Diabetes Detection with Machine Learning

I. Dataset and Data Analysis

A publicly available dataset from Kaggle (https://www.kaggle.com/datasets/ankitbatra1210/diabetes-dataset) was used to train the models.

Out of 13 types of diabetes, 4 major types were chosen, and the following mapping was applied to facilitate working with models that took only numberic values as input :

'Prediabetic': 0,

'Type 1 Diabetes': 1,

'Type 2 Diabetes': 2,

'Type 3c Diabetes ': 3

These types can be brielfy described.

i. Prediabetic

Prediabetes is a warning of Type 2 diabetes. Blood sugar levels are elevated, but not enough to be considered Type 2 diabetes. Lifestyle changes can be made to manage prediabetes, like getting more physical activity and adjusting eating patterns and habits.

Prediabetes usually have blood sugar levels ranging from 100 to 125 mg/dL.

According to the American Diabetes Association, for people 45 years old with prediabetes, the 10-year risk of developing Type 2 diabetes is 9% to 14%.

Family history of Type 2 diabetes, a BMI greater than 25, being physically active fewer than three times a week are general risk factors for prediabetes. Further, people who are 45 years of age or older are much likelier to fall into this category.

Some common symptoms experienced in this stage are : Increased thirst, Frequent urination, Increased hunger, Fatigue, Blurred vision, Numbness or tingling in the feet or hands, Frequent infections, Slow-healing sores and Unintended weight loss.

ii. Type 1 Diabetes

Type 1 diabetes is a chronic autoimmune disease that prevents the pancreas from making insulin. It requires daily management with insulin injections and blood sugar monitoring. Both children and adults can be diagnosed with Type 1 diabetes.

Lack of enough insulin causes a build uo of sugar in the blood, causing hyperglycemia (high blood sugar), resulting in the bosy not being able to use the food eaten for energy. This leads to serious health problems or even death if it’s not treated. People with Type 1 diabetes need synthetic insulin every day in order to live and be healthy.

Type 1 diabetes can appear at any age, but it appears at two noticeable peaks. The first peak occurs in children between 4 and 7 years old. The second is in children between 10 and 14 years old. Family history is another factor that is a risk factor.

Common symptoms include : Constant thirst, increased frequency of urination, constant hunger, significant weight loss, irritability, tiredness and blurry vision.

iii. Type 2 Diabetes

Type 2 diabetes happens when the body can’t use insulin properly. Without treatment, Type 2 diabetes can cause various health problems, like heart disease, kidney disease and stroke. It is a chronic condition that happens due to persistently high blood sugar levels (hyperglycemia). Patients have blood sugar levels of typically 126 mg/dL or higher.

Type 2 diabetes happens because the pancreas doesn’t make enough insulin, the body doesn’t use insulin properly, or both. Researchers estimate that this type of diabetes affects about 6.3% of the world’s population. It most commonly affects adults over 45, but people younger than 45 can have it as well, including children.

Chronic stress and a lack of quality sleep, genetics, excess body fat, physical inactivity, eating highly processed, high-carbohydrate foods and saturated fats frequently and long-term corticosteroid use can all cause type 2 diabetes.

iv. Type 3c Diabetes

Type 3c diabetes develops when the pancreas experiences damage that affects its ability to produce insulin. Conditions like chronic pancreatitis and cystic fibrosis can lead to pancreas damage that causes diabetes. Having the pancreas removed (pancreatectomy) also results in Type 3c diabetes.

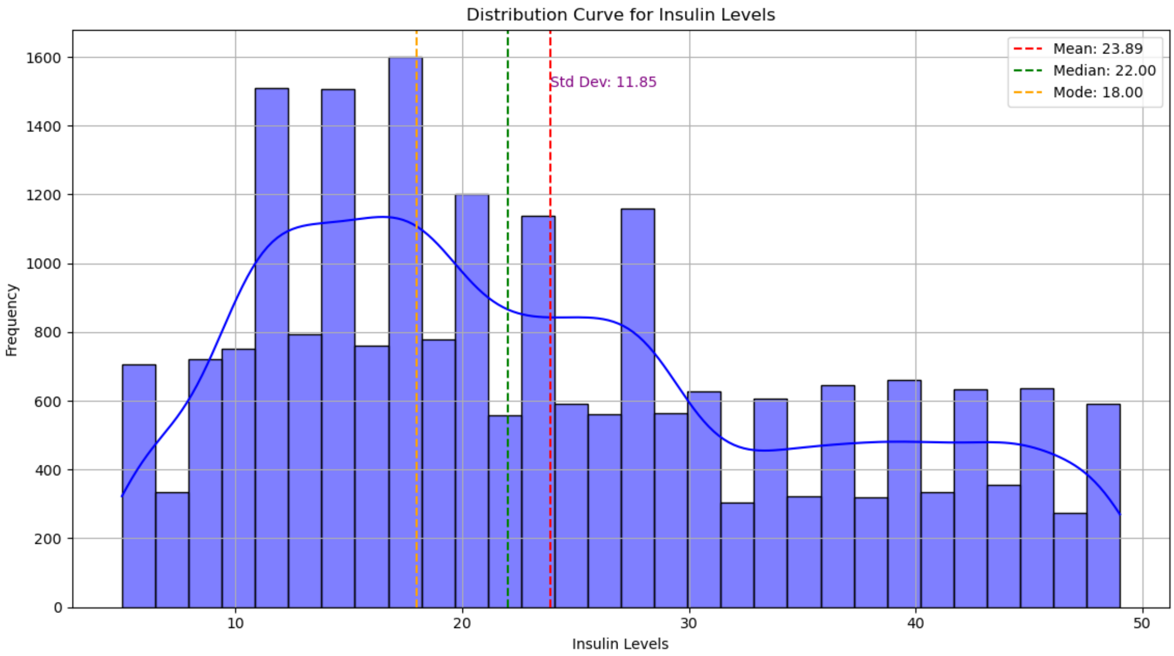
Pancreas damage that leads to Type 3c diabetes often also affects the pancreas’s ability to produce the enzymes that help with digestion. This condition is called exocrine pancreatic insufficiency (EPI).

Pancreatitis, pancreatic cancer, hemochromatosis are the main causes of type 3c diabetes. Symptoms similar to Type 1 Diabetes are visible in this type of diabetes.

After this filtering, data from 21,539 patients were available.

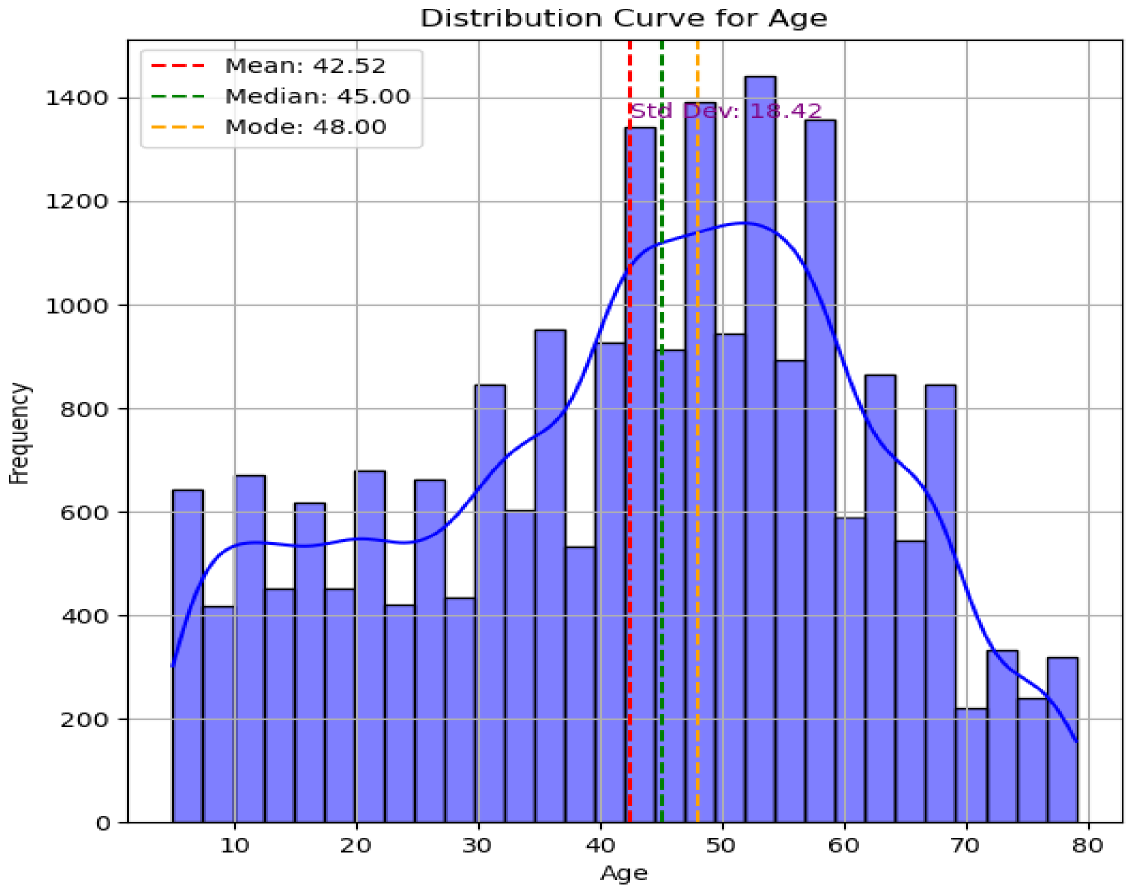
Further, out of the 33 available attributes, 9 were selected. These attributes have been described in brief. The Kernel Density Estimation curve, shown in blue, is a smoothed estimate of the probability density function of the data.

i. Insulin Levels



A study by Johnson, Duick, Chui and Aldasouqi [1] showed that at a fasting insulin level > 9.0 microIU/mL, prediabetes would be correctly identified in 80% of affected patients. Similarly, the lack of proper production of insulin is an indicator of Type 1 Diabetes.

ii. Age



Since there is an increasingly high risk of pancreatic cancer after the age of 50 years unrelated to the genotype [2], this age range is much likelier to develop Type 3c diabetes.

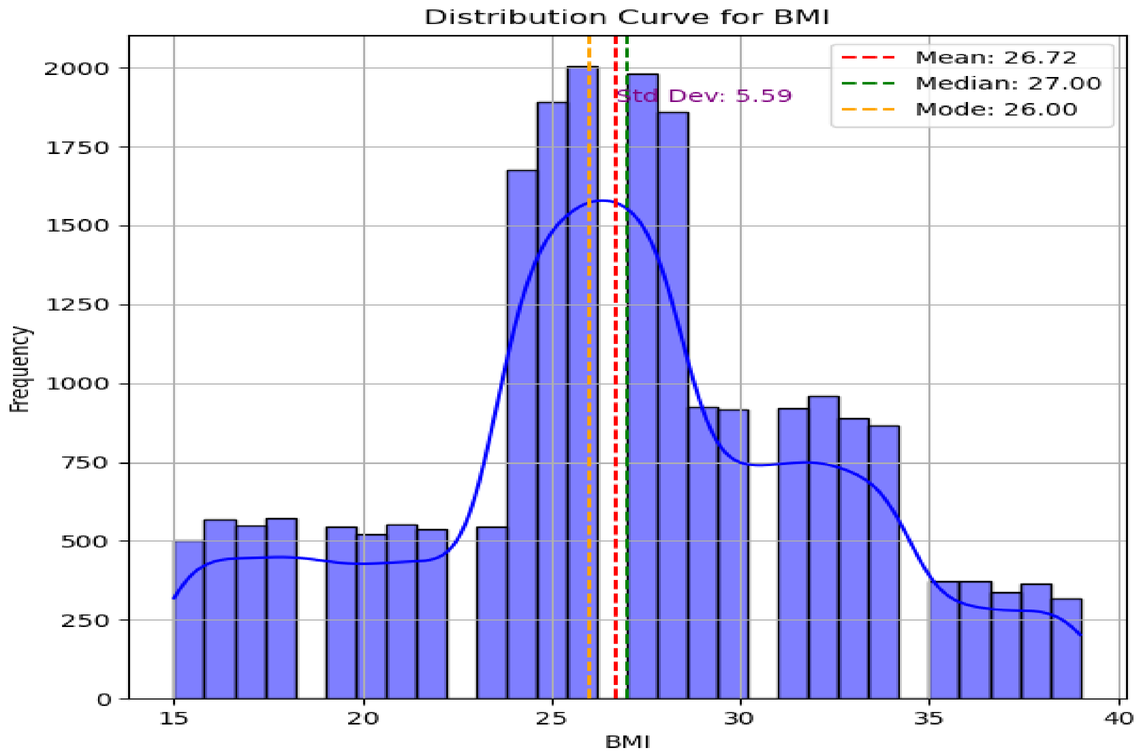
Type 1 diabetes mellitus can occur at any age, with a peak in incidence around puberty. Classification between T1D and type 2 diabetes becomes more challenging with increasing age of onset of T1D over time develops in genetically predisposed individuals. [3]

A study by Zoungas et. al showed that for type 2 diabetes, the mean age (±SD) was 65.8 ± 6.4 years, age at diagnosis was 57.8 ± 8.7 years and diabetes duration was 7.9 ± 6.4 years.[4]

iii. BMI

A study by Menon, Gill, and Hoey [5] confirms the abrupt onset of type 1 diabetes, the absence of a family history, and the importance of the classical symptoms of polyuria, polydipsia, and weight loss in the majority of cases.

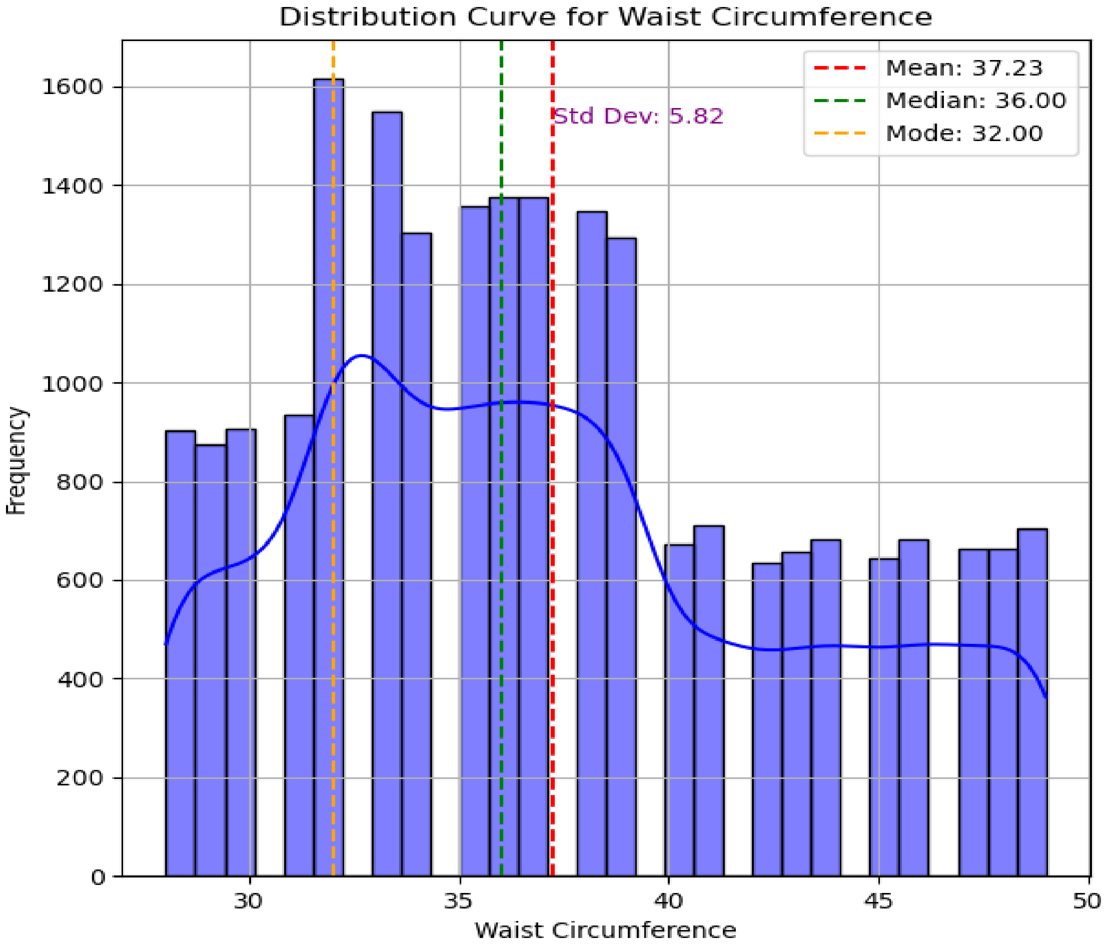
Prior to the release of the findings of the DCCT (Diabetes Control and Complications Trial) and the concomitant increase in the use of intensive insulin regimens, youth with T1D tended to be lean, mostly as a result of a loss of calories in the urine. However, corresponding to the increase of overweight and obesity in the general population, youth with T1D are also experiencing overweight and obesity (Liu et. al). [6]



iv. Waist Circumference

Has an almost direct correspondence with BMI.

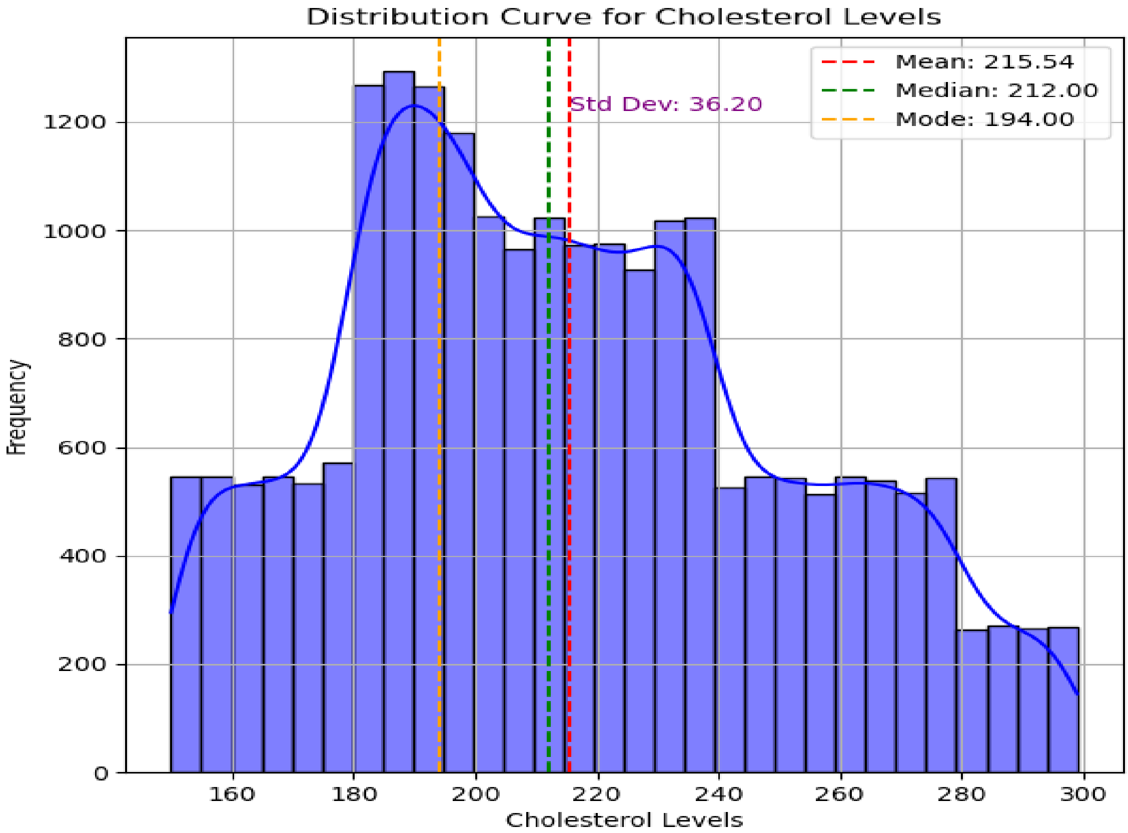
Janiszewski et al concluded in their study that waist circumference predicted diabetes, but not cardiovascular disease, beyond that explained by traditional cardiometabolic risk factors and BMI.



The findings lend critical support for the recommendation that waist circumference be a routine measure for identification of the high-risk, abdominally obese patient.

v. Cholesterol Levels

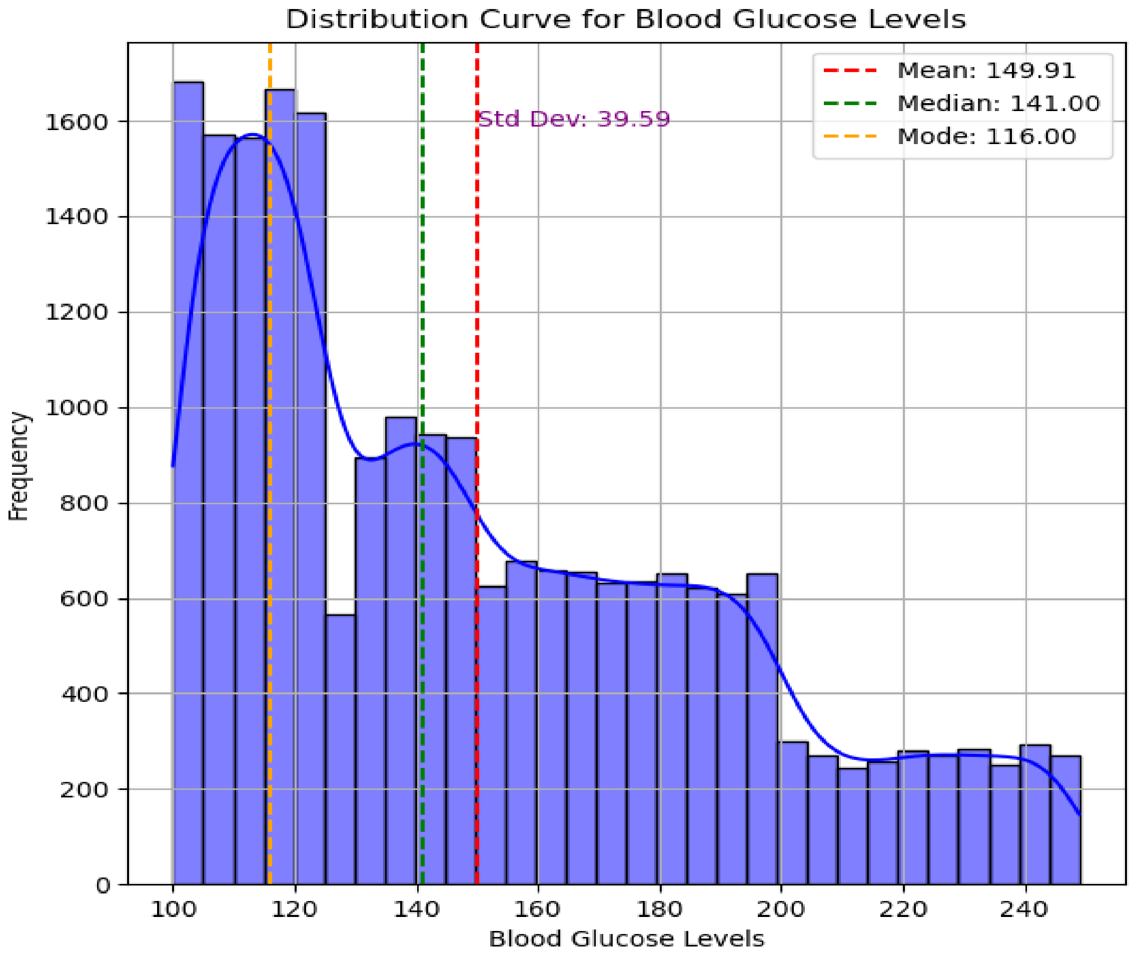
It has been shown that cholesterol absorption efficiency was increased by weight reduction, and the variables of glucose metabolism improved in obese diabetic subjects, suggesting that low cholesterol absorption is associated with insulin resistance and metabolic syndrome.[8]



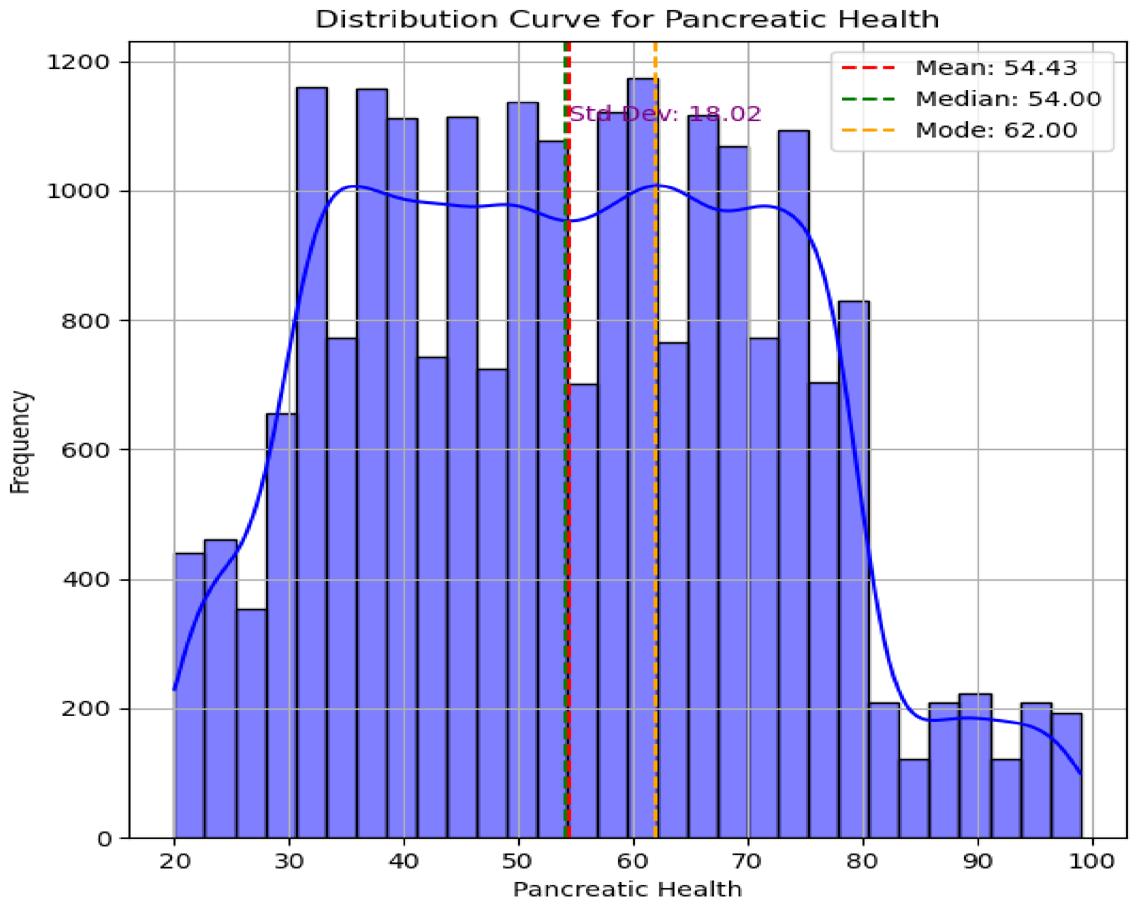
vi. Blood Glucose Levels

In literature there is a positive, but rather weak, association between the measures of blood glucose control and the risk of dying of patients with Type 2 DM. In the six larger studies (more than 100 deceased patients) that used a continuous categorization of glycaemia, the Risk ratio per unit varies from 1.03 to 1.12. [9]

<Emphasis to be added>



vii. Pancreatic Health



One frequently, though not consistently, reported feature of the pancreas in type 2 diabetes is its increased fat content, as determined by computed tomography (CT) and MRI. While Saisho and colleagues found pancreatic fat content increased with age, but not further with type 2 diabetes, multiple other studies documented additional lipid accumulation in individuals with type 2 diabetes and suggested that intra-organ fat might contribute to beta cell dysfunction.[10][11][12][13]

viii. Neurological Assesments

The number of neurological assesments might act as an indicator of impaired brained function.
If the blood sugar levels fall outside of the normal range, it can throw the command center off balance. In the same way diabetes can damage nerves in other parts of the body, it can damage nerves in the brain.
This can lead to problems with memory and learning, mood shifts, weight gain, and hormonal changes. Over time, it can also lead to other serious problems like Alzheimer's disease. Both high and low blood sugar levels can cause these harms.

ix. Glucose Tolerance Test

The glucose tolerance test measures the body's response to glucose.
This test can be used to screen for type 2 diabetes or prediabetes before you have symptoms of either condition. Or it can help find out whether diabetes is causing existing symptoms.
A healthy value is denoted by 0 (A healthy blood glucose level is lower than 140 mg/dL (7.8 mmol/L).), whereas values above the safe limit is denotes by a

II. Methodology

For model training, 60% of the data was utilised, and the remaining 40% was used for testing. Six different models were used for training and testing. It is very important to note that most of the other studies that use machine learning to predict the presence or absence of disease. However, our data consists of only the diseased population. This, combined with the significantly larger volume of data we trained our models on, means that our classification naturally shows a higher accuracy. Further improvement can be made by applying the datasets from other studies, to verify the validity of our model. A significant improvement can be achieved by obtaining datasets with non-diabetic parameters as well. Most of the refererred researches use the Pima Indians dataset ()

Results are very similar to those obtained in a study by Soni et. al[17].

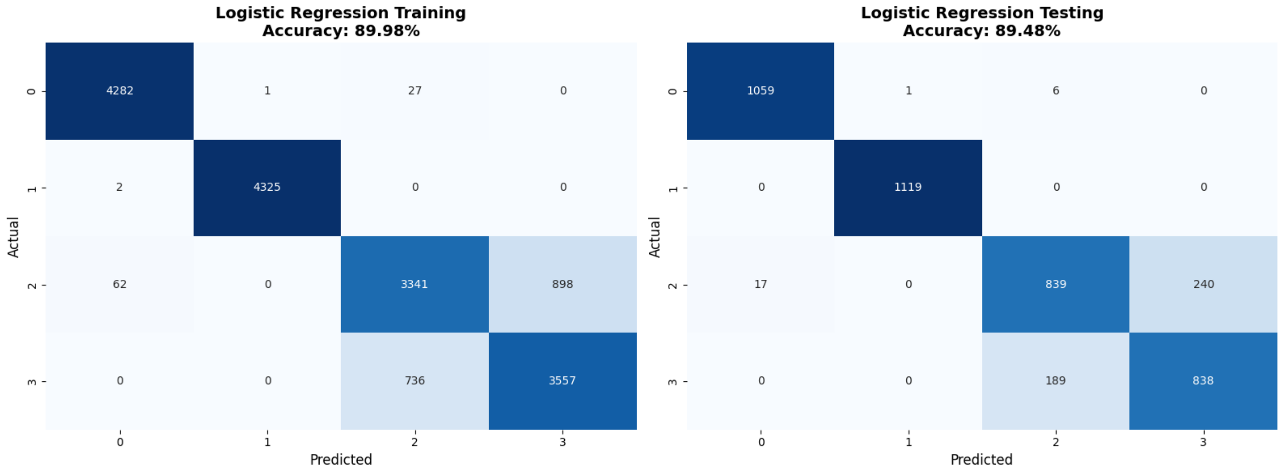
These parameters have been used to make comparisions :

The train accuracy: The accuracy of a model on examples it was constructed on.

The test accuracy : The accuracy of a model on examples it hasn't seen.

Confusion matrix: A tabulation of the predicted class against the actual class.

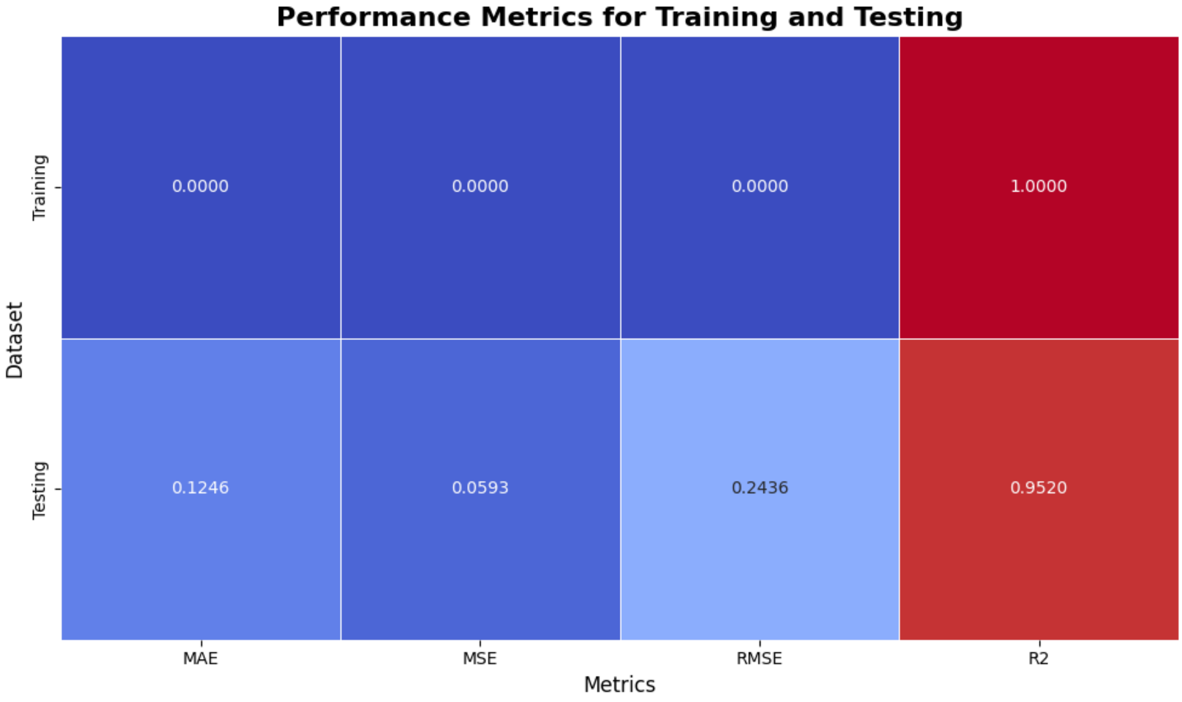
i. Logistic Regression



In a study by Rajendra and Latifi, the highest accuracy achieved was 78%. [14]

ii. K – Nearest Neighbours

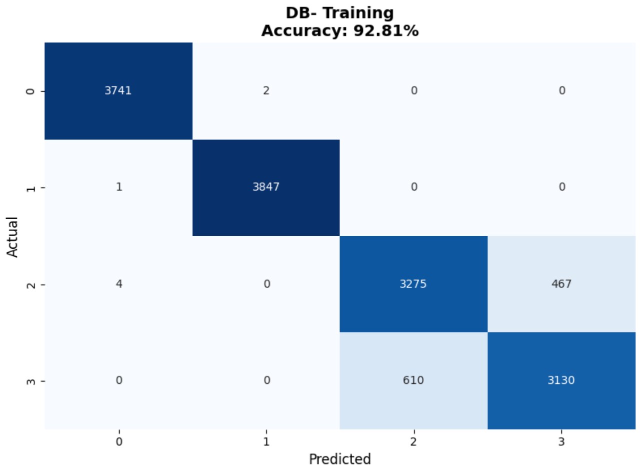
Some important definitions :
Mean Absolute Error (MAE) : MAE measures the average absolute difference between the actual and predicted values.
Mean Squared Error (MSE) : MSE calculates the average squared difference between the actual and predicted values.
Root Mean Squared Error (RMSE) : RMSE is the square root of the MSE and represents the standard deviation of prediction errors.
R-Squared (R^2): R^2 (coefficient of determination) measures the proportion of variance in the target variable that is explained by the model.

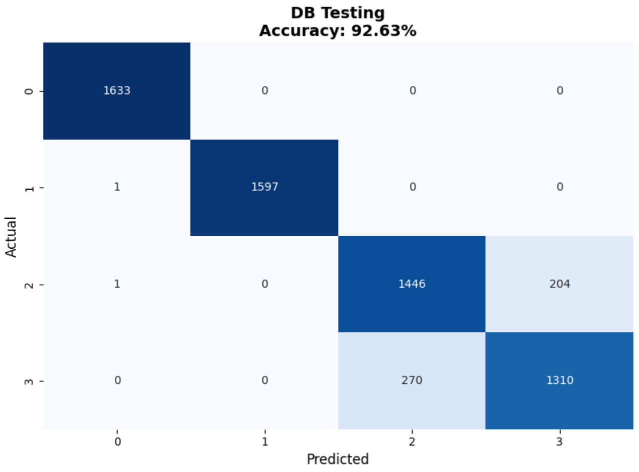


The important metric is R^2: 0.9520 for testing data, indicating the model explains 95.2% of the variance in the testing data.

The model is likely overfit on the training data, but yet manages to perform well on unseen data.

iii. Discrete Naive Bayes

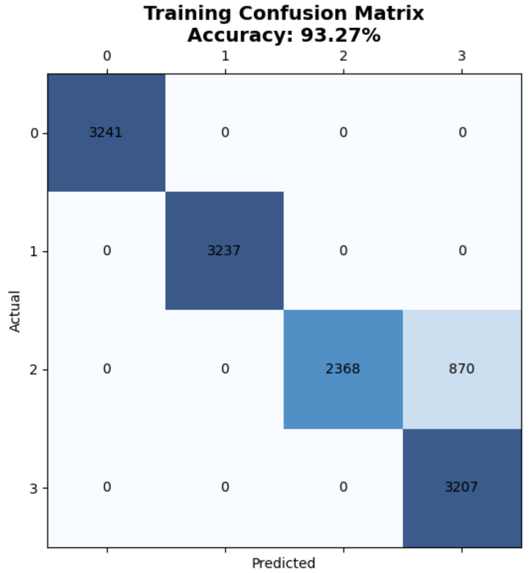


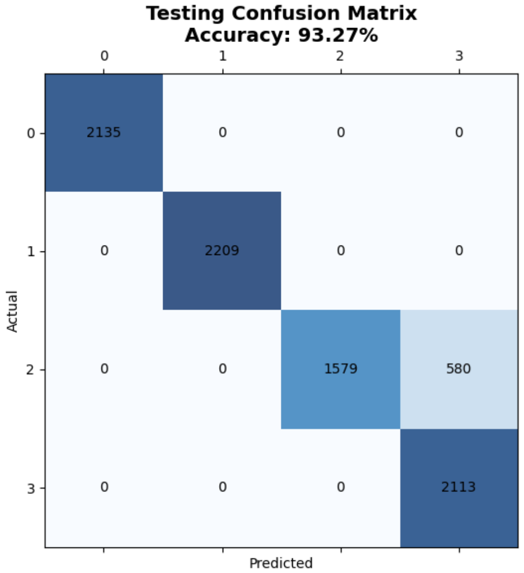


Here too, the confusion pattern between Type 2 and Type 3c persists, suggesting the model struggles to differentiate these classes due to similar features or overlapping data distributions.

A study by Priya et. al claims an accuracy of 96% in detecting diabetes in patients with Naive Bayes approach, but this study is currently under scrutiny by IEEE. [15]

iv. Decision Tree





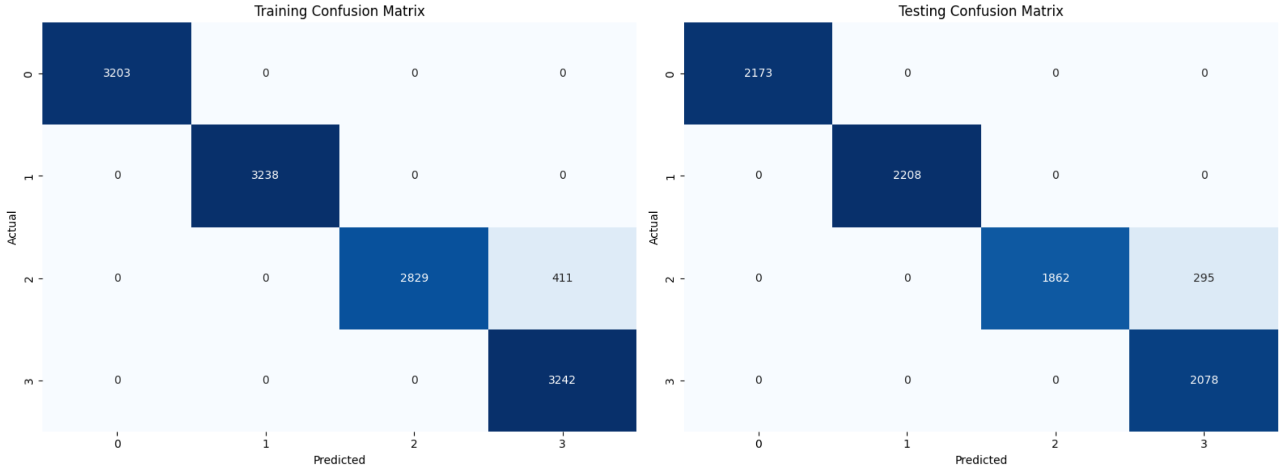
A significant improvement shown here from that previous models, that is the lack of misdiagnosis of Type 3c diabetes as Type 2 diabetes.

A study by Sonar and Jayamalini [16] reports an accuracy of 85% for decision tree approach. Mujumdar and Vaidehi reported two accuracies for this approach : On a private dataset with 8000 patients, they obtained an accuracy of 86%, and with the Pima Indians dataset, reported an accuracy of 74%[18]. This is a reasonable decrease in accuracy given that only 768 patients' data is available in the Pima Indians database,

v. Random Forest

Training Accuracy : 96.81 %

Testing Accuracy : 96.57 %



This ensemble method further improves on the decision tree's accuracy.

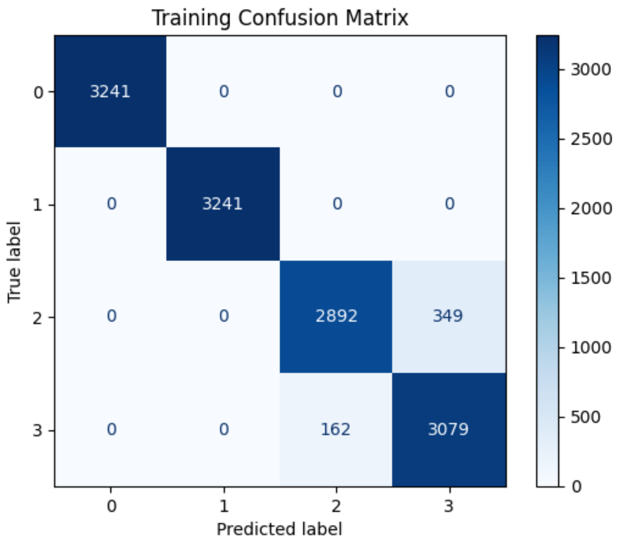
The optimal parameters for each tree was found to be : 50 estimators (trees), with a maximum depth of 10. Each node was split only if there were more than 60 samples, and a minimum of 60 samples in a leaf node was mad necessary in the model.

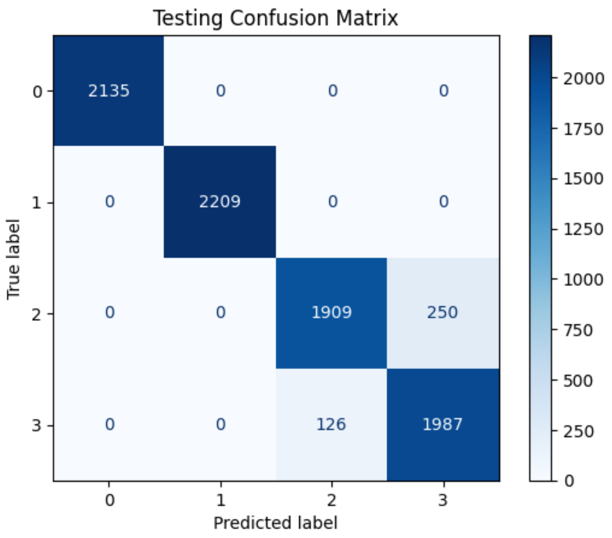
Butwall and Kumar [19] achieved an accuracy of 99.7% with 100 estimators, 20 splits and a maximum depth of 9. However, it is diffucult to substantiate these claims due to the small sample size of 306 used as their database.

vi. XGBoost

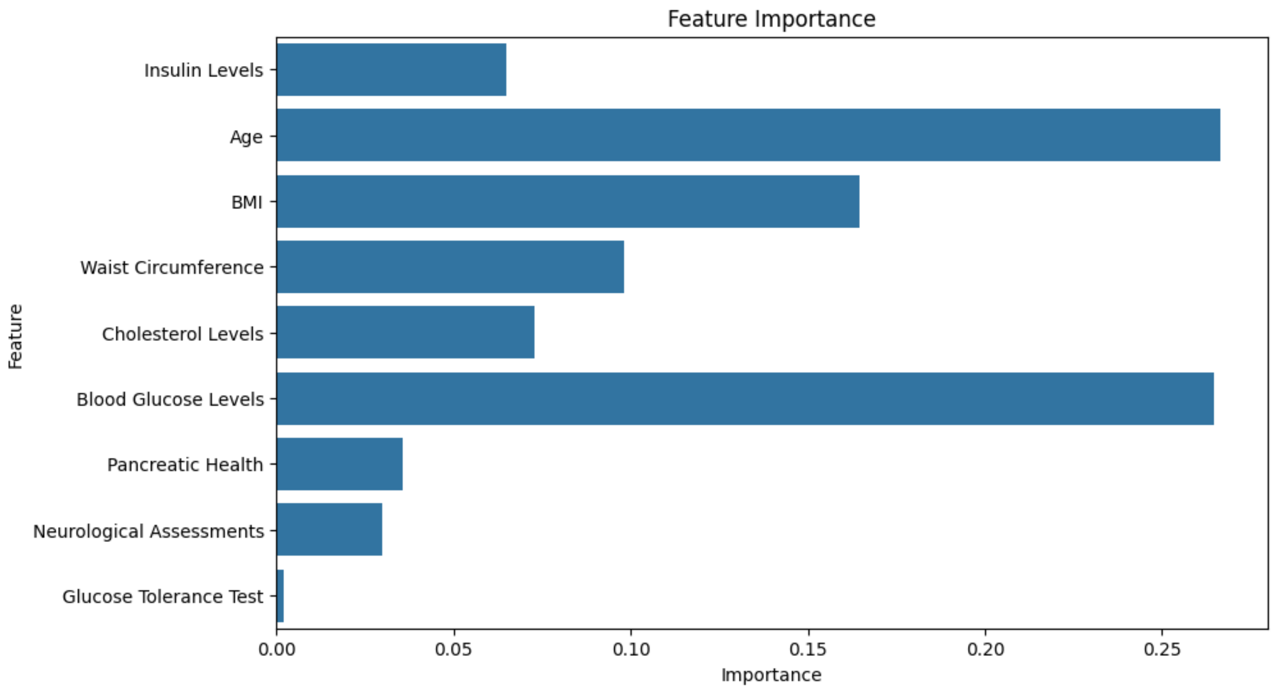
Training Accuracy: 96.06 %

Testing Accuracy: 95.64 %

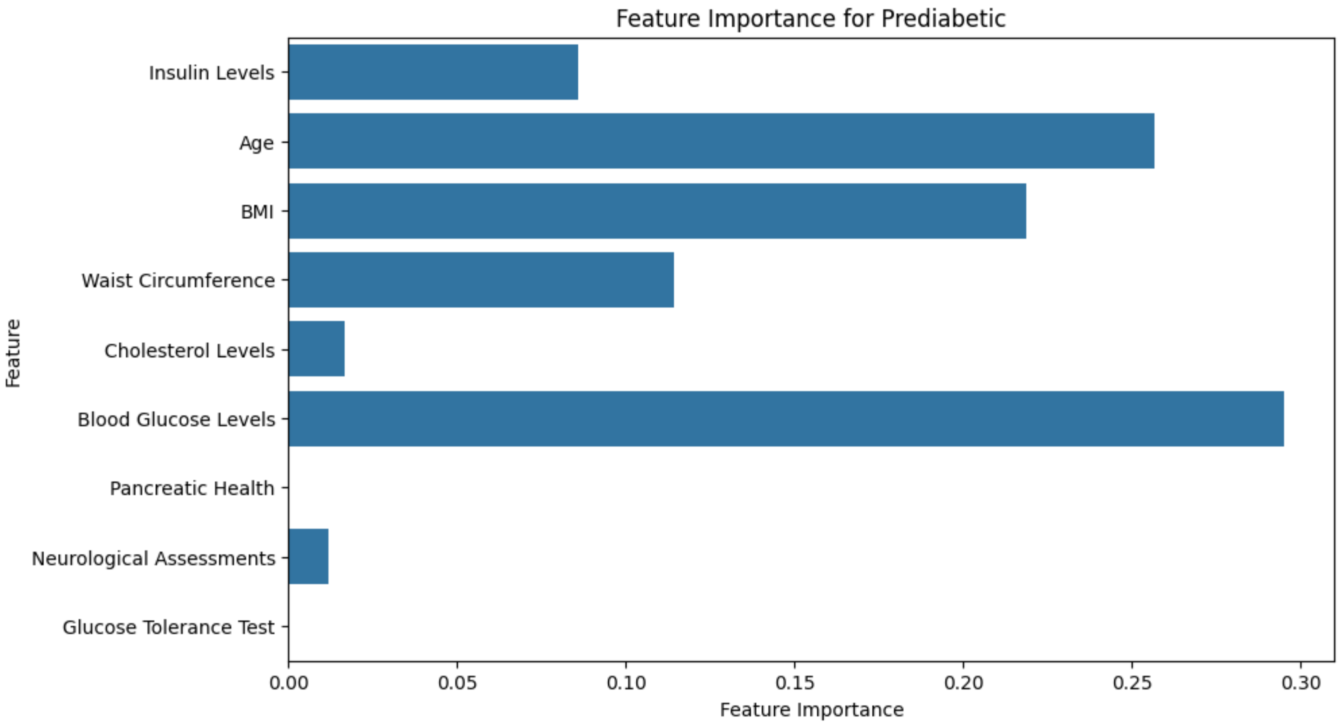


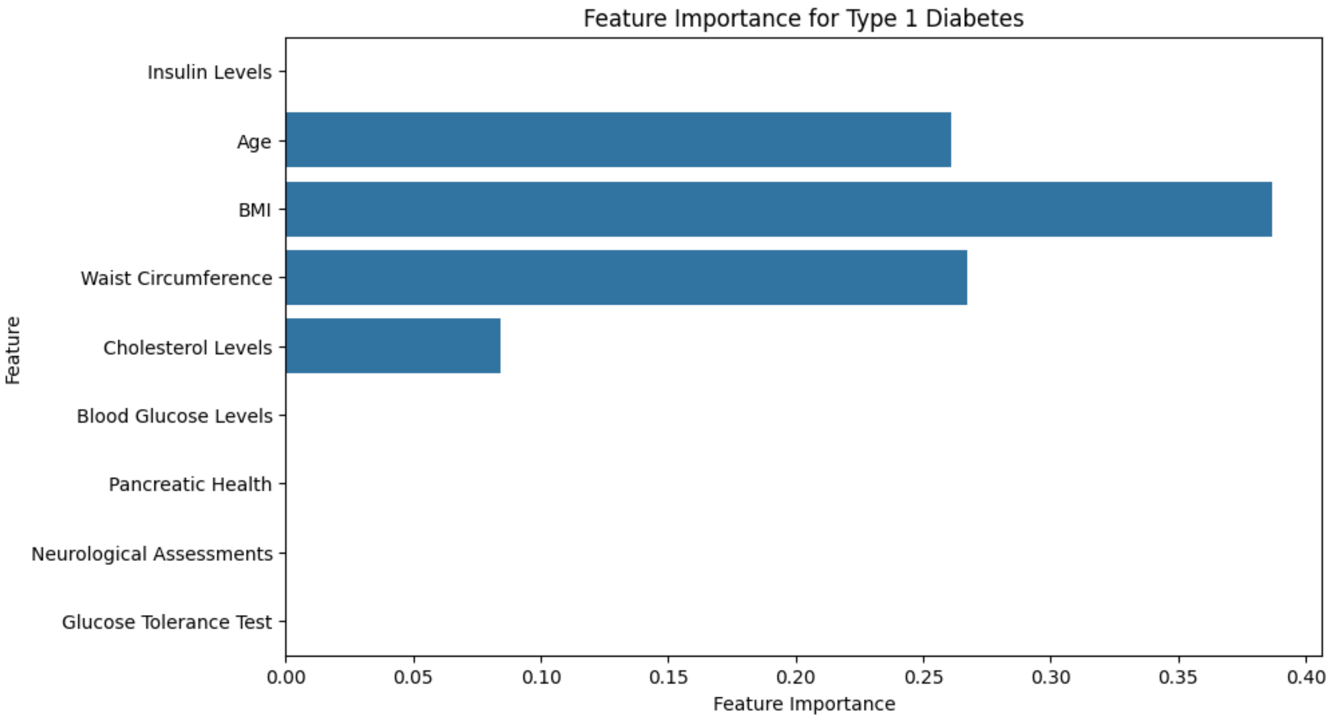


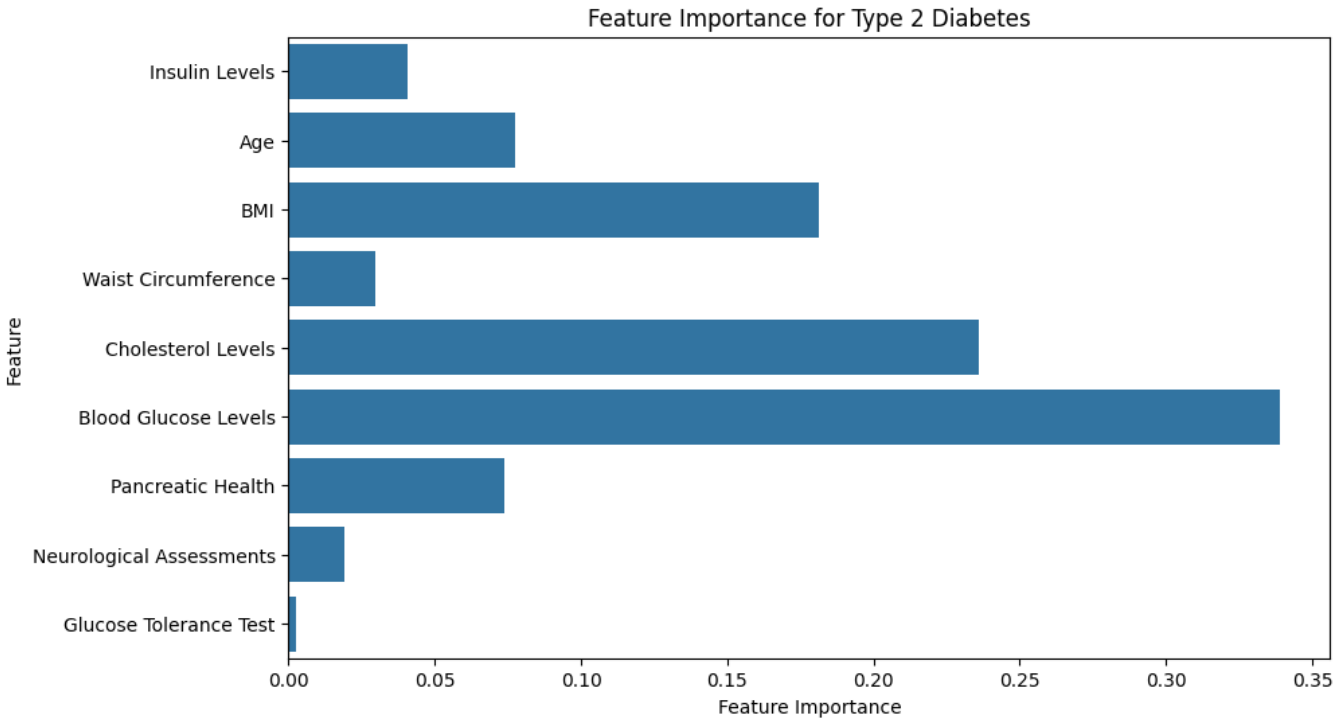
The seaborn library provides a method to assess the overall feature importance, and its output is given below :

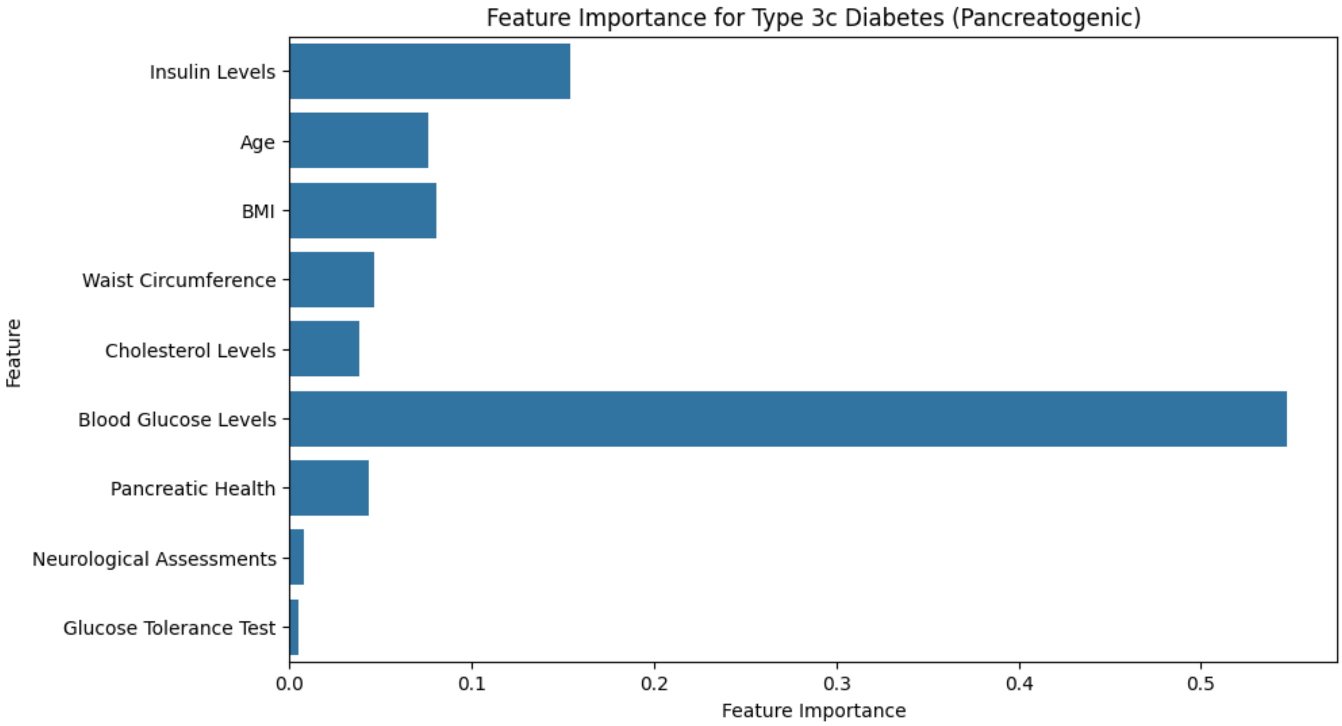


Feature importance for each individual type of diabetes were obtained :


Feature Importance for Prediabetic
Insulin Levels
Age
BMI
Waist Circumference
Cholesterol Levels
Blood Glucose Levels
Pancreatic Health
Neurological Assessments
Glucose Tolerance Test
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Feature Importance

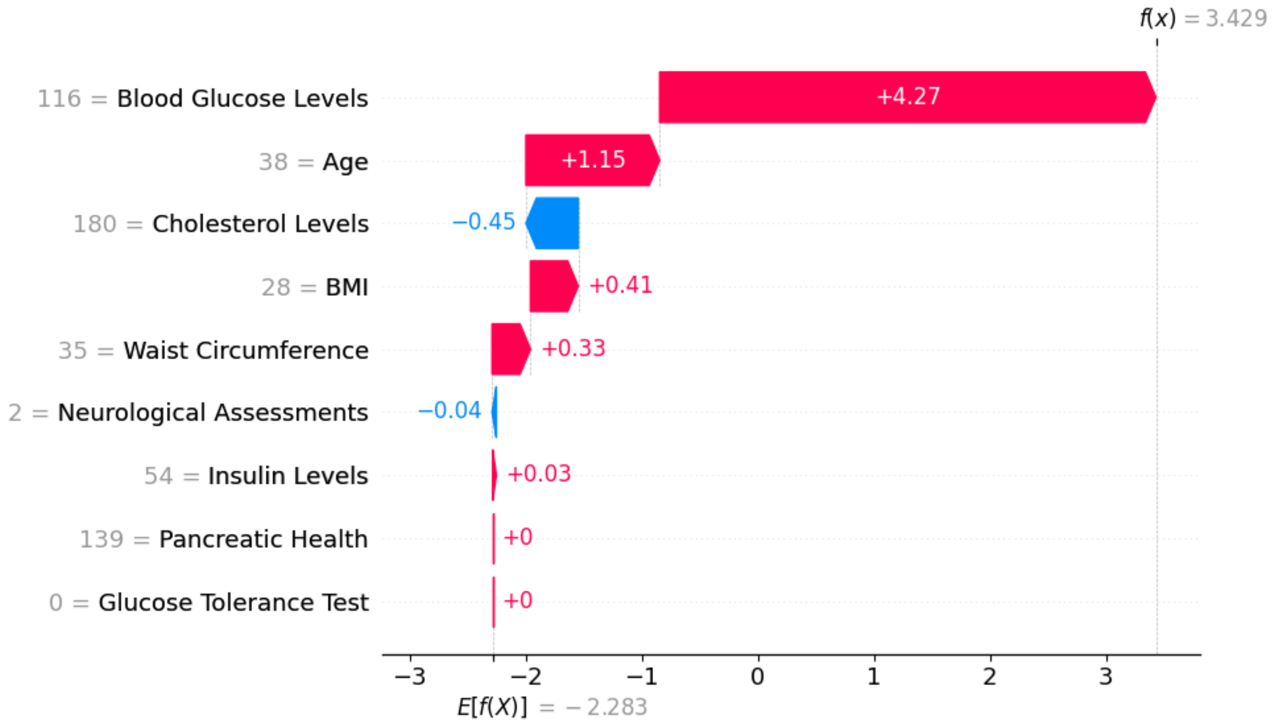

Feature Importance for Type 1 Diabetes
Insulin Levels
Age
BMI
Waist Circumference
Cholesterol Levels
Blood Glucose Levels
Pancreatic Health
Neurological Assessments
Glucose Tolerance Test
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Feature Importance





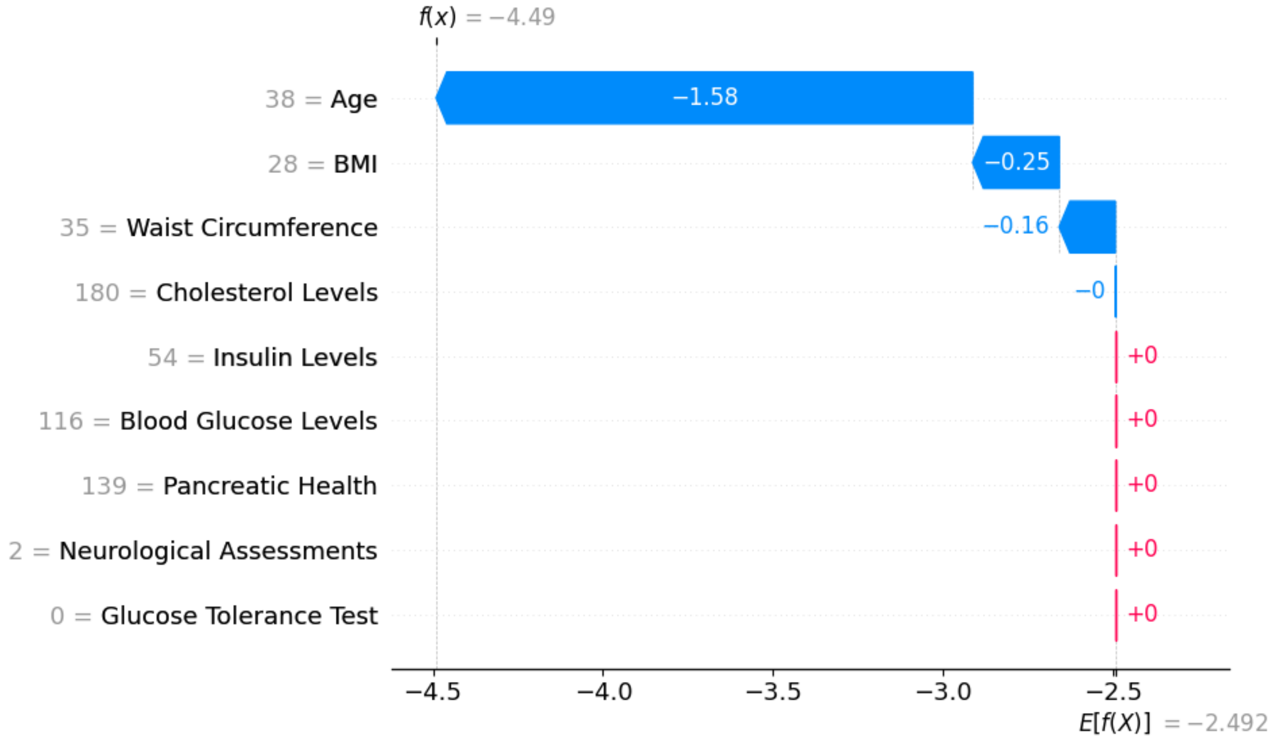
SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. This framework was used to further explain the effects of different attributes on the output and accuracy.

Prediabeties :

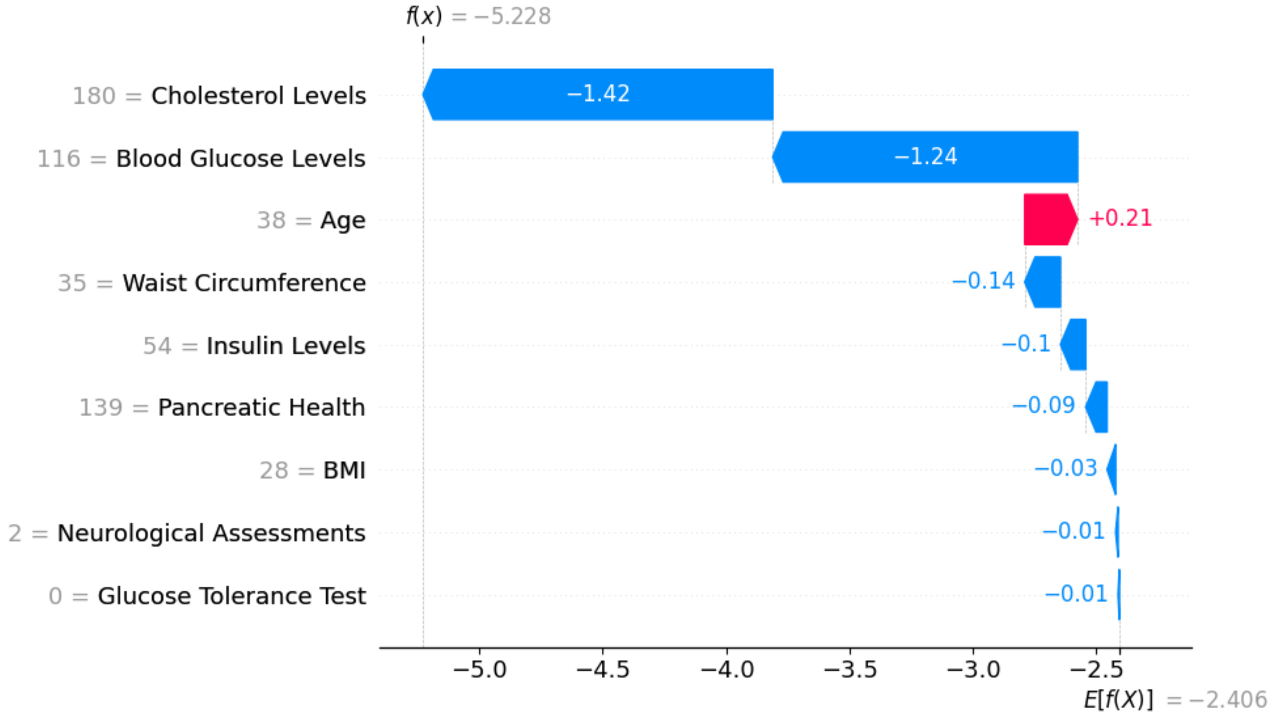


j

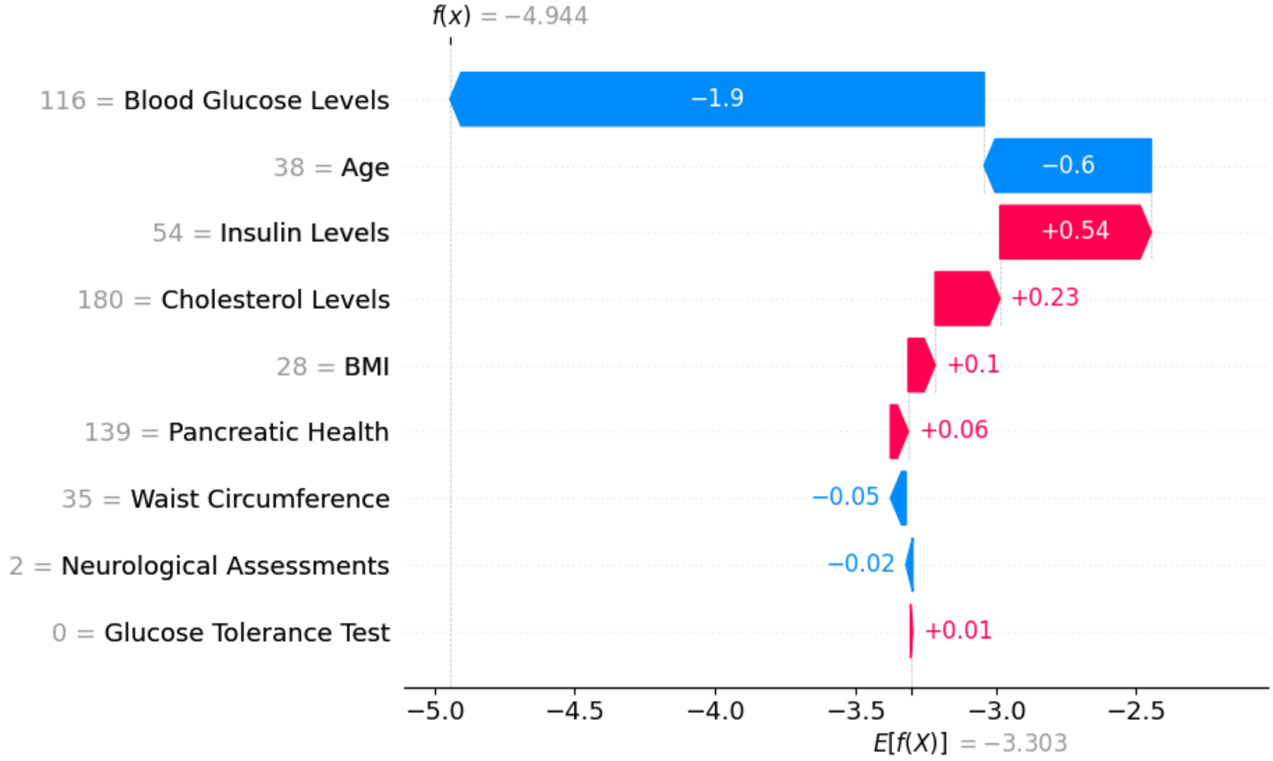
Type 1 Diabetes :



Type 2 diabetes :



Type 3c diabetes :



Bibliography

1) https://my.clevelandclinic.org/health/diseases/21498-prediabetes
2) https://my.clevelandclinic.org/health/diseases/21500-type-1-diabetes
3) https://my.clevelandclinic.org/health/diseases/21501-type-2-diabetes
4) https://my.clevelandclinic.org/health/diseases/24953-type-3c-diabetes
5) https://www.mayoclinic.org/tests-procedures/glucose-tolerance-test/about/pac-20394296

References

1. Johnson JL, Duick DS, Chui MA, Aldasouqi SA. Identifying prediabetes using fasting insulin levels. Endocr Pract. 2010 Jan-Feb;16(1):47-52. doi: 10.4158/EP09031.OR. PMID: 19789156

2. https://doi.org/10.1016/S1542-3565(04)00013-8.

3. https://link.springer.com/article/10.1007/s11892-013-0433-5

4. https://link.springer.com/article/10.1007/s00125-014-3369-7

5. Roche, E.F., Menon, A., Gill, D. and Hoey, H. (2005), Clinical presentation of type 1 diabetes. Pediatric Diabetes, 6: 75-78. https://doi.org/10.1111/j.1399-543X.2005.00110.x

6. Liu, L.L., Lawrence, J.M., Davis, C., Liese, A.D., Pettitt, D.J., Pihoker, C., Dabelea, D., Hamman, R., Waitzfelder, B., Kahn, H.S. and (2010), Prevalence of overweight and obesity in youth with diabetes in USA: the SEARCH for Diabetes in Youth Study. Pediatric Diabetes, 11: 4-11. https://doi.org/10.1111/j.1399-5448.2009.00519.x

7. Peter M. Janiszewski, Ian Janssen, Robert Ross; Does Waist Circumference Predict Diabetes and Cardiovascular Disease Beyond Commonly Evaluated Cardiometabolic Risk Factors?. Diabetes Care 1 December 2007; 30 (12): 3105–3109. https://doi.org/10.2337/dc07-0945

8. Simonen P, Gylling H, Howard AN, Miettinen TA: Introducing a new component of the metabolic syndrome: low cholesterol absorption. Am J Clin Nutr 72: 82–88, 2000

9. Groeneveld, Y., Petri, H., Hermans, J. and Springer, M.P. (1999), Relationship between blood glucose level and mortality in Type 2 diabetes mellitus: a systematic review. Diabetic Medicine, 16: 2-13. https://doi.org/10.1046/j.1464-5491.1999.00003.x

10. Saisho Y, Butler AE, Meier JJ et al (2007) Pancreas volumes in humans from birth to age one hundred taking into account sex, obesity, and presence of type-2 diabetes. Clin Anat 20(8):933–942. https://doi.org/10.1002/ca.20543

11. Ma J, Song Z, Yan F (2014) Detection of hepatic and pancreatic fat infiltration in type II diabetes mellitus patients with IDEAL-Quant using 3.0T MR: comparison with single-voxel proton spectroscopy. Chin Med J 127(20):3548–3552

12. Ou HY, Wang CY, Yang YC, Chen MF, Chang CJ (2013) The association between nonalcoholic fatty pancreas disease and diabetes. PLoS One 8(5):e62561.

13, Lee JS, Kim SH, Jun DW et al (2009) Clinical implications of fatty pancreas: correlations between fatty pancreas and metabolic syndrome. World J Gastroenterol 15(15):1869–1875. https://doi.org/10.3748/wjg.15.1869

14. https://doi.org/10.1016/j.cmpbup.2021.100032

15. K. L. Priya, M. S. Charan Reddy Kypa, M. M. Sudhan Reddy and G. R. Mohan Reddy, "A Novel Approach to Predict Diabetes by Using Naive Bayes Classifier," 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), Tirunelveli, India, 2020, pp. 603-607, doi: 10.1109/ICOEI48184.2020.9142959.

16. P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 367-371, doi: 10.1109/ICCMC.2019.8819841.

17. Soni, Mitushi, and Sunita Varma. "Diabetes prediction using machine learning techniques." International Journal of Engineering Research & Technology (IJERT) 9, no. 09 (2020): 2278-0181.

18. https://doi.org/10.1016/j.procs.2020.01.047

19. Butwall, Mani, and Shraddha Kumar. "A data mining approach for the diagnosis of diabetes mellitus using random forest classifier." International Journal of Computer Applications 120, no. 8 (2015).