6-2022

A Machine Learning Approach to Predicting the Onset of Type II Diabetes in a Sample of Pima Indian Women

Meriem Benarbia
The Graduate Center, City University of New York

A MACHINE LEARNING APPROACH TO PREDICTING THE ONSET OF TYPE II DIABETES IN A SAMPLE OF PIMA INDIAN WOMEN.

By

MERIEM BENARBIA

Master's Capstone Project submitted to the Graduate Faculty in Data Analysis and Visualization in partial fulfillment of the requirements for the degree of Master of Science, The City University of New York.

2022

A Machine Learning Approach to Predicting the Onset of Type II Diabetes in a Sample of Pima Indian Women

by

MERIEM BENARBIA

This manuscript has been read and accepted for the Graduate Faculty in Data Analysis and Visualization in satisfaction of the capstone project requirement for the degree of Master of Science.

Date	**Professor Howard Everson** **Capstone Advisor**
Date	**Professor Matthew Gold** **Program Director**

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

A Machine Learning Approach to Predicting the Onset of Type II Diabetes in a Sample of Pima Indian Women.

by

MERIEM BENARBIA

Advisor: Professor Howard T. Everson.

Type II diabetes is a disease that affects how the body regulates and uses sugar (glucose) as a fuel. This chronic disease results in too much sugar circulating in the bloodstream. High blood sugar levels can lead to circulatory, nervous, and immune systems disorders. Machine learning (ML) techniques have proven their-strength in diabetes diagnosis. In this paper we aimed to contribute to the literature on the use of ML methods by examining the value of a number of supervised machine learning algorithms such as logistic regression, decision tree classifiers, random forest classifiers, and support vector classifiers to identify factors and indicators (such as pregnancy, blood pressure, etc.) that may lead to more accurate predictions and classifications of Type II diabetes in women. By identifying these indicators, women will be able to take the necessary actions to prevent the onset of Type II diabetes. To apply these ML techniques ,the Pima Indian Women Diabetes dataset was downloaded from the Kaggle website. Different experiments were conducted on the dataset. Each machine learning algorithm was trained on unscaled data using a balanced and unbalanced dataset and trained again using scaled data with balanced and unbalanced dataset. Consequently, sixteen models were generated to evaluate the different ML classifiers' performance and select the best model. The results of these analyses are presented, and model-based findings are contrasted.

ACKNOWLEDGMENTS:

I would like to thank Professor Howard Everson for his patience, guidance, support, and all the information I have learned while studying in his class and conducting this capstone project. Also, I would like to thank him for his caring about the achievement of this project. Professor, you are an outstanding advisor.

I am thankful to Professor Matt Gold for making time to listen, and for all his advice.

I am also so grateful to Jason Nielsen for his support and presence throughout the program.

I want to thank Professor Johanna Devaney, whose Methods in advanced analysis class helped me conduct this capstone project.

Lastly, I want to thank my beloved Papa and Mama, my Sisters, Assia and Asma, and my Brother Tewfik for their daily support; without their encouragement and faith in me, I would not achieve my studies.

Table of Contents

List of Tables

List of Figures

Digital Manifest

I. PDF of Capstone Project White Paper

II. Git Repository zip file containing the Pima-Indians-diabetes Dataset, named Pima.csv, and all the analysis project Jupyter notebook

https://github.com/Atini2525/Data-Analysis-Capstone-Spring-2022

Note on Technical Specifications

The project's analyses were run using Python 3 and the Jupyter notebook architecture. It is simple to download Anaconda (Anaconda3), allowing easy software and package management. As an alternative to Jupyter Notebook, Jupyter may also be used. The following Python 3 packages were installed, to collect the data and run the models:

- Warnings
- NumPy
- Pandas
- Matplotlib
- Seaborn
- Sklearn
- Bioinfokit

1. **Introduction**:

Type II diabetes is often referred to as a lifestyle disease —a form of diabetes characterized by high blood sugar levels, a resistance to insulin, and a relative lack of insulin [1]. Common symptoms include increased thirst, frequent urination, and unexplained weight loss. Symptoms may also include increased hunger, tiredness, and bodily sores that do not heal [2]. Unfortunately, more than 34 million Americans have diabetes, and 90-95% of them have Type II diabetes. Type II diabetes most often develops in people over age 45, but more and more children, teens, and young adults are also diagnosed with diabetes [3].

Type II diabetes affects many major organs, including the heart, blood vessels, nerves, eyes, and kidneys. Also, it increases the risk of Alzheimer's disease and other disorders that cause dementia[4]. According to the Centers for Disease Control and Prevention (CDC), men are more likely to receive a diagnosis of diabetes than women. However, some research suggests that women with diabetes may be more likely to develop complications than men [3].

I was partially motivated to conduct these analyses to identify trends that could potentially help prevent further complications of diabetes because my grandmother lost her vision to this terrible disease. Additionally, my interest was reinforced by the different courses I took as a graduate student in the Data Analysis and Visualization Program at the CUNY Graduate Center, such as an Introduction to Machine Learning and Advanced Data Analysis Methods. In these courses, I learned how to use machine learning methods and statistical models for different classifications and predictions in the machine learning course. The data analysis methods course taught me how to select the right features that can positively impact my

prediction results. Both courses gave me the knowledge and the statistical foundation needed to conduct this project.

In this project, we investigated possible predictors of the onset of Type II diabetes in a sample of Pima Indian women using machine learning statistical methods. The dataset used in this study is the Pima Indians' Women Diabetes database sponsored by the National Institute of Diabetes, Digestive, and kidney diseases. The dataset is publicly available on the Kaggle website[5].

This database was chosen to build and test several machine learning models and to investigate whether the models would help predict the onset of Type II diabetes in women more generally. Many previous analyses were done using this sample of Pima Indian women. Therefore, we wanted to compare a select number of ML models with the models presented in these other studies. The details of our approach are discussed later in this paper

2. Prior Work:

There have been several studies using the publicly available Pima Indians dataset attempting to obtain a data driven classifier system that improves the diagnosis of Type II diabetes, supporting diabetic diagnosis. We chose two of those studies that had a high accuracy in predicting Type 2 Diabetes in the Pima Indian women to compare our analysis with and to determine if our research and the methodology we used produced more accurate classifications.

The first study was conducted by Kumar Bhai, S & Anshuman Abhishek, P. (2021, April. 28) [6]. This study is available at Turkish Journal of Computer website. Kumar & Anshuman generated different classification algorithms using all the features of Pima Indians dataset such as glucose level, blood pressure, skin thickness, insulin level, the body mass index (BMI), diabetes pedigree, pregnancy, and age. Here is a list of all the algorithms tested in that

research: Tree classifiers, Support Vector Machine, K-Nearest Neighbors, Naïve Bayes, Random Forest, Neural Network, Ad boost, Logistic Regression. The researchers did scale the data but did not balance the data (more details about the balanced and unbalanced dataset and their meaning to the analytics are presented in section 3.3).The researchers concluded that logistic regression was the modeling method that provided the highest prediction accuracy.

The second study was conducted by Thammi Reddy, A & Nagendra, M. (2019, Oct. 12) [7] . and it is publicly available at Semantic Scholar website. Thammi & Nagendra generated three classifications models to predict diabetes using Principal Component Analysis (PCA). Here is a list of all the algorithms tested in that research: Naïve Bayes, Decision Tree and SVM. The three models used all features of the Pima Indians' dataset. The researchers did scale the data but did not balance the dataset(more details about scaling the data are discussed in the section 3.2.3 Scaling the Data:). They concluded that the Naïve Bayes [8] model provided the highest accuracy.

These prior studies had a high accuracy in predicting diabetes in the Pima Indian women. Both studies conducted an analysis on scaled data and unbalanced datasets, with all the feature of the dataset. Thus, we decided to conduct our analysis on the selected feature (more details of the selected features are discussed in the section 3.4. Selection of the Relevant Feature:) with scaled and unscaled data and on both unbalanced and balanced datasets, to assess whether we can get better accuracy in predicting Type II Diabetes in the female population of the Pima Indian dataset.

3. Methodology:

In this project, we used fours algorithms to fits our data: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine "SVM" because comparing the machine learning

models is a significant and essential task since we need to compare the performance of these models using different subsets of different features. Moreover, comparing the models allows for testing which model fits the data well and delivers the best accuracy in predicting the outcome. Furthermore, there is a comparison within the same model besides comparing different models. Often data scientists compare the same model with different parameters (In our project, we ran the same model on scaled and unscaled data) to evaluate which parameters allow the model to have the best performance. In addition to comparing the different ML models, we compared the same model with different parameters.

Lastly, we ran the same model on scaled and unscaled data to test which type of data better fits our outcome and The same algorithm was run on balanced and unbalanced datasets to test if the difference in the number of observations improved the accuracy. In short, we fit all the algorithms used in this study : logistic regression, random forest, decision tree, and support vector machine on four different versions of the data as shown below:

- Unbalanced dataset and Unscaled data.
- Unbalanced dataset and Scaled data.
- Balanced dataset and Unscaled data.
- Balanced dataset and Scaled data.

This project aims to identify factors and indicators (such as pregnancy, blood pressure, etc.) That has a high risk of predicting Type II Diabetes in women. Therefore, this section describes the features of the Pima Indians dataset. It illustrates the several methodologies used to predict the onset of Type II diabetes in females of Pima Indians heritage. This section also describes the quartile approach used to eliminate the outliers and the KNN (K-Nearest Neighbors)approach used to fill the missing values in the dataset. Also, it describes the

correlational analyses used to select the most relevant and essential features for use in the ML models. Moreover, in this section, we fit the data using the most popular machine learning models in classification: Logistic regression, SVM (Support Vector Machine), decision tree models, and random forest to find the optimal model. We hope to identify the indicators that would allow women to take the necessary actions to prevent the onset of Type II Diabetes.

3.1 Description of Pima Indians Dataset Before Preprocessing :

The dataset used, called the Pima Indian Diabetes database, was sponsored and published by the National Institute of Diabetes, Digestive, and Kidney Diseases. Pima Indians are 'North American Indians who traditionally lived along the Gila and Salt rivers in Arizona.

The dataset is publicly available on the Kaggle website (https://www.kaggle.com/uciml/pima-indians-diabetes-database) [5]; it is an open-source dataset consisting of records of female patients. The dataset has 768 cases, where each case represents a female Pima Indian participant. Each case contains a binary indicator—non-diabetic (0) and diabetic (1). The dataset contains 500 non-diabetic cases and 268 with a diabetic classification. In addition, the dataset contains eight individual difference variables (that we call features), which may serve as predictors of our binary dependent variable (diabetes or non-diabetes). The dataset includes the following features:

3.1.1 Pregnancies:

This variable illustrates the number of times a Pima Indian female got pregnant. The range is from 0 to 17 in the dataset, and the average was 3.84.

3.1.2 Glucose Level:

It measures the plasma glucose concentration over 2 hours in an oral glucose tolerance test. Scores range in the Pima Indians Dataset is from 0 to 199, where 0 means

signifies a missing value (more details of the missing value are discussed in the section 3.2.1. Missing Values:). Its average is 120.89. The glucose tolerance test, also known as the oral glucose tolerance test, measures the body's response to sugar (glucose)[9].

3.1.3 Blood pressure:

Blood pressure is the force that moves blood through the circulatory system. Both high blood pressure and low blood pressure can have grave consequences, and severe changes in blood pressure may be a precursor to death [10]. The metric used in the dataset for diastolic blood pressure is ((mm Hg)). The range in the dataset is from 0 to 122 where 0 means a missing value . The average is 69.10.

3.1.4 Skin Thickness:

Triceps skinfold thickness, the metric used in the dataset for this variable is (mm). This measure ranges from 0 where 0 means a missing value to 99; the average is 20.53. The variable of ‘Skin Thickness mean triceps skinfold thickness. It provides a good estimate of obesity and body fat distribution [1].

3.1.5 Insulin:

The metric used in the Pima Indians to measure the two-hours serum insulin level (mu U/ml). The variable of 'Insulin ‘in this dataset means 2-hour serum insulin. Based on one's insulin levels after a meal, we can tell if there is a metabolic disorder and whether there is a defect in islet function which are related with diabetes.

Insulin is a peptide hormone produced by beta cells of the pancreatic islets, which is the main anabolic hormone of the body. It regulates the metabolism of carbohydrates, fats, and protein by promoting glucose absorption from the blood into the liver, fat, and skeletal muscle cells [12].

The range of the insulin in the dataset is from 0 to 846, where 0 means a missing value and the mean is 79.79.

3.1.6 BMI:

Body mass index (BMI), a measure of obesity and health, is commonly used in statistical analysis. The degree of obesity cannot be judged directly by the absolute value of the weight, and it is naturally related to height. So, BMI is defined as the body mass divided by the square of the body height. [13]. The formula for calculating BMI in the actual dataset is (weight in kg/(height in m)*2). The range of BMI in the Pima Indians dataset is from 0 to 67.10 , where 0 signifies a missing value. The mean of BMI is 32.00

3.1.7 Diabetes Pedigree Function:

This variable is called **DBF**, and its scores range in the Pima Indian dataset is from 0.07 to 2.42, and the average is 0.47. The DBF variable represents the likelihood of getting diabetes based on family history.

3.1.8 Age:

Age (years) the range in the dataset is from 21 to 81. The mean is 33 years

3.1.9 Outcome:

Classification variable where 0 means that a female does not have Type II diabetes, and a 1 indicates the participant has Type II diabetes.

Table 1: Summary of Statistics About the Pima Indians' Variable Before Preprocessing

	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

3.2 Data Preprocessing

Data preprocessing is one of the most data mining tasks. It includes preparing and transforming data into a suitable form for data mining using machine learning methods. The data preprocessing aims to reduce the size of the dataset, find relations between and among the features in the dataset, normalize their values, remove outliers, and extract features for further data processing and analysis. It includes several techniques like data cleaning, integration, data transformation, and data reduction [14]. To make the dataset more productive, preprocessing was done using Anaconda, Python NumPy, and Panda's libraries.

3.2.1. Missing Values:

When the dataset was closely analyzed using the function "Is null" of the library Pandas of python and the boxplot, it was detected that the dataset had many missing values and many with zeros values(as demonstrated in Table 1: Summary of Statistics About the Pima Indians' Variable Before Preprocessing).

General practitioners and doctors know that the Glucose, Insulin, BMI, and blood pressure range can never start from Zero. However, the dataset showed many missing and zeros records, which were replaced with estimated nearest neighbor values using the K-nearest neighbor (KNN) algorithm Configuration of KNN imputation often involves selecting the distance measure (e.g., Euclidean) and the number of contributing neighbors for each prediction, the k hyperparameter of the KNN algorithm. We decided to choose The KNN model to impute our missing because it was proven to be effective in many experiments.[15]

3.2.2 Outliers:

The interquartile range is often used to find outliers in data. Outliers are defined as observations that fall below Q1 − 1.5 IQR or above Q3 + 1.5 IQR. Using boxplots, the highest and lowest occurring values within this limit are indicated by the "whiskers" of the boxplot (frequently with an additional bar at the end of the whisker), and outliers are identified as individual points [16](as demonstrated in Figure 1: Boxplot of The Blood Pressure Feature.) Also, From the Table 1: Summary of Statistics About the Pima Indians' Variable Before Preprocessing, we could observe that our features had many outliers , an example the insulin had 846 U/ml as a maximum value which scientifically cannot be possible Therefore, the outliers were found and deleted from the dataset using the minimum and maximum values and the variable's upper and lower quartile values.

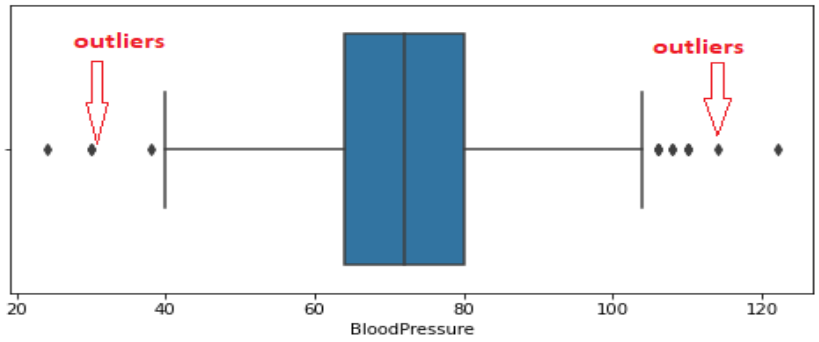

Figure 1: Boxplot of The Blood Pressure Feature.

3.2.3 Scaling the Data:

Feature scaling in machine learning is one of the most critical steps during the preprocessing of data before fitting the machine learning model. Scaling can make a difference between a weak machine learning model and a better one. The most common techniques of feature scaling are normalization and standardization. Normalization is used to bound our values between two numbers, typically between [0,1] or [-1,1]. While standardization transforms the data to have zero mean and a variance of 1, they make our data unitless[17].In this project, the function StandardScaler() of the Sklearn package in Python was used to scale the data [19]. The function StandardScaler() standardizes all the predictors' features (Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree, Pregnancy, and Age) by subtracting the mean, then scaling it to unit variance. Unit variance means dividing all the values by the standard deviation "is a measure of the amount of variation or dispersion of a set of values" [18]—scaler results in a distribution with a standard deviation equal to 1. Finally, StandardScaler makes the mean of the distribution approximately 0, and the standard deviation equal to 1. This amounts to using z scores to rescale the data.

3.3 Balancing the dataset by under-sampling technique:

In the Pima Indians dataset, the "Outcome" (or dependent variable) is a binary variable, where one (1) as "outcome" signifies that a Pima Indian female has diabetes and zero (0) indicates that the female does not have diabetes.

In this dataset, the number of diabetic Pima women "outcome" (1) is 268, while the number of those who do not have diabetes "outcome" (0) is 500, which is almost double that of non-diabetic females. As a result, we concluded that the Pima Indian dataset is unbalanced and may need to be balanced in some applications of our ML models.

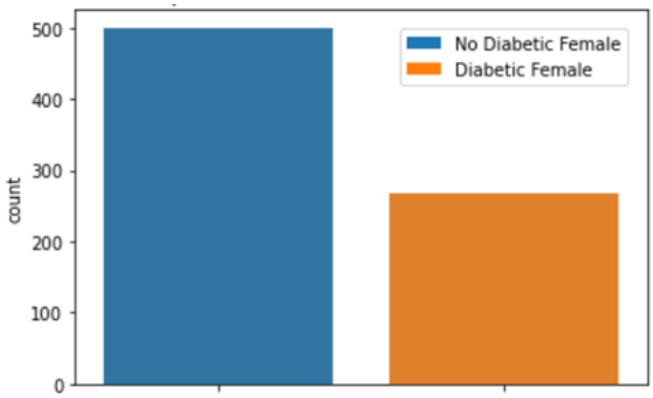

Figure 2: Frequency of Type II Diabetes in the Pima Indian Women Dataset.

An unbalanced dataset is when the outcome/target variable has more instances/observations in one specific class than the others(as shown in Figure 2). The class has a more considerable number of instances called a "major" class (in our case, the "outcome" 0-not having diabetes), while the one having a smaller number of instances is called a "minor" class (in our case, the "outcome" 1-having diabetes).

In an unbalanced dataset, most of the classifiers/models are biased towards the major class and show extremely poor classification rates for minor classes. Moreover, it is also possible that the classifier predicts everything as a major class and ignores the smaller, smaller class.

Various techniques have been proposed to solve the problems associated with class unbalance. Two of these different techniques are known: over-sampling and under-sampling. In simple terms, under-sampling means dropping records from the major class. And over-sampling means adding records to the minor class. [20].

As previously mentioned, an unbalanced dataset can be the origin of a poor prediction of the different machine learning methods applied to the data. Consequently, an under-sampling technique was used in this research for this dataset to redress the unbalance problem to avoid predicting the major class, which is not having diabetes. Consequently, both Outcome classes had 268 records after applying the under-sampling technique.

3.4. Selection of the Relevant Feature:

The goal of feature selection, in general, is to reduce the number of features when developing a predictive model. This feature reduction process allows a classifier to reduce the computational cost of modeling and reach optimal performance.

In this project, we used the statistical correlations to determine the critical features that may contribute meaningfully to the ML modeling and reach the optimal accuracy of the model.

3.4.1 Correlation

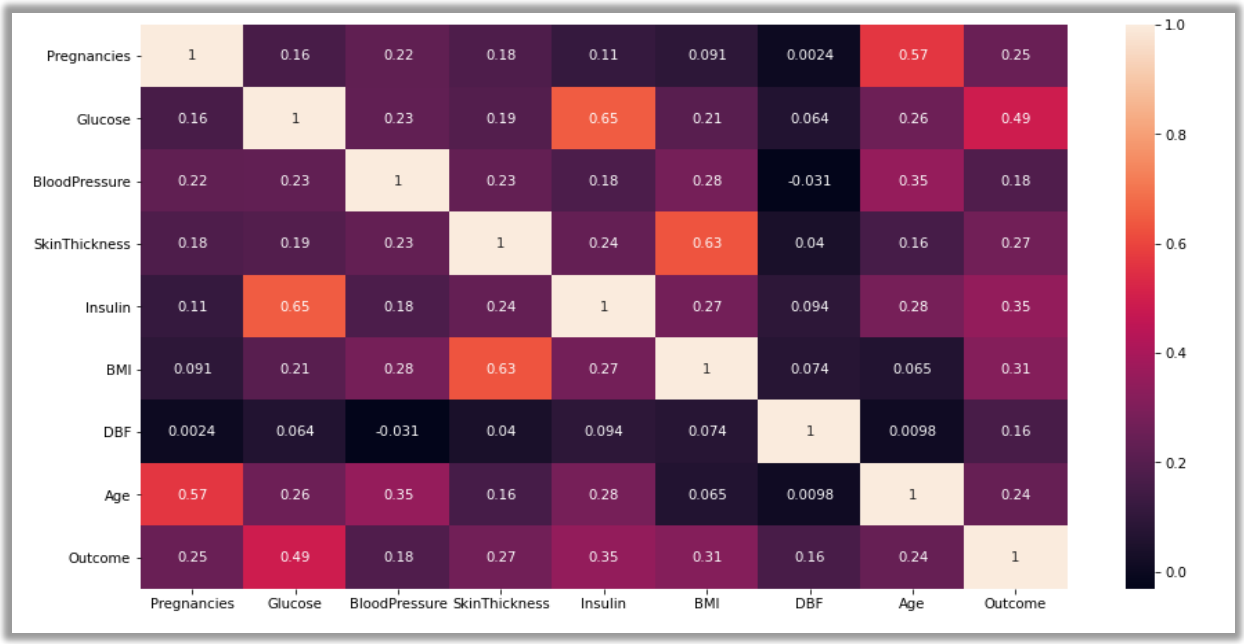

Figure 3:Correlation Heatmap for the Features Selection

In statistics, the correlation (r) is used to measure the strength of the relationship between two features which is essential in real life because we can predict the value of one variable with the help of other features, which is being correlated with it. It is a type of bivariate statistic because two features are involved here. A correlation matrix is a table of all the bivariate or zero-order correlations among and between the features in the dataset. [21]. Correlation coefficients with magnitudes between 0.9 and 1.0 indicate features are very highly correlated (very strong association). Correlation coefficients whose magnitude are between 0.7 and 0.9 indicate the two features are highly associated. Correlation coefficients with magnitudes between 0.5 and 0.7 indicate features that can be considered moderately correlated. Correlation coefficients with magnitudes between 0.3 and 0.5 indicate features with a low correlation. Correlation coefficients whose magnitude is less than 0.2 have a little (linear) correlation and may contribute little to the prediction of important outcome variables.

The Heatmap plot in Figure 3 displays the correlations among and between all features related to Pima Indian Diabetes. From the Heatmap, we can glimpse that the correlation magnitude between the target variable "Outcome" and the independent features: Blood pressure, and DBF is smaller than "0.2." As we have already discussed, a magnitude smaller than 0.2 is a low association with the outcome. Thus, we dropped these two features from our primary dataset. In short, the following features: Glucose, Pregnancies, Age, Insulin, Skin Thickness and BMI were determined to be the most relevant features on which our ML classifiers were trained

Table 2: Summary of Statistics of The Relevant Features After Preprocessing.

	Glucose	Pregnancies	Age	Insulin	BMI	SkinThickness
Count	644.00	644.0	644.0	644.00	644.0	644.0
Mean	119.76	03.90	33.13	141.49	31.88	28.34
std	29.55	03.33	11.74	073.21	06.34	08.61
Min	44.00	00.00	21.00	015.00	18.20	07.00
25%	99.00	01.00	24.00	087.45	27.30	22.70
50%	114.00	03.00	29.00	130.00	32.00	29.00
75%	138.00	06.00	40.25	181.85	35.80	33.45
Max	198.00	17.00	81.00	392.00	50.00	50.00

3.5. Split of the data:

Data splitting is partitioning available data into two portions: One portion of the data is used to develop a predictive model (Training set) and the other to evaluate the model's performance (Testing set)[22]. The results of the evaluation allow us to compare the performance of different algorithms for predictive modeling problems. Therefore, in this project, the function Train_Test_Split( ) of the library sci-kit-learn of Python was used to split the data into the training set and testing set by a ratio of 80% and 20%.

3.6. Algorithms for Prediction of Diabetes:

Since the preprocessing step and the training/testing sets split were done, we processed to fit the ML models. Therefore, this section discusses multiple supervised learning algorithms chosen in this project for classifying the participants with and without diabetes. Supervised machine learning creates models that specifically map the given inputs (independent variables, features, or predictors) to the given output (Outcome, dependent variable, or target).
A classification is a task that necessitates machine learning algorithms, which learn how to allocate a class label to a specific occurrence. In our project, an occurrence is classifying Pima Indian women as "*Diabetic"* or "*Not Diabetic*."

There are many diverse types of classification methods, one of these types (the one used in our project) is binary classification. Binary classification is a classification task with two class labels (0 or 1, True or false, Being diabetic or not, etc.). We have chosen the following ML algorithms to predict our feature since they are the most popular algorithms used for binary classification: Logistic Regression, Random Forest, Decision Trees, and Support Vector Machine. We have used these four algorithms to train our data, and we have created four versions (different parameters) of each algorithm to select the optimal model and parameters. The same algorithm was trained on an unbalanced dataset with scaled and unscaled data and on a balanced dataset with scaled and unscaled data. Consequently, sixteen (16) different models were created to discern the model that fit our dataset best and had the best performance score.

3.6.1 Logistic Regression:

This is a classification algorithm in machine learning that uses one or more independent features to determine an outcome. The outcome is measured with a dichotomous or binary variable, meaning it will have only two outcomes[23]. And, since our dependent variable

'Outcome' has only the values 0 and 1, logistic regression was the most straightforward method to use to train our dataset.

3.6.2 The Decision Tree Classifier:

A decision tree is a decision support tool that uses a tree-like model of decisions and their consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm containing only conditional control statements. Moreover, a decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g., whether a coin flip produces heads or tails), each branch represents the outcome of the test. Each leaf node represents a class label (a decision taken after computing all attributes, such as the example of whether a participant is diabetic or not.) The paths from the root to the leaf represent classification rules [24]. A decision tree can be used for both classification and regression problems. Since our dependent variable 'Outcome' is a classification variable, using the categorical decision tree method was one of the suitable algorithms to fit to the Pima Indians dataset.

3.6.3 Random Forest (RF):

Random forests are a combination of tree predictors. Each tree in the forest depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The training algorithm used by RF is a bootstrapping aggregation or bagging method [25]. Each tree in the random forest produces a “class” prediction, and the class with the most votes becomes the model's prediction[26].

3.6.4. Support Vector Machine:

The support vector machine "SVM" supports only the binary class classification. And since our dependent variable 'Outcome' has only the values 0 and 1, the SVM approach was a

suitable method for training our dataset. The SVM is a classifier that represents the training data as points in space separated into categories by a gap as wide as possible. New points are then added to space by predicting which category (class or outcome) they fall into and to which space they will belong [27].

4. Analysis and Results.

Evaluating a machine learning model is as critical as building it. Therefore, the following section discusses the different metrics used to evaluate our models and the analytics used to choose our best predicting model.

4.1. Evaluation of the Performance Measures:

Many metrics are used to evaluate the models. However, a confusion matrix is not a metric to evaluate a model, but it provides insight into the predictions. It is essential to calculate the confusion matrix to comprehend other classification metrics such as accuracy, precision, and recall. The confusion matrix goes deeper than classification accuracy by showing each class's correct and incorrect (i.e., true, or false) predictions[28].

Table 3: Confusion Matrix

	Predicted Positive	**Predictive Negative**
Actual Positive	TP	FP
Actual Negative	FN	TN

- **TP (True Positive):** The model predicted positive, and it is true. In our case, The model predicted that a Pima Indian woman would have diabetes, and she is diabetic.

- **TN (True negative):** The model predicted negative, and it is true.in our project, means The model predicted that a Pima Indian woman is not diabetic, and she is not diabetic.
- **FP (False positive):** The model predicted positive, and it is false. In our case, the model predicted that a Pima Indian woman is diabetic, but she is not.
- **FN (False negative):** the model predicted negative, false.in our case, the model predicted that a Pima Indian woman is not diabetic, but she is.

TP, TN, FP, FN are used to calculate the performance measurement or metric of the classification method. It is important to understand different relevance metrics, to choose the right ML model and make decisions based on the predictions. The following performance metrics were used to select the optimal model: [29]

- **Accuracy**: is one way to evaluate classification models. It is the percentage of predictions each model got right. In other words, it is the number of accurate predictions divided by the total number of predictions. Mathematically, Accuracy = (TP +TN) / (TP + TN + FP +FN)
- **Precision**: The precision of a model describes how many detected items are truly relevant (true positive) to the total predicted positive observations. In our study, which would be the measure of the Pima Indians women that we correctly identified having diabetes out of all the women having it. Mathematically, Precision =TP / (TP + FP)
- **Recall**: measures how many relevant elements were detected. Therefore it divides true positives by the number of relevant elements. In our project, for all the women who have diabetes, recall tells us how many we correctly identified or predicted as having diabetes. Mathematically, **Recall** =TP / (TP + FN). Recall should be high as possible.

4.2 Results:

The following Table4 summarizes all the performance metrics of all the models used in our study.

Table 4: Comparative Performance of Classification Algorithms on Various Measures.

Parameters	Models/Metric	Accuracy	Precision	Recall
Unbalanced Dataset and Unscaled Data	Logistic Regression	**0.82**	**0.70**	**0.60**
	Support Vector Machine	**0.81**	**0.85**	**0.53**
	Decision Tree Classifier	0.72	0.61	0.60
	Random Forest Classifier	0.79	0.73	0.66
Unbalanced Dataset and Scaled Data	Logistic Regression	0.73	0.71	0.71
	Support Vector Machine	0.69	0.68	0.63
	Decision Tree Classifier	0.73	0.73	0.65
	Random Forest Classifier	0.75	0.72	0.76
Balanced Dataset and Unscaled Data	Logistic Regression	0.73	0.71	0.71
	Support Vector Machine	0.69	0.68	0.63
	Decision Tree Classifier	0.73	0.73	0.65
	Random Forest Classifier	0.75	0.72	0.76
Balanced dataset and Scaled data	Logistic Regression	0.71	0.74	0.60
	Support Vector Machine	**0.80**	**0.75**	**0.86**
	Decision Tree Classifier	0.70	0.65	0.71
	Random Forest Classifier	0.76	0.71	0.84

4.3Analysis of Model Performance:

When creating classification models for any disease, keeping the end-user in mind is crucial. In this project, the classifications models we tested are used to detect Type II diabetes. The features chosen to determine the early detection of this disease are shown earlier in Table 2. Table 4 summarizes the different metrics of the sixteen models of the four algorithms. As we mentioned earlier, each algorithm was evaluated under four conditions: (1) unbalanced dataset

and unscaled Data, (2) unbalanced dataset and scaled data, (3) balanced dataset and unscaled data, and (4) balanced dataset and scaled data. By observing all the metric's results in Table 4, we can analyze that the three best accuracies were: 82%, 81%, and 80%, respectively, belonging to the following models: Logistic Regression tuned on unbalanced with unscaled data, Support Vector Machine tuned on unbalanced with unscaled data, and finally Support Vector Machine tuned on a balanced dataset with a scaled data. However, both logistic regression models and Support Vector Machine tuned on unbalanced dataset unscaled data did not have the best Recall metric. They had respectively around 60% and 53%, which is bad for our project since the models identified correctly [53% - 60%] of the women having diabetes, which can be dangerous for the [47 % - 40%] of the Pima Indian women who had diabetes, Yet the model detected them as not diabetic. Nevertheless, the recall metric of the model SVM tuned on a balanced dataset with scaled data was 86%, implying that the model correctly identified most of the women having diabetes.

In this project, recall performance is particularly important since it measures how accurately our model can identify the relevant data (True Positive). Imagine that our Pima female is diabetic, but there is no treatment given to her because our model predicted that she was not diabetic. That is a situation we would like to avoid. Thus, we relied on performance recall and accuracy to select the optimal model. As a result, the Support vector machine trained on balanced dataset with scaled data was the optimal model for predicting diabetes disease in the females of Pima Indians.

5.Conclusion:

Type II diabetes is often referred to as a lifestyle disease, affecting many major organs, including the heart, blood vessels, nerves, eyes, kidneys, and many other organs. This project aimed to identify factors and indicators that have a substantial risk of predicting Type II Diabetes

in women to allow the women who have those factors to take the necessary actions to prevent the onset of the disease.

In this paper, we have predicted diabetes using the female Pima Indians diabetes dataset. We have selected glucose, pregnancies, age, insulin, and BMI as the features to perform the prediction of the outcome. Various supervised learning algorithms have been used to evaluate the best performance, such as logistic regression, random forest, decision tree, and SVM were trained to evaluate the best model. Each algorithm was trained on unscaled data with an unbalanced and balanced dataset and on scaled data with an unbalanced and balanced dataset. Therefore, four models were generated for the same algorithm; consequently, sixteen models generated the training and testing dataset. The metrics: accuracy, recall, and precision were used to select the best model. The results of the sixteen models suggest that the SVM model trained on the balanced dataset with scaled data performed best in comparison to other models.

If this project were to be expanded in the future, it would be judicious to gather more data since the dataset had only 798 records. In addition, we ought to collect records of men. Moreover, we would add more features—for example, daily exercise and dietary records. Having more and different data may help better predict Type II diabetes, allowing both men and women to take the necessary actions to prevent this onset of the disease.

Appendix:

Table 5: List of Variables

Variable Name	Description
class_0	returns the number of instances related to class 0 (nondiabetic person)
class_1	returns the number of instances related to class 1(diabetic person)
DF	pandas' data frame containing all features
DF2	pandas' data frame containing only selected features
Imputer	imputation for completing missing values using k-nearest neighbors
Detect_Outliers	elimination of outliers by calculating interquartile 25 and interquartile 75
CorrMatrix	creates a matrix of correlation
Test_Under	creates a balanced dataset.
X	variable features to train and to model
y	target variable
X_train	training features variable , 'glucose', 'pregnancies', 'age,' 'insulin', 'bmi','skin thickness'
y_train	training target variable 'outcome'
X_test	testing features variable
y_test	testing target variable
train_test_split	split arrays or matrices into random train and test subsets
logreg	implements logistic regression and regularized logistic regression models.
y_pred	prediction values of default logistic regression on unscaled data and unbalanced dataset
Cnf_Matrix	creates confusion matrix from y_test and model of logistic regression on unscaled and unbalanced dataset
SVM1	instance of the SVM classifier model
y_pred1	prediction values of the support vector classifier model on unscaled and unbalanced dataset
cm1	creates confusion matrix from y_test and the svm classifier model on unscaled and unbalanced dataset

DecisionTree2	implements the decision tree classifier
y_pred2	prediction values of the decision tree classifier model on unscaled and unbalanced dataset
cm2	creates confusion matrix from y_test and the decision tree classifier model on unscaled and unbalanced dataset
Rforest3	Instance of the Random Forest Classifier model
y_pred3	prediction values of the random forest classifier model on unscaled and unbalanced dataset
cm3	creates confusion matrix from y_test and the random forest classifier model on unscaled and unbalanced dataset
logreg4	implements the Logistic Regression model
y_pred4	prediction values of default logistic regression on scaled and unbalanced dataset
cm4	creates confusion matrix from y_test and model of logistic regression on scaled and unbalanced dataset
SVM5	implements the Support Vector Classifier model on scaled and unbalanced dataset
y_pred5	prediction values of the support vector classifier model on scaled and unbalanced dataset
cm5	creates confusion matrix from y_test and model the support vector classifier model on scaled and unbalanced dataset
DecisionTree6	implements the Decision Tree Classifier model
y_pred6	prediction values of the Decision Tree Classifier model on scaled and unbalanced dataset
cm6	creates confusion matrix from y_test and the decision tree classifier model on scaled and unbalanced dataset.
RForest7	implement the Random Forest Classifier model on scaled and unbalanced dataset
y_pred7	prediction values of the Random Forest Classifier model on scaled and unbalanced dataset
cm7	creates confusion matrix from y_test and the Random Forest Classifier model on scaled and unbalanced dataset

logreg8	implements of the Logistic Regression model
y_pred8	prediction values of default logistic regression on unscaled and balanced dataset
cm8	creates confusion matrix from y_test and model of logistic regression on unscaled and balanced dataset
SVM9	implements the Support Vector Classifier model on unscaled and balanced dataset
y_pred9	prediction values of the Support Vector Classifier model on unscaled and balanced dataset
cm9	creates confusion matrix from y_test and model the support vector classifier model on unscaled and balanced dataset
DTree10	implements the Decision Tree Classifier model on unscaled and balanced dataset
y_pred10	prediction values of the decision tree classifier model on unscaled and balanced dataset
cnf_matrix10	creates confusion matrix from y_test and the decision tree classifier model on unscaled and balanced dataset
RForest11	implements the random office Classifier model on unscaled and balanced dataset
y_pred11	prediction values of the random forest classifier model on scaled and unbalanced dataset
cm11	creates confusion matrix from y_test and the random forest classifier model on unscaled and balanced dataset
logreg12	instance of the logistic regression model with scaled dataset and balanced dataset
y_pred12	prediction values of default Logistic Regression on scaled and balanced dataset
cm12	creates confusion matrix from y_test and model of logistic regression on scaled and balanced dataset
SVM13	implements the Support Vector classifier model on scaled and balanced dataset
y_pred13	prediction values of the Support Vector Classifier model on scaled and balanced dataset
cm13	creates confusion matrix from y_test and model the support vector classifier model on scaled and balanced dataset

DTree14	implements of the Decision Tree Classifier model on scaled and balanced dataset
y_pred14	prediction values of the decision tree classifier model on scaled and balanced dataset
cm14	creates confusion matrix from y_test and the decision tree classifier model on scaled and balanced dataset
RForest 15	implements of the random forest classifier model on scaled and balanced dataset
y_pred15	Prediction values of the Random Forest Classifier model on scaled and balanced dataset
cm15	creates confusion matrix from y_test and the Random Forest Classifier model on scaled and balanced dataset

References

1. Type 2 diabetes . Wikipedia. [accessed 2022 April 12]. https://en.wikipedia.org/wiki/Type_2_diabetes

2. Symptoms & Causes of Diabetes | NIDDK. National Institute of Diabetes and Digestive and Kidney Diseases. 2014 December 1 [accessed 2022 December 16]. https://www.niddk.nih.gov/health-information/diabetes/overview/symptoms-causes

3. Type 2 Diabetes -CDC. Centers for Disease Control and Prevention. [accessed 2022 September 12]. https://www.cdc.gov/diabetes/basics/type2.html

4. Type 2 diabetes - Symptoms and causes - Mayo Clinic. Mayo Clinic. 2021 November 9 [accessed 2021 November 17]. https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193

5. Pima Indians Diabetes Database . Kaggle. [accessed 2022 September 1]. https://www.kaggle.com/uciml/pima-indians-diabetes-database

6. Kumar Bhoi S, Anshuman Abhisek P. Prediction of Diabetes in Females of Pima Indian Heritage: A Complete Supervised Learning Approach. Turkish Journal of Computer and Mathematics Education. 2021 April 27:3074-3084.doi: 10.17762/turcomat.v12i10.4958

7. Thammi Reddy A, Nagendra M. Minimal Rule-Based Classifiers using PCA on Pima-Indians-Diabetes-Dataset. International Journal of Innovative Technology and Exploring Engineering. 2019 October 12;8(12):4414-4420. doi:10.35940/ijitee.l2476.1081219

8. Naive Bayes classifier. Wikipedia. [accessed 2022 January 9]. https://en.wikipedia.org/wiki/Naive_Bayes_classifier

9. Glucose tolerance test . Mayo Clinic. [accessed 2021 October 16]. https://www.mayoclinic.org/tests-procedures/glucose-tolerance-test/about/pac-20394296

10. Hypertension. World Health Organization. [accessed2021 October 16]. https://www.who.int/news-room/fact-sheets/detail/hypertension

11. Ramírez-Vélez R, López-Cifuentes M, González-Ruíz K. Triceps and Subscapular Skinfold Thickness Percentiles and Cut-Offs for Overweight and Obesity in a Population-Based Sample of Schoolchildren and Adolescents in Bogota, Colombia. Nutrients. 2016;8(10):595. doi:10.3390/nu8100595

12. Insulin . Wikipedia. [accessed 2021 October 16]. https://en.wikipedia.org/wiki/Insulin

13. Body_mass_index . Wikipedia. [accessed 2021 October16]. https://en.wikipedia.org/wiki/Body_mass_index.

14. Bhaya WS. Review of Data Preprocessing Techniques in Data Mining. Journal of Engineering and Applied Sciences. 2017 September ;12(16):4102-4102. https://www.researchgate.net/publication/319990923_Review_of_Data_Preprocessing_Techniques_in_Data_Mining. doi:10.3923/jeasci.2017.4102.4107

15. Brownlee J. kNN Imputation for Missing Values in Machine Learning. 2020 June 1 [accessed 2022 October 16]. https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/

16. Taylor C. What Is the Interquartile Range Rule? Thought Co. 2018 April 26 [accessed 2021 December 7]. https://www.thoughtco.com/what-is-the-interquartile-range-rule-3126244

17. Roy B. All about Feature Scaling. Scale data for better performance. Towards Data Science. 2020 April 6 [accessed 2022 January 15]. https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35

18. Standard deviation . Wikipedia. [accessed 2022 January 2]. https://en.wikipedia.org/wiki/Standard_deviation

19. Hale J. Scale, Standardize, or Normalize with Scikit-Learn. Towards Data Science. 2019 March 6 [accessed 2022 January 4]. https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02

20. Longadge R, SewakDas Dongre S. Class Imbalance Problem in Data Mining Review. International Journal of Computer Science and Network. 2013;2(1):2277-2278. https://www.researchgate.net/publication/236651567_Class_Imbalance_Problem_in_Data_Mining_Review

21. Algina J, Keselman HJ. Comparing squared multiple correlation coefficients: Examination of a confidence interval and a test significance. Psychological Methods. 1999;4(1):76-83. doi:10.1037/1082-989x.4.1.76

22. Picard RR, Berk KN. Data Splitting. The American Statistician. 1990;44(2):140-145. doi:10.2307/2684155

23. Logistic regression . Wikipedia. [accessed 2022 January 3]. https://en.wikipedia.org/wiki/Logistic_regression

24. Decision tree - Wikipedia. [accessed 2022 January 1]. https://en.wikipedia.org/wiki/Decision_tree

25. Random forest - Wikipedia. [accessed 2022 January 2]. https://en.wikipedia.org/wiki/Random_forest

26. Yiu T. Understanding Random Forest. How the Algorithm Works and Why it. Towards Data Science. 2019 June 10 [accessed 2022 February 1]. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

27. Waseem M. How To Implement Classification In Machine Learning? Edureka. 2022 May 10 [accessed 2022 April 12]. https://www.edureka.co/blog/classification-in-machine-learning/

28. Brownlee J. What is a Confusion Matrix in Machine Learning. Machine Learning Mastery. 2016 November 15 [accessed 2022 January 4]. https://machinelearningmastery.com/confusion-matrix-machine-learning/

29. Joshi R. Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures. Exsilio Solutions. 2016 September 9 [accessed 2022 January 1]. https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/