**Diabetes Predictions by Using**

**Machine Learning Models**



**Abstract**

Diabetes is a common chronic disease in the United States. The main focus of this paper is to discover which factors are strongly related to diabetes and to design a model that can predict diabetes. BRFSS (The Behavioural Risk Factor Surveillance System) was the source of the original data. For the study, we used a cleaned dataset on Kaggle for the year 2015, which includes 253,680 survey results from the CDC (Centres for Disease Control and Prevention)’s BRFSS related to the target variable diabetes and 10 feature variables. The Chi-square test was used to find out if feature indicators are linked to diabetes and several machine learning models like logistic regression, random forest and decision tree were constructed to help predict the diabetes. The chosen model is Logistic Regression Classifier which has an accuracy of 70.58% on the testing set. In addition to accuracy, precision, recall, F1 score and AUC are other metrics used to assess the model. As per the feature importance applied by the Logistic Regression Classifier, the most important 5 features were high blood pressure (HighBP), General Health (GenHlth), high cholesterol (HighChol), BMI (Body Mass Index) and Age.

**Table of Contents**

|  |  |
| --- | --- |
| **1. Introduction and Problem Description**………………………………………………………………………………… | 1 |
| **2. ABI Technique Review**………………………………………………………………………………………………………….. | 2 |
| 2.1 Machine Learning Classifier ………………………………………………………………………………………….. | 2 |
| **3. Solution Design and Development**………………………………………………………………………………………. | 3 |
| 3.1 Data Description……………………………………………………………………………………………………………. | 3 |
| 3.2 Exploratory Data Analysis……………………………………………………………………………………………… | 4 |
| 3.3 Data Preprocessing………………………………………………………………………………………………………. | 9 |
| 3.4 Model Building……………………………………………………………………………………………………………... | 13 |
| 3.5 Model Evaluation and Validation…………………………………………………………………………………… | 18 |
| **4. Comparison with Existing Research**…………………………………………………………………………………….. | 20 |
| **5. Conclusion and Future Development**…………………………………………………………………………………... | 21 |
| **6. References**……………………………………………………………………………………………………………………........ | 23 |
| **7. Appendix** …………………………………………………………………………………………………………………………….. | 25 |
| 7.1 Appendix 1 - Dataset Attribute Information…………………………………………………………………………… | 25 |
| 7.2 Appendix 2 – CODE…………………………………………………………………………………………………………………… | 26 |

**1.** **Introduction and Problem Description**

Diabetes is a common and highly widespread chronic illness, and it is estimated that over 422 million individuals have diabetes all over the world (Hasan & Yasmin, 2025). Taking care of diabetes well is important to reduce the chance of eye, foot and mouth problems, kidney disease and some types of cancer. Early detection of diabetes is one of the best ways to handle it (Chang et al., 2022). Two methods are used to diagnose diabetes– the first is done by medical professionals and the second is with the help of technology. Early symptoms of diabetes can be hard to find just by conducting manual tests (Chaki et al., 2020). In this context, early identification of diabetes and people at higher risk helps prevent complications and manage the disease well.

The goal of this research is to recognize the early signs of diabetes and take preventive measures using information from actual patients in the USA. The data for this study is a portion of the Behavioural Risk Factor Surveillance System (BRFSS) data published by Kaggle under the name "Diabetes Health Indicators Dataset". It consists of 253,680 observations. The study focused on 11 variables: 10 feature variables and 1 label variable that showed whether someone had diabetes or not. If the binary target variable is “1”, it means the result is positive for diabetic and if it is “0”, it means the result is negative for non-diabetic.

In this study, Logistic Regression, Random Forest and Decision Tree are the machine learning algorithms applied by R software. Besides comparing accuracy, other factors such as precision, recall, F1 score and AUC are evaluated to find out which model performs best. This study consists of exploratory data analysis, data preprocessing, result analysis and evaluation, a comparison with other existing studies and a conclusion.

**2. ABI Technique Review**

It is possible to make real-time decisions with Adaptive Business Intelligence (ABI), as it uses data mining, machine learning and optimization (Michalewicz et al., 2007). ABI systems are different from traditional ones since they keep learning and adjusting to new data which makes them fit for use in areas like healthcare analytics.

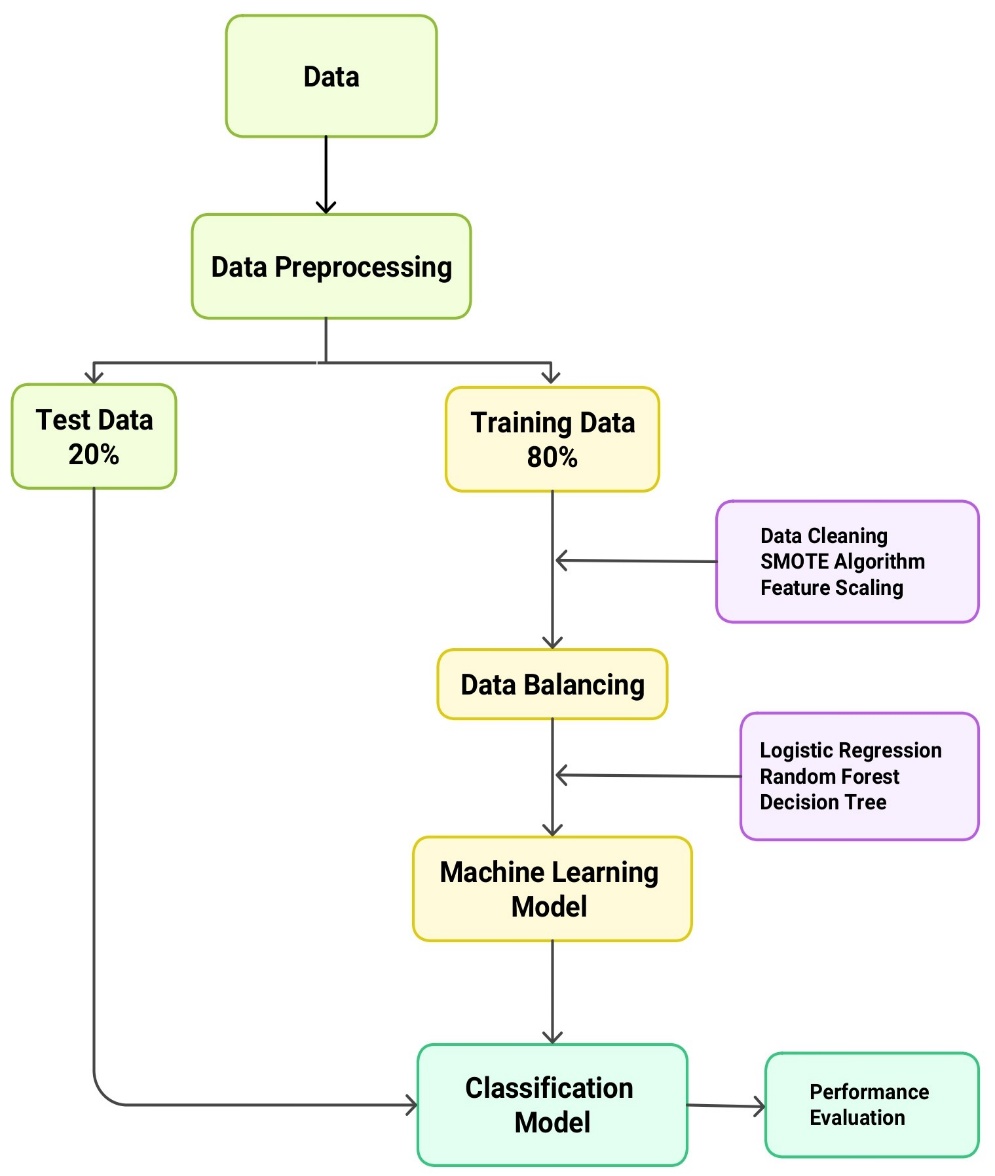
In this case, Logistic Regression, Random Forest and Decision Trees will be used as machine learning methods to find patterns and predict diabetes.

**2.1 Machine Learning Classifier**

* **Logistic Regression:** Logistic regression is used when the dependent variable is binary. It can be used to see how a binary result varies with different independent variables that may be nominal, ordinal, interval or ratio (Hosmer et al., 2013). This type of algorithm uses likelihood and can only produce two results: yes or no, true or false, 0 or 1, high or low. If the probability is less than 0.5, it becomes 0 and if it is more than 0.5, it becomes 1 (Chang et al., 2022).
* **Random Forest:** Random Forest (RF) combines a number of decision trees to build a forest as an ensemble technique. Each tree in the forest is built by using a random part of the data and selecting a set of features. RF is powerful for both making predictions and for determining which independent variables are most important for predicting a dependent outcome (Mehta & Patnaik, 2021).
* **Decision Tree:** Decision Trees help obtain data by applying decision rules to a large number of available datasets (Chang et al., 2022). It is a useful way to show an algorithm that is only made up of conditional control statements. (Bhargav & Sashirekha, 2023).

**3. Solution Design and Development**

The step-by-step problem solution and development are outlined in Figure 1. It represents the stages of data collection, analysing it and making sense of what it means.



**Figure 1: Architecture Design**

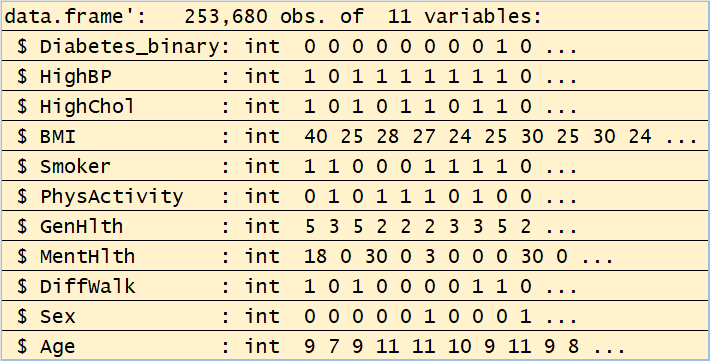
**3.1 Data Description**

The BRFSS collected the US Diabetes Health Indicators Dataset which is now published and shared by Alex Teboul on Kaggle. There are 253,680 observations in the dataset that represent adult respondents from the United States. This study examined 10 feature variables from the dataset such as High Blood Pressure, High Cholesterol, Body Mass Index (BMI), Smoker, Physical Activity, General Health Conditions, Mental Health, Difficulty Walking, Sex and Age (detailed in Appendix 1). The target variable is set to binary with 1 meaning diabetes and 0 meaning no diabetes.

**3.2 Exploratory Data Analysis**

Before using machine learning models, it is essential to use Exploratory Data Analysis (EDA) to discover the structure and patterns in the dataset.

**3.2.1 Dataset Overview**



**Table 1: Dataset Structure**

Table 1 shows the layout of the diabetes dataset which consists of 253,680 observations and 11 variables. The target variable serves to indicate whether a person is non-diabetic (0) or diabetic (1). Most predictors are health-related and demographic features, including high blood pressure, high cholesterol, smoker, physical activity, difficulty walking and Sex and are either binary or ordinal. BMI, MentHlth (mental health days) and Age are all continuous variables with a high range of values.

**3.2.2 Missing Value**

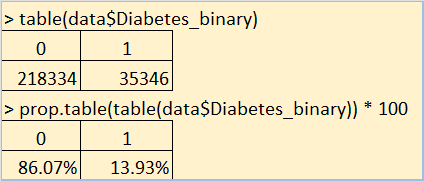
A table with numbers and letters

AI-generated content may be incorrect.

**Table 2: Missing Value Check**

Table 2 shows that all the variables are clean, no missing values, which suggests the data is in good for early diabetes detection.

**3.2.3 Class Distribution of Target Variable**



**Table 3: Class Distribution**

The data demonstrates that, among all individuals, 86.07% are non-diabetic (label 0) and only 13.93% are diabetic or prediabetic (label 1), meaning the dataset is highly imbalanced (Table 3). If the data is not balanced, machine learning models may favour the majority class (non-diabetic) which can result in not identifying as many people at risk for diabetes. A solution to this is achieved with algorithms like the SMOTE (Synthetic Minority Over-sampling Technique).

**3.2.4 Descriptive Statistics**

A table with numbers and letters

AI-generated content may be incorrect.

**Table 4: Descriptive Statistics**

The descriptive statistics point out that BMI, HighBP, HighChol, GenHlth and MentHlth are important factors related to diabetes risk. The BMI average is 28.38 and there is a clear right skew (2.12) which points to most people being overweight and a higher risk of diabetes. Moreover, 43% of people have high blood pressure and 42% have high cholesterol, both problems found in people with diabetes.

**3.2.5 Visualize Distributions (Histograms)**

**A graph of a graph

AI-generated content may be incorrect.**

**Figure 2: BMI Distribution**

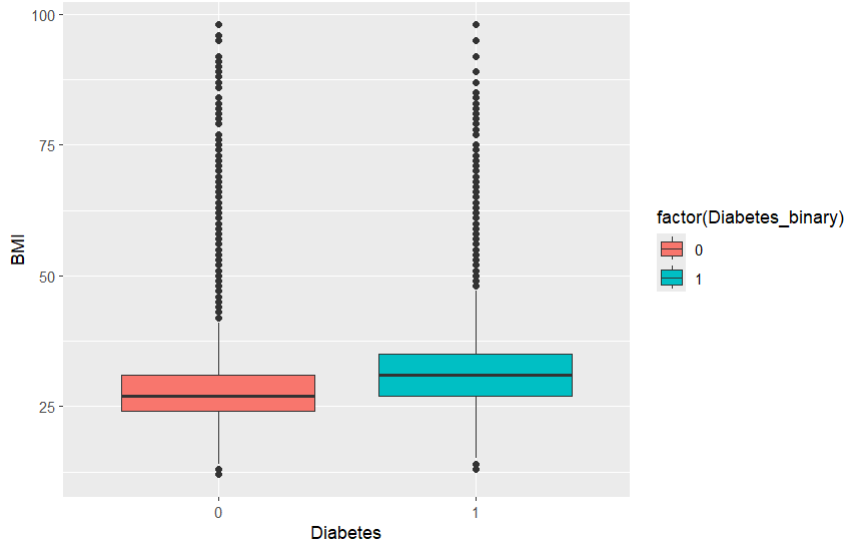
**A graph with red squares

AI-generated content may be incorrect.**

**Figure 3: Mental Health Days**

Figure 2 demonstrates that the majority of individuals in this group have a BMI between 25 and 35 which suggests they are mainly overweight. As Figure 3 shows, the data is right-skewed, indicating that a lot of people experience no poor mental health days, but a significant number report struggling with mental health daily for as long as 30 days. Such patterns suggest that both physical and mental wellness play a role in identifying people who are more likely to get diabetes.

**3.2.6 Group Comparisons with Boxplots**



**Figure 4: BMI by Diabetes Status**

In the boxplot (figure 4), the median BMI is generally higher among individuals with diabetes (group 1) than among non-diabetic individuals (group 0). The interquartile range is wider and the BMI range more extreme for the diabetic group which suggests a larger variation among them and a higher likelihood of obesity.

**3.2.7 Bivariate Bar plots (Categorical Features)**

A graph showing a couple of blue squares

AI-generated content may be incorrect.

**Figure 5: Diabetes vs High Blood Pressure**

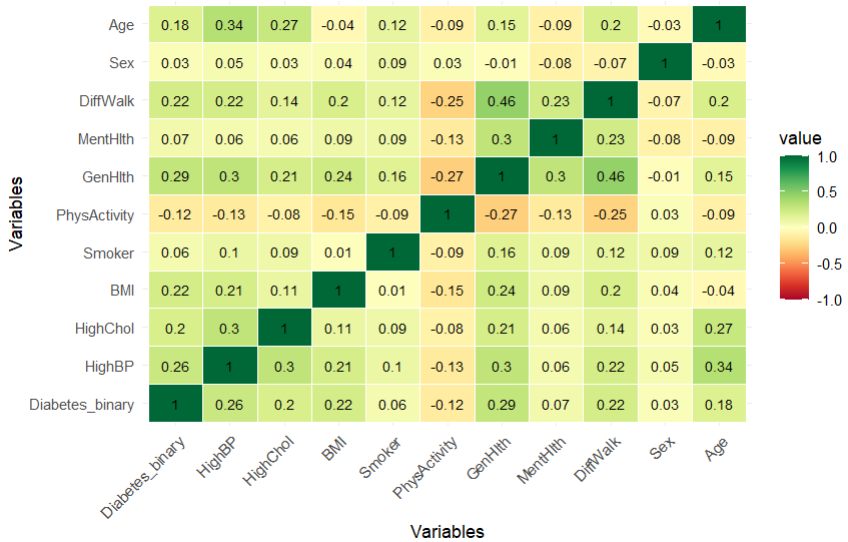
A graph of a diabetes

AI-generated content may be incorrect.

**Figure 6: Diabetes vs Physical Activity**

In Figure 5, we can see that more diabetic individuals (label 1) have high blood pressure which means that high blood pressure is a frequent issue and a major risk factor in people with diabetes. According to Figure 6, diabetics are much less likely to be physically active than non-diabetics which might suggest that being inactive can lead to a higher risk of diabetes. As a result, the early diabetes prediction models should take into account both high blood pressure and physical activity.

**3.2.8 Correlation Matrix**



**Figure 7: Correlation Heatmap of Diabetes Dataset**

The correlation heatmap (Figure 7) shows that there is a strong connection between Diabetes\_binary and HighBP (0.26), GenHlth (0.29), BMI (0.22) and DiffWalk (0.22), meaning that individuals with these conditions have a higher chance of being diabetic. While there is a slight positive correlation (0.12) between physical activity and diabetes, a negative correlation (−0.12) also exists, meaning that lower levels of physical activity may increase the risk of diabetes. Because of these patterns, including these variables helps create a reliable early diabetes detection model.

**3.3 Data Preprocessing**

**3.3.1 Feature selection**

It is important to select the proper features because this helps to remove irrelevant ones and to select those best suited for the target variable. Application of feature selection makes the classifier to reduce the execution time and perform better.

* **Feature correlation**

To identify relationships between the target variable “Diabetes\_binary” and the features, the correlation plot (Figure 8) was used. Things like general health, high blood pressure, walking difficulty, BMI and BMI, as well as age, were strongly connected to diabetes, whereas physical activity showed a weak negative relationship (−0.12). The study pointed out the main factors that can lead to diabetes.

**A graph of a number of purple bars

AI-generated content may be incorrect.**

**Figure 8: Correlation Diagram**

* **Chi-square test**

As correlation does not guarantee useful features, a Chi-square test was conducted to investigate whether each feature predicted the target variable with statistical significance (Bahassine et al., 2020).

A table with numbers and letters

AI-generated content may be incorrect.

**Table 5: Feature with Chi-square Score**

The Chi-square test allows us to find which categorical features have a strong link to the target variable. Since the Chi-square scores were high for the top 8 features, they were used for model training and ‘smoker’ and ‘sex’ with low scores were eliminated from the data. The features that were eliminated are found to have a lower correlation in the previous correlation matrix.

**3.3.2 Data splitting**

Splitting the dataset must be done before training the model. All the values in the dataset are cleaned and no missing values. Rearranging the dataset randomly ensures that all the data is distributed randomly. If the dataset is not shuffled in machine learning, the model might develop a bias, reduce its effectiveness in generalization and the accuracy of training. Low correlation variables were not included anymore and the top 8 features were used for training and testing the model. The test set consists of 20% of the data and the remainder 80% goes to the train set (Table 6).

A yellow and black rectangular table with numbers and text

AI-generated content may be incorrect.

**Table 6: Train and Test data**

A table with numbers and letters

AI-generated content may be incorrect.

**Table 7: Proportion of Diabetic and Non-Diabetic in Train & Test Dataset**

**3.3.3 SMOTE Algorithm**

As shown in Table 7, there are more non-diabetic records (86.07%) than diabetic records (13.93%). Bias and poor training results may happen if the training dataset is not well balanced. For this reason, the data imbalance must be corrected before creating the model. The SMOTE (Synthetic Minority Oversampling Technique) preprocessing algorithm was used to address this issue. The main point of SMOTE is to produce extra synthetic samples for the minority group until both classes have equal numbers (Mishra, n.d.). It also enables classifiers to better generalise the outcomes of their training. Figure 9 shows the change in how diabetes is distributed before and after oversampling.

A graph of a diabetes distribution

AI-generated content may be incorrect.

A graph of a diabetes distribution

AI-generated content may be incorrect.

**Figure 9: Diabetes Distribution Before and After Oversampling.**

**3.3.4 Feature scaling:**

One of the last things to do in data preprocessing is feature scaling which helps models perform better by treating all numeric features as if they are on the same scale. We only used standardization for BMI, MentHlth and Age in this study. Some ranges are wider, for instance, BMI can be above 90 and Mental Health ranges only from 0 to 30. If the model is not scaled, the effects of large amounts could make things less fair. We didn’t change the other variables because they are binary (0 or 1).

A table with numbers and letters

AI-generated content may be incorrect.

**Table 8: Mean and Standard deviation of BMI, Mental Health, Age after Feature Scaling**

According to Table 8, the features BMI, mental health and age are all scaled so that their means are near 0 and their standard deviations are 1. The result of this process is that the data is now standardized, all three features are on the same scale and models can treat each feature equally during training.

**3.4 Model Building**

**3.4.1 Logistic Regression Model**

* **Model Description**

A logistic regression model is created by using the training data in this study. It forecasts the possibility that a person has diabetes based on the measured indicators of health. The training data (train\_data) contains 80% of the observations which is 202,945 and the testing data (test\_data) contains the remaining 20% or 50,735 observations. Training data is used to develop the model and testing data is used to see how well it performs.

A table with numbers and a few digits

AI-generated content may be incorrect.

**Table 9: Summary of the Logistic Regression Model**

From table 9, we can say that the model’s predictors are significant (p-value < 0.05) and thus have a meaningful impact on predicting diabetes. Of all these, the strongest are high blood pressure, high cholesterol, general health condition and BMI. Thus, people with higher blood pressure, higher cholesterol, worse general health and greater BMI are more likelihood of getting diabetes.

A screenshot of a calculator

AI-generated content may be incorrect.

**Table 10: Confusion Matrix of Logistic Regression Model**

The confusion matrix (Table 10) shows how the model predicted diabetes cases and how these predictions match with the actual results in the test data. Using the model led to 70.58% accuracy which means it was able to predict diabetes status in about 71 out of every 100 cases. The balanced accuracy is 74.29% which means the model does quite well for both diabetic and non-diabetic groups. The model was correct in identifying around 69% of true diabetic cases (sensitivity) and also accurate in detecting about 79% of non-diabetic cases (specificity).

* **Feature Importance**

According to the feature importance plot (Figure 12), high blood pressure, general health, and high cholesterol play the biggest role in diagnosing diabetes, and BMI and Age are ranked lower. The absolute values for these features are the highest which means they are strongly related to the chances of having diabetes. Conversely, mental health, physical activity and difficulty walking play a smaller role which means their impact on the model is not as strong. It allows diabetes risk to be assessed by focusing on the main health indicators.

A graph with blue bars

AI-generated content may be incorrect.

**Figure 12: Feature Importance**

**3.4.2 Random Forest (RF)**

* **Model Description**

Random forest was applied with 100 trees and the target variable is Diabetes\_binary. For this model, the confusion matrix (table 11) reveals an accuracy of 65.35% which means about 65 out of 100 patients were given the correct diagnosis. Sensitivity for diabetic cases is 63.45% which is lower than the specificity of 77.13% for non-diabetics. A very high percentage, 94.49%, of those who are predicted to have diabetes are actually found to have it. Nevertheless, the low negative predictive value (25.46%) indicates that diagnosing someone as not having diabetes may not be very accurate. A moderate balance between detecting classes is shown by the balanced accuracy of 70.29%.

A screenshot of a computer

AI-generated content may be incorrect.

**Table 11: Confusion Matrix of Random Forest Model**

* **Feature Importance**

As shown in Figure 10, HighBP, GenHlth and HighChol contribute most to how accurate the model is, with Age and BMI as the next factors. These features play the biggest role in deciding if someone has diabetes.

A comparison of a graph

AI-generated content may be incorrect.

**Figure 10: Variable importance plots**

**3.4.3 Decision Tree**

Machine learning often uses decision trees as a common method. Table 12 reflects a moderate ability to predict diabetes. The model’s overall accuracy is 76.59%, so it correctly predicts if someone has diabetes. The model can identify actual diabetic cases well (sensitivity of 0.7916), but it has a lower specificity (0.6073) which means it struggles to detect non-diabetic cases. Since the positive predictive value is high (0.9257), the model is generally correct when it predicts a patient has diabetes. The low negative predictive value (0.3205) means it can be challenging to confidently exclude diabetes in this case. Since there is class imbalance, the balanced accuracy of 0.6994 gives a fair judgment by considering sensitivity and specificity together.

A screenshot of a calculator

AI-generated content may be incorrect.

**Table 12: Confusion Matrix of Decision Tree**

**3.5 Model Evaluation and Validation:**

* **Confusion Matrix Analysis**

Data in the dataset is cleaned and normalized before being given to the model. We do not use imbalanced datasets since they may show higher accuracy because they favor the majority group, without being able to identify the minority group which is more significant in healthcare (Krawczyk, 2016). All three models show good accuracy with this imbalanced data set, but the AUC scores are lower than those seen with balanced data sets (Table 13). Since the training data is imbalanced, the decision tree only predicts class 0 and ignores the minority class (1). For these reasons, we avoid imbalanced data results. SMOTE is used to increase the number of minority class data points in a dataset. The Pareto principle calls for the dataset to be split 80:20. When the training is done, the classifiers are evaluated using various metrics found in the confusion matrix. Accuracy, Precision, recall, specificity, F1 score, AUC and ROC-AUC curve are some of the important model evaluation metrics (Table 14).

A table with numbers and letters

AI-generated content may be incorrect.

**Table 13: Performance Matrix before Implementing Data Balancing Technique**

**A screenshot of a graph

AI-generated content may be incorrect.**

**Table 14: Performance Matrix after Implementing Data Balancing Technique**

* **ROC Curve & AUC**

A graph with different colored lines

AI-generated content may be incorrect.

**Figure 11: ROC Curve Comparison**

Logistic Regression is preferred over Random Forest and Decision Tree because it gives the most balanced and dependable performance in important evaluation areas. While the Decision Tree is more accurate (76.59%), Logistic Regression gives the highest AUC (81.63%), making it more capable of detecting the difference between diabetic and non-diabetic cases. This model also gets the highest recall (79.43%), allowing more real diabetic patients to be detected which is vital for medical diagnosis. Besides, its reliable F1-score (42.93%) and specificity (69.14%) indicate it can achieve a good balance between recall and precision. Because of these combined strengths, Logistic Regression is the most suitable and trusted choice for predicting diabetes in this case.

* **Prediction Testing on Simulated Patient Input**

We first trained the model and saved it and later we used it to check predictions for new data. To choose our test cases, we first looked at the head and tail of the dataset (Figure 12 and 14). Next, a function was made to add patient information manually, with risk factors such as high blood pressure, high cholesterol, BMI, level of physical activity, general health, mental health, difficulty walking and age.

By applying this setup, we verified the model on the 5th patient and found it classified the patient as a non-diabetic (0) which was in line with the original label in the data (Figure 13). A white background with black and blue text

AI-generated content may be incorrect.

**Figure 12: First Five Observations of the Diabetes Dataset**



**Figure 13: Prediction Output**

In the same way, the model predicted the 253,677th patient to be diabetic (1) and this proved to be accurate since the true result was also diabetic (Figure 15).

A list of different colored letters

AI-generated content may be incorrect.

**Figure 14: Last Five Observations of the Diabetes Dataset**

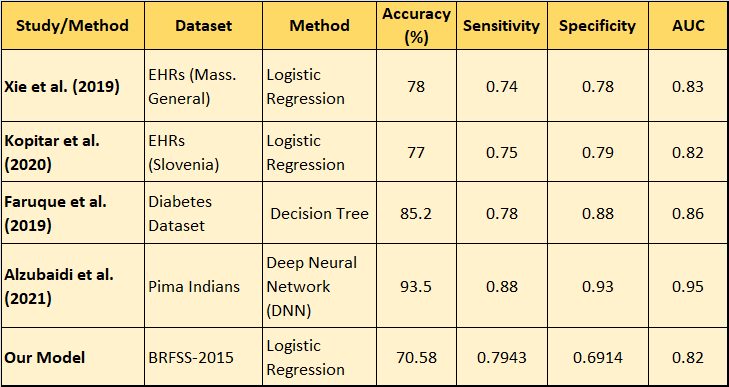


**Figure 15: Prediction Output**

Such reliable predictions prove that logistic regression can handle the classification of diabetic and non-diabetic patients.

**4. Comparison with Existing Research**

Looking at our logistic regression model and findings from previous studies lets us check how well our model functions compared to different datasets and algorithms. It gives a quick look at how the models perform on different datasets.



**Table: Comparison of Early Studies**

Xie et al. (2019) obtained 78% accuracy by applying logistic regression to EHR data from Massachusetts General Hospital and the AUC was 0.83. Also, Kopitar et al. (2020) performed logistic regression on Slovenian EHRs and got 77% accuracy, with an AUC value of 0.82. Faruque et al. (2019) trained a decision tree on diabetes data and received 85.2% accuracy which matched with high sensitivity (0.78) and specificity (0.88). Alzubaidi et al. (2021) relied on a deep neural network on the Pima Indian dataset and obtained high results: accuracy of 93.5%, sensitivity of 0.88, specificity of 0.93 and an AUC of 0.95. Instead, our model got 70.58% accuracy, 0.7943 sensitivity, 0.6914 specificity and an AUC score of 0.82.

**5. Conclusion and Future Development:**

This study uses a dataset from BRFSS in 2015 that gives the distribution of diabetes disease. The objective is to discover which features affect diabetes and to develop a predictive model with reasonable accuracy. The selected 8 features were chosen using the correlation diagram and the Chi-square test from among 11 total features. After finishing the preprocessing of the data and the training of the model, it is suggested that the Logistic Regression classifier is suitable for identifying diabetes using the selected features. But the model still has some limitations. The accuracy of the model is not what was expected. An explanation could be that the data is not equally balanced. The training data was adjusted using SMOTE to make it balanced, but the testing set has not changed and remains unbalanced (it should not be used SMOTE for the entire dataset, since this affects the purity of the dataset and cause the model to be overfitted ). When the test set is not balanced, the model may not be as accurate on it. Improving the performance of the Logistic Regression Classifier in the future will probably require gathering more data from diabetic patients.

**6. References**

‌Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00444-8

‌Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2020). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University - Computer and Information Sciences*, *32*(2), 225–231. https://doi.org/10.1016/j.jksuci.2018.05.010

Bhargav, P., & Sashirekha, K. (2023). A Machine Learning Method for Predicting Loan Approval by Comparing the Random Forest and Decision Tree Algorithms. *Journal of Survey in Fisheries Sciences*, *10*(1S), 1803–1813. https://doi.org/10.17762/sfs.v10i1S.414

‌Chaki, J., Thillai Ganesh, S., Cidham, S. K., & Ananda Theertan, S. (2020). Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. *Journal of King Saud University - Computer and Information Sciences*, *34*(6). https://doi.org/10.1016/j.jksuci.2020.06.013

Chang, V., Ganatra, M. A., Hall, K., Golightly, L., & Xu, Q. A. (2022). An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics*, *2*, 100118. https://doi.org/10.1016/j.health.2022.100118

Faruque, M. F., Asaduzzaman, & Sarker, I. H. (2019). *Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus*. ArXiv.org. https://arxiv.org/abs/1902.10028

Hasan, M., & Yasmin, F. (2025, May 11). *Predicting Diabetes Using Machine Learning: A Comparative Study of Classifiers*. https://doi.org/10.48550/arXiv.2505.07036

‌Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-68771-z

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

‌Mehta, S., & Patnaik, K. S. (2021). Improved prediction of software defects using ensemble machine learning techniques. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-021-05811-3

Mishra, S. (n.d.). Handling Imbalanced Data: SMOTE vs. Random Undersampling. In *International Research Journal of Engineering and Technology*. Retrieved May 26, 2025, from https://www.irjet.net/archives/V4/i8/IRJET-V4I857.pdf

TEBOUL, A. (2022). *Diabetes Health Indicators Dataset*. Www.kaggle.com. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Preventing Chronic Disease*, *16*. https://doi.org/10.5888/pcd16.190109

‌

**7. Appendix**

**7.1 Appendix 1 -Dataset Attribute Information**

