Prediction of Diabetes in Women using Machine Learning Models

Bhavesh Gujjula
Masters in Data Science
University of North Texas
Texas, USA
bhaveshgujjula@my.unt.edu

Srinivas Pallapu
Masters in Data Science
University of North Texas
Texas, USA
srinivaspallapu@my.unt.edu

Jaya Simha Reddy pocha
Masters in Data Science
University of North Texas
Texas, USA
jayasimhareddypocha@my.unt.edu

Vidyadhari bheemarpu
Masters in Data Science
University of North Texas
Texas, USA
vidyadharibheemarpu@my.unt.edu

Farishta, Inara Karim
Masters in Advanced Data Analytics
University of North Texas
Texas, USA
inarafarishta@my.unt.edu

***Abstract*—Type II diabetes mellitus is a condition that interferes with how the body manages and utilizes sugar (glucose) as energy. Far too much sugar is flowing into the bloodstream due to this chronic condition. Disorders of the immunological, neurological, and circulatory systems can result from high blood sugar levels. Approaches of Machine Learning (ML) demonstrated their effectiveness in the diagnosis process of diabetes. The variables and metrics, equivalent to pregnancy, and blood pressure, may give rise to more precise prediction and classification systems of Type II diabetes in women. This paper examined the value of decision tree classifiers and supervised machine learning algorithms. Women could be able to undertake the decisive steps to stop the emergence of Type II diabetes by recognizing these indications. The Pima Indian Women Diabetes dataset was taken from the Kaggle website to apply Machine Learning algorithms. We conducted several comparative experiments to assess. Algorithms like decision tree(ID3), naive Bayes and Random Forest will be developed to compare the effectiveness of the decision tree classifiers and choose the top model. This paper also discusses the various benefits and drawbacks of prediction techniques and correlates the degree of data categorization accuracy.**

I. Introduction

One of the chronic, incurable diseases known as diabetes is brought on by a deficiency or absence of the hormone insulin [1]. It is a vital hormone that the pancreas secretes that enables cells to absorb glucose (blood sugar) from dietary sources to give them the energy they require [2]. Hyperglycemia is the medical name for the condition with elevated blood sugar levels. There are two potential causes for this situation: When the body cannot produce the insulin that the blood cells need, it is also unable to react to insulin as it should. For blood glucose to enter body cells and be used as fuel, insulin is required by the body. However, glucose builds up in the blood and causes hyperglycemia if the body cannot use it to make energy. Stroke, nonketotic hyperosmolar cardiovascular disease, and diabetic ketoacidosis are just a few of the significant health issues that can result from this. The World Health Organization estimates that 422 million people worldwide have diabetes, making it one of the top causes of death. It resulted in 1.6 million fatalities in 2016 [3]. Type 1 and type 2 diabetes are the two primary subtypes. 5 to 10 percent of all cases of diabetes are diabetes type 1. The pancreas only partially functions in diabetes, typically in childhood or adolescence. As long as the pancreas is still partially functional, type 1 diabetes does not initially manifest symptoms. Only 80–90 percent of pancreatic insulin-producing cells are damaged before the condition is diagnosed [4].

90 percent of instances of diabetes are type 2 diabetes. Chronic hyperglycemia and the body's inability to control blood sugar levels, which result in an excessively high blood glucose (sugar) level, are characteristics of this kind of diabetes. Most of those affected by this condition are obese or overweight older individuals [5].

According to medical professionals and recent studies, the likelihood of recovery is higher if the sickness is found early. Machine learning and deep learning approaches are now very helpful in disease analysis and early prediction because of the ongoing improvement of technology. The methods employed in the study to predict diabetes include Logistic Regression, Decision Tree, k-nearest neighbor algorithm (k-NN), naïve Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF) [6].

Recently, several academics have concentrated on applying deep learning and machine learning to predict diabetes. For example, writers in [7] provided theoretical research based

on four machine learning algorithms: Linear Discriminant Analysis (LDA), K-nearest Neighbor (KNN), Support Vector Machine (SVM), and Random Forest (RF). The Pima Indians Diabetes data collection has been used for testing the proposed method. The suggested system's high classification accuracy (87.66 percent) demonstrates how well the RF works to aid medical professionals in disease prediction.

Using machine learning methods, including a k-nearest neighbor, decision trees, random forests, AdaBoost, Naive Bayes, XG Boost, and multilayer perceptrons, the study's authors [3] have presented a model for early diabetes prediction. For example, they removed an outlier, checked for null values, chose characteristics, and used K-fold cross-validation to test the model, among other data pre-processing techniques. Respectively. [8], diabetes is predicted using Decision Trees, SVM, and NB. With Naive Bayes algorithms, they achieved an AUC of 0.819 percent as their maximum.

Three ML methods were applied to the Pima Indian Diabetes dataset analysis in [9]. Accuracy, precision, sensitivity, specificity, F-score, and area under the curve were the six metrics utilized to assess the outcomes. A binary classification model's accuracy is measured by the standard F-score, which is created by dividing the product of precision and sensitivity by two. The 3-factor subset, 5-factor subset, and entire dataset are classified or predicted using the J48 decision tree, which is an implementation of the ID3 (Iterative Dichotomiser 3) decision tree technique in R. A random forest model improves classification accuracy over a single tree model by using random groups of attributes. The "naive" assumption that all predictor variables are independent and will not affect one another underlies the oversimplified Naive Bayes model. An essential performance metric for classification models, the receiver operating characteristics (ROC) curve and resulting area under the curve (AUC) describe the degree of class separability. For accuracy, precision, specificity, f-score, and AUC on the total dataset, the random forest classifier outperformed the Naive Bayes and J48 decision tree models, with the J48 having the highest sensitivity. However, the 3-factor and 5-factor data subsets showed more excellent performance for the Naive Bayes model. Although the models used in this experiment are over 80

In [10], authors suggested a hybrid diabetes prediction approach that combines the ensemble method with bagging and 10 Folds cross-validation Techniques to predict diabetes using machine learning algorithms. As classification approaches utilize ensemble methods, the author has used Nave Bayes, J48, Support Vector Machine, RF, and logistic regression in this work. This particular machine learning probabilistic classifier is entirely founded on the Bayes Theorem. It is praised for its efficacy and simplicity. J48 is a selection tree classifier that builds the selection tree from labelled educational records using the facts entropy. It can help with both continuous and specific phases of tree formation. A type of statistical regression known as a logistic regression model employs one or more unbiased variables to forecast a dependent variable. It has many applications in social science, bioinformatics, and health. Given a two-elegance schooling pattern, a guiding vector device is employed to determine the best, highest-margin separation. The decision is made some distance from the historical causes in each category. The method for predicting diabetes uses machine learning algorithms that combine bagging, ensemble learning, and ten-fold cross-validation. The ensemble method's accuracy for the dataset is 77.60 percent, the highest among the others but not significantly better than the outcomes from using Random Forest.

Another recent study described in [11] used a rule-based classifier crucial in today's diabetes diagnosis. A method of data reduction known as Principal Component Analysis (PCA) offers fresh features that sharply differentiate between groups. PCA has been widely utilized in machine learning to identify and eliminate duplicate variables and can be used for data compression. The most effective linear classifier is probably a support vector machine (SVM). The input space is divided into two areas using a hyperplane, and because of the more significant margin, it has the best generalization properties. When given a complete data set as input, greedy divide-and-conquer Decision Tree algorithms look for a split that maximizes the "separation" of the classes. The naive Bayes algorithm counts the number of observations to identify patterns or relationships between data. It then develops a model that can be used to forecast various goals. The rule-based system's classification accuracy in diabetes data sets without applying PCA between the three classification algorithms is 76 percent, 75 percent, and 68 percent, respectively. The bar graph demonstrates that, compared to the decision tree and SVM classifiers without PCA, the accuracy of the Naive Bayes classifier increased by 1.36 percent.

Our main goal is to analyze the effectiveness of traditional machine-learning algorithms for diabetic prediction. We will use ID3, naive Bayes, and Random Forest in conjunction with traditional machine-learning techniques. So, to create a decision support system, we chose these algorithms and tested how well they detected diabetes. We will conduct numerous experiments on our data collection. An equal number of train and test samples for each approach will be chosen. Python 3.6 and the Anaconda 5 environment are used for the analysis and visualization.

II. The methodology

In this research, we employed four different algorithms to fit our data: decision tree(ID3), Random forest,Navie Bayes; comparing machine learning models to the performance of these models using various subsets of various variables. Additionally, contrasting the models enables testing to determine which model has the most accuracy in outcome prediction and better fits the data. This project focuses on indicators and factors (such as pregnancy and blood pressure) that have a high likelihood of predicting Type II Diabetes in women. We want to find the warning signs so women can take the appropriate preventative measures before Type II Diabetes develops. As a result, the characteristics of the Pima Indians dataset are described in this section. It exemplifies the many

approaches taken to forecast the onset of Type II diabetes in female Pima Indians. The mean imputation method for filling in the dataset's missing values are also covered in this section.

Description of Pima Indians Dataset Before Preprocessing: The National Institute of Diabetes, Digestive, and Kidney Diseases sponsored and published the dataset, which is known as the Pima Indian Diabetes database. Native Americans from North America called Pima lived in Arizona's Salt and Gila rivers historically. The open-source dataset, which contains patient records for female patients, is accessible to the general public on the Kaggle website (https://www.kaggle.com/uciml/pima-indians-diabetes-database) . The dataset contains 768 cases, each of which represents a participant who is a female Pima Indian. There is a binary indicator for each case: non-diabetic (0) and diabetes (1). The dataset includes 268 diabetes cases and 500 non-diabetic cases. The dataset also includes eight distinct difference factors (which we refer to as features), which could be predictors of our binary dependent variable (diabetes or non-diabetes). The dataset has the following characteristics:

Pregnancies This variable shows how many times a Pima Indian woman became pregnant. The dataset's range is 0 to 17, and the average value was 3.84.

Glucose Level: The body's reaction to sugar (glucose) is measured by the glucose tolerance test, commonly known as the oral glucose tolerance test. An oral glucose tolerance test measures plasma glucose concentration over two hours. The Pima Indians Dataset scores range from 0 to 199, with 0 denoting a missing value. It has a 120.89 average.

Blood pressure: The force that propels blood through the circulatory system is blood pressure. Extreme blood pressure variations may signify impending death, and both high and low blood pressure can have serious repercussions. The dataset's diastolic blood pressure metric is ((mm Hg)). The dataset's range is 0 to 122, with 0 denoting a missing value. The standard value is 69.10.

Skin Thickness: The measure used in the dataset for this variable is triceps skinfold thickness (mm). This measurement ranges from 0 to 99, where 0 denotes a missing value, with a mean of 20.53. "Skin Thickness" refers to the thickness of the triceps skinfold. It accurately estimates obesity and body fat distribution.

Insulin: We can determine if a person has a metabolic disease and whether islet function is deficient, both of which are associated with diabetes, based on their insulin levels after eating. In this dataset, "Insulin" refers to 2-hour serum insulin. The Pima Indians' unit of measurement for the two-hour blood insulin level is mu U/ml. The primary anabolic hormone in the body, Insulin, is a peptide hormone made by beta cells of the pancreatic islets. By facilitating glucose absorption from the circulation into the liver, fat, and skeletal muscle cells, it controls the metabolism of carbs, lipids, and protein . The insulin data set's range is 0 to 846, with the mean being 79.79 and 0 denoting a missing value.

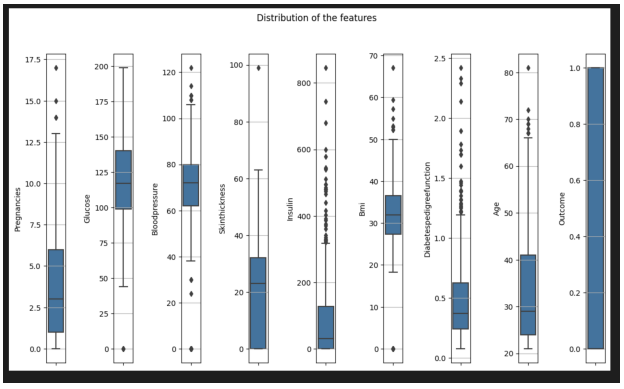
BMI: In statistical analysis, the body mass index (BMI), which measures obesity and health, is frequently utilized. The degree of obesity is correlated with height and cannot be determined solely by the weight's absolute value. BMI is calculated by dividing the body mass by the square of the body height. The dataset's BMI calculation formula is (weight in kg/(height in m)*2). The Pima Indians dataset's BMI ranges from 0 to 67.10, with 0 denoting a missing value. BMI average is 32.00.

Diabetes Pedigree Function: The Pima Indian dataset contains a variable named DBF, whose scores range from 0.07 to 2.42, with an average of 0.47. Based on family history, the DBF variable predicts the risk of developing diabetes. 3.1.8 age: Age (years) (years) The dataset ranges from 21 to 81. The median age is 33. 3.1.9 Outcome: Classification variable, where 0 denotes the absence of Type II diabetes in females and 1 denotes the condition in the individual.

Table 1: Before preprocessing, a Summary of Statistics Regarding the Pima Indians' Variable

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DBF	Age
count	768.00	768.00	768.00	768.00	768.00	768.00	768.00	768.00
mean	3.84	120.89	69.10	20.53	79.80	32.00	0.47	33.24
std	3.36	31.97	19.35	15.95	115.24	7.88	0.33	11.76
min	0.00	0.00	0.00	0.00	0.00	0.00	0.07	21.00
25%	1.00	99.00	62.00	0.00	0.00	27.30	0.24	24.00
50%	3.00	117.00	72.00	23.00	30.50	32.00	0.37	29.00
75%	6.00	140.25	80.00	32.00	127.25	36.60	0.62	41.00
max	17.00	199.00	122.00	99.00	846.00	67.10	2.42	81.00

Distribution of Features



Data Preprocessing One of the most common data mining tasks is data pretreatment. It entails utilizing machine learning techniques to prepare and transform data into a format appropriate for data mining. Reducing the size of the dataset, identifying relationships between and among the dataset's features, standardizing their values, weeing out outliers, and extracting features for additional data processing and analysis are the goals of data preparation. It incorporates several methods, including data integration, transformation, and reduction. Preprocessing was carried out on the dataset

using the libraries from Anaconda, Python NumPy, and Pandas.

Missing Values: As shown in Table 1: Summary of Statistics About the Pima Indians' Variable Before Preprocessing, it was discovered when the dataset was carefully examined using the function "Is null" of the Python library Pandas and the dataset contained a large number of missing values and many values with zeros. missing values were replaced with mean values by mean imputation.

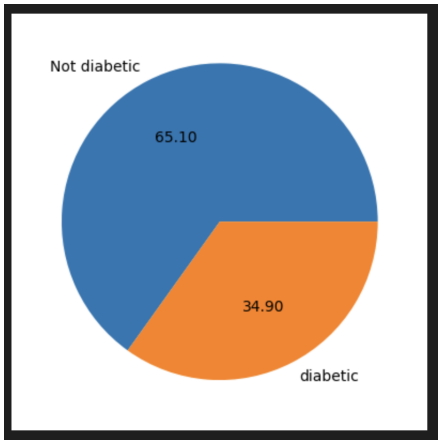
Scaling the Data: One of the essential phases in data preparation before fitting the machine learning model is feature scaling in machine learning. A machine learning model's strength can be changed through scaling, from poor to better. Normalization and standardization are the two methods of feature scaling that are most frequently used. Our values are restricted by normalization between two numbers, usually between [0,1] and [-1,1]. Our data become unitless through normalization, which changes the data to have a mean of 0 and a variance of 1. The data in this experiment was scaled using the Python method StandardScaler() of the Sklearn package . By subtracting the mean and scaling it to unit variance, StandardScaler() standardizes all predictor factors (glucose, blood pressure, skin thickness, Insulin, BMI, diabetes pedigree, pregnancy, and age). Scaler produces a distribution with a standard deviation of one when all the values are divided by the unit variance, defined as "a measure of the amount of variation or dispersion of a group of values". Last, StandardScaler sets the distribution's mean to approximately 0 and its standard deviation to 1. This amounts to rescaling the data using z scores.

Balancing the dataset by under-sampling technique: The "Outcome" (or dependent variable) in the Pima Indians dataset is a binary variable, where one (1) as "outcome" denotes a Pima Indian female having diabetes and zero (0) denotes the female not having diabetes. In this dataset, there are 268 Pima women who "outcome" (1) as having diabetes, compared to 500 women who "outcome" (0) as not having diabetes or nearly twice as many non-diabetic women. Consequently, we concluded that the Pima Indian dataset is uneven and may require balancing in some applications of our ML models.

When the outcome or target variable has more occurrences or observations in one class than the others, the dataset is unbalanced (as shown in Figure 2). A "major" class is one with a more significant number of instances (in our case, the "outcome" 0-not having diabetes), whereas a "minor" class is one with a lesser number of instances (in our case, the "outcome" 1-having diabetes).

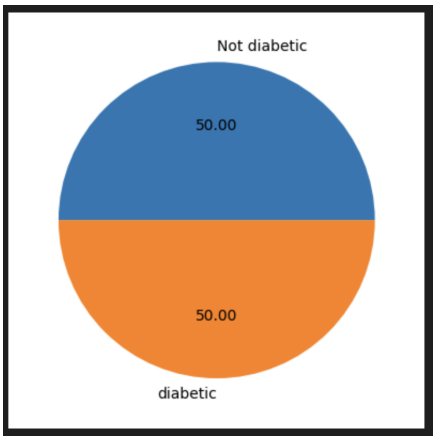
Most classifiers/models in an unbalanced dataset are biased toward the significant class and exhibit meager classification rates for minor classes. Furthermore, it is also feasible that the classifier treats every category as the main class while ignoring the minor categories. Different approaches have been suggested to address the issues brought on by class unbalance. Over-sampling and under-sampling are two of these numerous well-known methods. Under-sampling is the practice of omitting records from the main class. Adding records to the minor class is another result of oversampling.

Distribution of Outcome Feature(imbalanced)



As was already noted, a dataset that needs to be balanced might lead to inaccurate predictions from the various machine-learning techniques used to analyze the data. In order to correct the imbalance issue and prevent predicting the primary class, which is not having diabetes, an under-sampling strategy was utilized in this study for this dataset. Consequently, both Outcome classes had 268 records after using the under-sampling method.

Distribution of Outcome Feature(Balancing with imblearn library UnderSampling)



Selection of the Relevant Feature: When creating a predictive model, feature selection often aims to minimize the number of features. A classifier can lower the computational expense of modeling and achieve optimal performance by using this feature reduction approach. In this research, we employed statistical correlations to identify the key features that might contribute to ML modeling and achieve the best

model accuracy.

Correlation

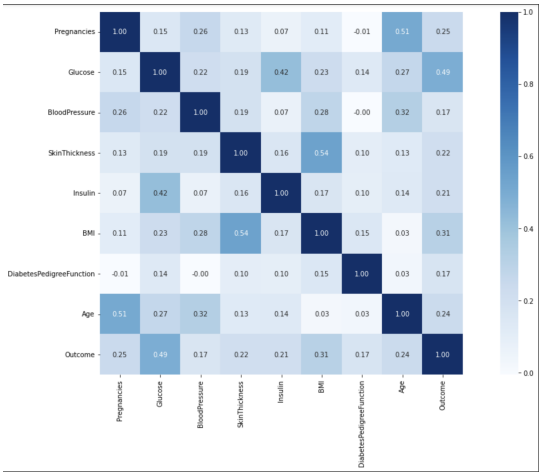


Figure 3: Feature Selection Correlation Heatmap

Since we can predict the value of one variable with the aid of other correlated qualities, correlation (r), which is used to evaluate the strength of the relationship between two features in statistics, is crucial in real life. Due to the presence of two features, it falls within the category of a bivariate statistic. A correlation matrix is a table that contains all of the bivariate or zero-order correlations that exist between and among the dataset's attributes. Features that are highly connected have correlation coefficients between 0.9 and 1.0. (influential association). The two properties are strongly correlated when the correlation coefficients are between 0.7 and 0.9 in magnitude. Correlation coefficients indicate features strongly correlated with magnitudes between 0.5 and 0.7. Correlation coefficients indicate low correlation features between 0.3 and 0.5 in magnitude. Less than 0.2 correlation coefficients have a weak (linear) association and may not be very helpful in predicting significant outcome factors.

The associations among and between all characteristics associated with Pima Indian diabetes are shown in Figure 3's Heatmap plot. We can see from the Heatmap that there is less than a 0.2 correlation between the independent parameters Blood pressure and DBF and the target variable "Outcome." As we have already discussed, a magnitude less than 0.2 represents a low association with the outcome. We consequently removed these two features from our primary dataset. We concluded that the most pertinent features for our ML classifiers to be trained on were: Glucose, Pregnancy, Age, Insulin, Skin Thickness, and BMI.

Split of the data: Data splitting divides the available data into two groups: the Training set is used to create a predictive model, and the Testing set is used to assess the model's performance. Thanks to the evaluation's findings, we can evaluate the effectiveness of various algorithms for solving problems involving predictive modeling. Therefore, in this project, the data was divided into the training set and testing set by a ratio of 80 percent and 20 percent using Train Test Split() of the sci-kit-learn Python package.

Algorithms for Prediction of Diabetes: Models are developed using supervised machine learning that explicitly relates the inputs (independent variables, features, or predictors) to the outputs (Outcome, dependent variable, or target). We processed to fit the ML models after the preprocessing stage and the training/testing sets split were completed. As a result, this section goes through the various supervised learning methods selected for this study to categorize the participants as having diabetes or not. Machine learning algorithms are required for the classification task since they must be taught how to assign a class label to a particular event. Classifying Pima Indian women as "Diabetic" or "Not Diabetic" frequently occurs in our project. There are many different techniques, and binary classification is one of them (the kind we use in this project). A classification problem known as binary classification uses two class labels (0 or 1, True or false, Being diabetic or not). Given that they are the most widely used ML algorithms for binary classification, we have selected Decision tree,Navie bayes, Random Forest to forecast our feature.

Decision Tree(Iterative Dichotomiser 3)

With no backtracking, the top-down greedy search method of the ID3 algorithm creates decision trees by traversing the space of potential branches. As the name implies, a greedy algorithm always selects the option that, at the time, appears to be the best.

ID3 Algorithm implementation

1. Choose the root node(S)or nodes depending on the lowest entropy and most significant information gain

Information Gain is an impurity-based criteria. Entropy (information theory) is the impurity metric in information Gain.

$$\text{Entropy(y,S)} = - \sum \frac{|\sigma_{y=c_j} S|}{|S|} \log_2 \frac{|\sigma_{y=c_j} S|}{|S|}$$

2. Considering that every node is unused, an algorithm calculates the Entropy and Information gain on each iteration.
3. Choose a node based on either the highest information gain or the lowest entropy.
4. Splits set S to create the data subsets.
5. A decision tree is created using an algorithm that continuously iterates through each subset to ensure

Random Forest

The supervised learning approach is used by the well-known machine learning algorithm Random Forest. It can be applied to classification and regression problems in machine learning. It is based on ensemble learning, a technique for combining various classifiers to solve a challenging problem and improve the performance of the model. According to its name, "Random Forest is a classifier that contains numerous decision trees on different subsets of a given dataset and takes the average to enhance the predicted accuracy of that dataset."

The random forest gathers the predictions from each decision tree and predicts the final result based on the majority votes of predictions, as opposed to relying solely on one decision tree. A bootstrapping aggregation or bagging approach is the training algorithm employed by RF . They combine various tree predictors results in random forests. The values of a random vector sampled random and with the same distribution across all of the forest's trees determine the values of each tree. The "class" forecast made by each tree in the random forest becomes the model's prediction based on whatever class receives the most votes.

Naive bayes

Based on the Bayes theorem, the Naive Bayes method is a approach to supervised learning that deals with classificatio problems. The most frequent application of it is in tex categorization with a big training set. In order to create quick machine learning models that are capable of making precise predictions, the Naive Bayes Classifier is a simple and efficient classification technique. Being a probabilistic classifier, it bases its predictions on how likely the object is to occur. Bayes's theorem The Bayes theorem, also referred to as Bayes' rule or Bayes' law, is a mathematical formula for estimating the likelihood of a hypothesis given available data. The factor that decides is conditional probability. Bayes' theorem's formulation is as follows: P(A/B) in this case Probability of a hypothesis A being true in relation to an event B that has been observed. classifier is utilized for classification task.

Implementation :

We trained the decision tree, Naive bayes and random forest models with train set and tested models with a test set , we drawn a confusion matrix, Accuracy score , and classification report for each model.

For the decision tree model, we got an Accuracy Score of 67.5925

confusion matrix

```
from sklearn.metrics import confusion_matrix
confMat = confusion_matrix(y_test, OutcomePred)
confMat
```

```
array([[33, 21],
       [14, 40]])
```

Classification report

```
from sklearn.metrics import classification_report
reportDT=classification_report(y_test,OutcomePred)
print(reportDT)
```

```
              precision    recall  f1-score   support

           0       0.70      0.61      0.65        54
           1       0.66      0.74      0.70        54

    accuracy                           0.68       108
   macro avg       0.68      0.68      0.67       108
weighted avg       0.68      0.68      0.67       108
```

For the Random Forest model, we got an Accuracy score of 75.92592592592592

confusion matrix

```
from sklearn.metrics import confusion_matrix
confMatRM=confusion_matrix(y_test,OutcomeRFRM)
confMatRM
```

Python

```
array([[37, 17],
       [ 9, 45]])
```

Classification report

```
from sklearn.metrics import classification_report
reportRM=classification_report(y_test,OutcomeRFRM)
print(reportRM)
```

```
              precision    recall  f1-score   support

           0       0.80      0.69      0.74        54
           1       0.73      0.83      0.78        54

    accuracy                           0.76       108
   macro avg       0.77      0.76      0.76       108
weighted avg       0.77      0.76      0.76       108
```

For the Naive bayes model, we got an Accuracy score of 77.77777777777779

confusion matrix

```
from sklearn.metrics import confusion_matrix
confMatNB=confusion_matrix(y_test,OutcomeGTNB)
confMatNB
```

Python

```
array([[41, 13],
       [11, 43]])
```

Classification report

```
from sklearn.metrics import classification_report
reportsclassification_report(y_test,OutcomeGTNB)
print(report)
```

```
              precision    recall  f1-score   support

           0       0.79      0.76      0.77        54
           1       0.77      0.80      0.78        54

    accuracy                           0.78       108
   macro avg       0.78      0.78      0.78       108
weighted avg       0.78      0.78      0.78       108
```

III. The results

After implementing All three models with balanced data (with under-sampling technique) (Decision tree, Random forest, Naive Bayes), we can see that with the decision tree, we got an accuracy score of 67 percent. We got an Accuracy score of 75 percent with the Random Forest model. With the Naive Bayes model, we got an Accuracy score of 77. The highest accuracy score was achieved by the Naive Bayes model.

IV. Conclusion

The heart, blood arteries, nerves, eyes, kidneys, and many other important organs are among those affected by type II diabetes, which is frequently referred to as a lifestyle disease. The goal of this project was to identify risk factors and indicators for Type II Diabetes in women so that those women might take the required precautions to delay the disease's emergence. Using the female Pima Indians' diabetes dataset, we predicted diabetes in this study. To do the result prediction, we have chosen the features of glucose, pregnancies, age, insulin, and BMI. The greatest performance has been assessed using a variety of supervised learning algorithms, including decision trees, random forest, and Naive Bayes.

Since the dataset only had 798 records, it would be wise to collect more information if this research were to be expanded. We should also compile men's records, too. Additionally, we would include extra features, such as daily nutrition and exercise logs. The ability of both men and women to take the required precautions to stop the emergence of Type II diabetes may be improved by having access to more and various types of data.

REFERENCES

[1] S. K. Reddy, T. Krishnaveni, G. Nikitha, and E. Vijaykanth, "Diabetes prediction using different machine learning algorithms," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 1261–1265, IEEE, 2021.

[2] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," in *2018 4th International conference on computing communication and automation (ICCCA)*, pp. 1–4, IEEE, 2018.

[3] S. Sakib, N. Yasmin, I. K. Tasawar, A. Aziz, M. A. B. Siddique, and M. M. R. Khan, "Performance analysis of machine learning approaches in diabetes prediction," in *2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC)*, pp. 1–6, IEEE.

[4] S. Ghane, N. Bhorade, N. Chitre, B. Poyekar, R. Mote, and P. Topale, "Diabetes prediction using feature extraction and machine learning models," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1652–1657, IEEE, 2021.

[5] M. Pal, S. Parija, and G. Panda, "Improved prediction of diabetes mellitus using machine learning based approach," in *2021 2nd International Conference on Range Technology (ICORT)*, pp. 1–6, IEEE, 2021.

[6] G. Tripathi and R. Kumar, "Early prediction of diabetes mellitus using machine learning," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pp. 1009–1014, IEEE, 2020.

[7] R. Barhate and P. Kulkarni, "Analysis of classifiers for prediction of type ii diabetes mellitus," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1–6, IEEE, 2018.

[8] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A decision support system for diabetes prediction using machine learning and deep learning techniques," in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, pp. 1–4, IEEE, 2019.

[9] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima indians diabetes mellitus classification based on machine learning (ml) algorithms," *Neural Computing and Applications*, pp. 1–17, 2022.

[10] P. Goyal and S. Jain, "Prediction of type-2 diabetes using classification and ensemble method approach," in *2022 International Mobile and Embedded Technology Conference (MECON)*, pp. 658–665, IEEE, 2022.

[11] A. Thammi Reddy and M. Nagendra, "Minimal rule-based classifiers using pca on pima-indians-diabetes-dataset," *Int. J. Innov. Technol. Explor. Eng*, vol. 8, no. 12, pp. 4414–4420, 2019.