import Scikit-learn, classification report, and confusion matrix functions. The confusion matrix is utilized to show how a classification model is applied to test data. The confusion matrix provides visual analysis and highlights the amount of true positives, true negatives, false positives, and false negatives, showing the model's performance.

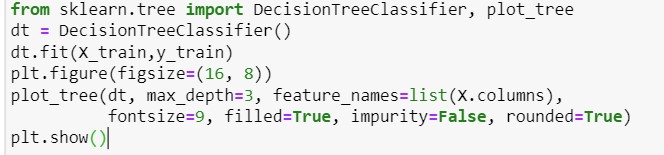






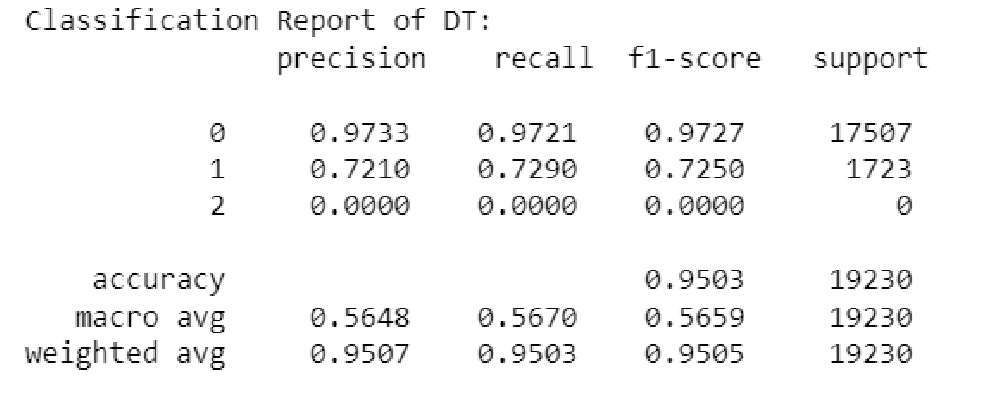
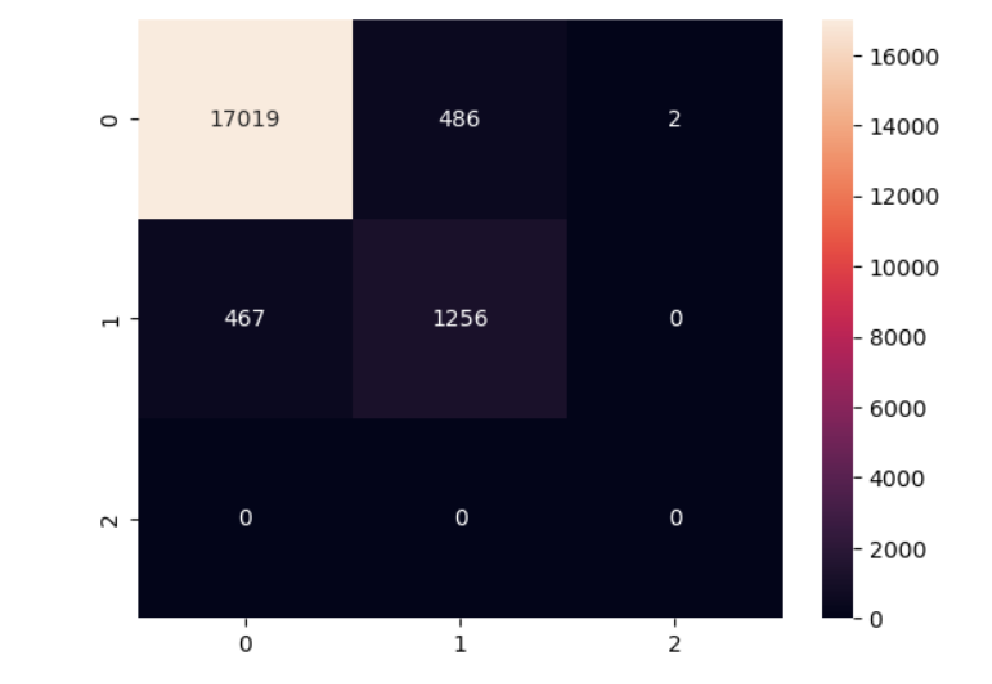
# Decision Tree

To run the Descion tree, the first Scikit-learn method must be imported. This can be done by applying the sklearn.tree import DecisionTreeClassifier. The Decision Tree is fitted to the training set of data (X\_train,y\_train). This will train the data and identify the connections between the (y\_train) and (X\_train).



# Confusion Matrix

The confusion matrix is applied to the predictions of the Decision Tree from sklearn.metrics import classification\_report, confusion\_matrix, this imports Scikit-Learn methods, classification report, and confusion matrix functions. Visualization will show the confusion matrix and how the classification model works when used on the data.



**VI. CONCLUSION**

# Achievements / Challenges

In conclusion, all objectives were achieved. Two logistic regressions and a decision tree were applied successfully. Necessary cleaning was done during preprocessing and data cleaning. The demonstration of data was clearly presented in various forms. Models were analyzed and checked for precision. The classification report for logistic regression does well for accuracy but somewhat needs to improve and produce imbalance. The decision tree does well for class 0, somewhat for class 1, and poorly for class 2. Overall, we achieved what we wanted to do.

The challenges presented were a large dataset, to begin with; only a few datasets with small data were found. However, for this instance was helpful for testing Logistic Regression and Decision Trees.

# VII. REFRENCES

1. “What is Diabetes.”

Center for Disease Control and Prevention– 2023

URL: <https://www.cdc.gov/diabetes/basics/diabetes>

1. “Diabetes Statistics”

National Institute of Diabetes and Digestive and Kidney diseases-2023

URL: [Diabetes Statistics - NIDDK (nih.gov)](https://www.niddk.nih.gov/health-information/health-statistics/diabetes-statistics)

1. “Diabetes prediction dataset”

Mohammed Mustafa – 2023

URL: [Diabetes prediction dataset (kaggle.com)](https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset)

1. “Diabetes Classification Using Machine Learning

Techniques.”

Methaporn Phongying and Sasiprapa Hiriote– 2023

URL: [Computation | Free Full-Text | Diabetes Classification](https://www.mdpi.com/2079-3197/11/5/96#:~:text=In%20many%20research%20studies%2C%20well-known%20machine%20learning%20techniques%2C,widely%20used%20in%20diabetes%20classification%20%5B%206%20%5D.)

[Using Machine Learning Techniques (mdpi.com)](https://www.mdpi.com/2079-3197/11/5/96#:~:text=In%20many%20research%20studies%2C%20well-known%20machine%20learning%20techniques%2C,widely%20used%20in%20diabetes%20classification%20%5B%206%20%5D.)

1. “An Effective Machine Learning Approach for Diabetes

Prediction”

Lakshmi,Thokala,Kumar,Reddy,Reddy-2023

URL: [An Effective Machine Learning Approach for](https://ijsrcseit.com/CSEIT2390442)

[Diabetes Prediction (ijsrcseit.com)](https://ijsrcseit.com/CSEIT2390442)

1. “Machine Learning Based Diabetes Classification And

Prediction for Healthcare Applications”

Butt, Letchmunan, Ali, Hassan, Baqui, Sherazi– 2021

URL: [Machine Learning Based Diabetes Classification and](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8500744/)

[Prediction for Healthcare Applications - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8500744/)

1. “What is Data Exploration?”

Fasih Khan – 2023

URL: [Data Exploration: Importance + 10-Step Guide |](https://www.astera.com/type/blog/data-exploration/)

[Astera](https://www.astera.com/type/blog/data-exploration/)

1. “Data Preprocessing”

Ryan Miller – 2019

URL: [Data Preprocessing: what is it and why is important -](https://ceoworld.biz/2019/12/13/data-preprocessing-what-is-it-and-why-is-important/)

[CEOWORLD magazine](https://ceoworld.biz/2019/12/13/data-preprocessing-what-is-it-and-why-is-important/)

1. “What is a decision tree”

IBM –

URL: [What is a Decision Tree? | IBM](https://www.ibm.com/topics/decision-trees)

1. “what is logistic regression”

IBM–

URL: [Data Preprocessing: what is it and why is important -](https://ceoworld.biz/2019/12/13/data-preprocessing-what-is-it-and-why-is-important/)

[CEOWORLD magazine](https://ceoworld.biz/2019/12/13/data-preprocessing-what-is-it-and-why-is-important/)

1. “ExtraTreesClassifier”

Naman Bhandari – 2018

URL: [ExtraTreesClassifier. How does ExtraTreesClassifier reduce… | by Naman Bhandari | Medium](https://medium.com/@namanbhandari/extratreesclassifier-8e7fc0502c7)