Diabetes Detection Using Machine Learning Classification Methods

Nour Abdulhadi
Computer Engineering Department
Princess Sumaya University for Technology
Amman, Jordan
nou20180738@std.psut.edu.jo

Amjed Al-Mousa
Computer Engineering Department
Princess Sumaya University for Technology
Amman, Jordan
a.almousa@psut.edu.jo

2021 International Conference on Information Technology (ICIT) | 978-1-6654-2870-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIT52682.2021.9491788

***Abstract*— The main objective of this research is to predict the possible presence of diabetes -specifically in females- at an early stage using different machine learning techniques. Early detection of diabetes can significantly prevent the progression of the disease and reduce the risk of serious complications such as heart and kidney diseases, making the proper lifestyle changes at the right time can help avoid diabetes and all the illnesses associated with it. So, there is a crucial need for a tool that can better assist doctors to detect this deadly disease at an early stage and consequently stop its progression. Finally, this model produced an accuracy of 82% based on the random forest classifier model.**

Keywords—Machine learning; diabetes; mellitus; Pima Indians dataset; classification;

I. INTRODUCTION

Diabetes is a disease where glucose, or blood sugar, is not metabolized by the body which increases the glucose rate to alarmingly high levels. Normally, a hormone called insulin helps control the amount of glucose in one's bloodstream, people with diabetes either don't produce insulin (type 1 diabetes) or don't respond to insulin the way they should (type 2 diabetes). Approximately, 90% of all diagnosed cases of diabetes are that of type 2. [1]

According to [2], the number of people living with diabetes more than tripled between 1990 and 2010, and the number of new cases doubled every year. Why are the numbers rising so fast? Obesity is believed to account for 80-85% of the risk of developing type 2 diabetes [3] and the World Health Organization (WHO) studies have shown that worldwide obesity has nearly tripled since 1975 [4], this leads to the belief that the escalating rates of obesity and type 2 diabetes are directly linked to each other. One of the reasons for the global rise in obesity is that people are eating more high-calorie, high-fat foods and are less physically active [5] because new technological advancements provide entertainment, education, communication, and all type of purchases right on the spot.

This model focuses on the early detection of type 2 diabetes because it is more common. [1] To carry out the training and testing of the machine learning model, the Pima Indians dataset from the National Institute of Diabetes and Digestive and Kidney Diseases was used. All patients in this dataset are females that are at least 21 years old and with Pima Indian Heritage, Pima Indians in the United States have the world's highest recorded prevalence and incidence of type 2 diabetes [6] which is why the study was conducted on this specific group to generate this dataset. It consists of 8 medical predictor variables (attributes), and a single target, outcome. The outcome is the variable that specifies if a patient has been diagnosed with type 2 diabetes or not. The dataset contains 768 instances.

The remaining part of the paper is organized as follows: Section II includes previous work that addressed the same problem. Section III introduces the complex details of the used dataset, the process of preparing the data and making it suitable for a machine learning model, and the machine learning algorithms used. Moreover, the results of every technique used and the associated accuracy of it are presented in Section IV. At last, a conclusion is outlined in section V.

II. RELATED WORK

Machine learning has been used successfully for the prediction of many outcomes, ranging from the likelihood of being admitted to a university [7], predicting what books you might like based on your history [8], or even predicting who tweeted a certain tweet [9]. But a more relevant case is the use of machine learning in the detection of heart disease using majority ensemble methods [10]. In addition, multiple algorithms and models have been trained in the field of diabetes detection, and several methods have been used to perform data pre-processing. In [11], a dataset consisting of 178131 instances has been used to train a model that reached an accuracy of 80.8%. The model used the Random Forest Classifier method when using all 14 physical examination features that included: age, pulse rate, height, weight, fasting glucose, etc.
A model using the Pima Indians Dataset used the k-nearest neighbor (KNN) algorithm and tried different k-values ranging from 1 to 100 to reach a maximum receiver operating characteristic accuracy of 74% when k was set to 0. [12]
Moreover, the paper in [13] is a study to build an effective prediction model to identify Canadian patients at risk of having Diabetes Mellitus based on patient demographic data and the laboratory results during their visits to medical facilities, it has been trained on a dataset that contains 13309 Canadian patients with their ages ranging between 18 and 90 years. The Gradient Boosting Machine (GBM) technique performed best according to the evaluation of *area under the receiver operating characteristic curve* (AROC), the AROC for this model is 84.7% with a sensitivity of 71.6%.

Table 1: PIMA Indian Dataset Attributes Description

Attributes	Range	Description
Pregnancies	0-17	Number of times pregnant
Glucose	0-199	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Blood Pressure	0-122	Diastolic blood pressure (mm Hg)
BMI	0-67.1	Body mass index = (weight in kg/(height in m)^2)
Skin Thickness	0-99	Triceps skin fold thickness (mm)
Diabetes Pedigree Function	0.078-2.42	A function that scores the likelihood of diabetes based on family history
Age	21-81	Age in years
Insulin	0-846	2-Hour serum insulin (mu U/ml)
Outcome	0-1	Class variable, diagnoses classes: 0 = healthy, 1 = diagnosed with diabetes

Finally, the research in [14] used multiple techniques on different datasets. The algorithms used included Naïve Bayesian, Random Forest (RF), KNN and used evaluation techniques like K-fold Cross-Validation. The highest accuracy achieved on the Pima Indian dataset (which was used as an example of a numeric-only dataset in the research) was 64.47% using the k-fold cross-validation technique.

III. EXPERIMENTAL SETUP

The main purpose of this paper is to build a model that predicts diabetes at an early stage using the previously mentioned dataset. It is a real-world dataset taken from a specific group in a specific area as previously mentioned. Part of the data will be used to train the model, and the other to test it making it able to adapt to new unknown data to predict the outcome.

A. Dataset Attribute Information

Each of the 768 instances in the dataset has 9 attributes, one of them being the target variable. A description of each attribute is present in Table 1.

To get a further insight into the data, correlation values were calculated to know how much an attribute affects the target attribute (Outcome) or if other attributes are affected by it. Correlation values were calculated using the Pearson (product-moment) correlation coefficient equation. It computes the ratio of the covariance of both features to the product of their standard deviations consequently finding the measure of the linear relationship between those two features. Correlation values are shown in Table 2.

Table 2: Correlation with Outcome (Target)

Attribute	Correlation Value
Pregnancies	0.22
Glucose	0.49
Blood Pressure	0.17
BMI	0.22
Skin Thickness	0.21
Diabetes Pedigree Function	0.31
Age	0.17
Insulin	0.24

The heat map of the calculated correlation values is shown in Figure 1 below.

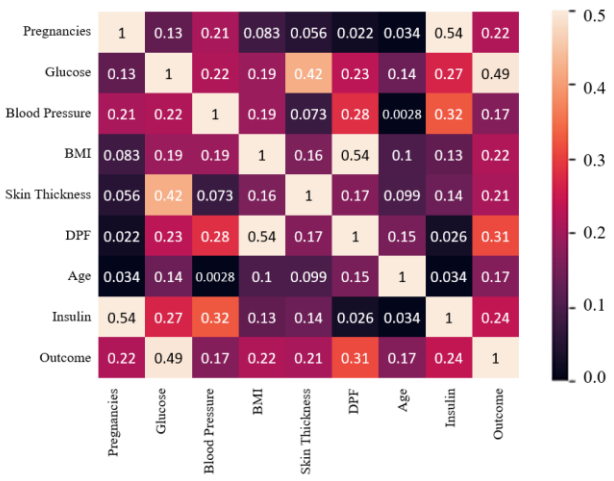

Figure 1: Heat map to show the correlation between features

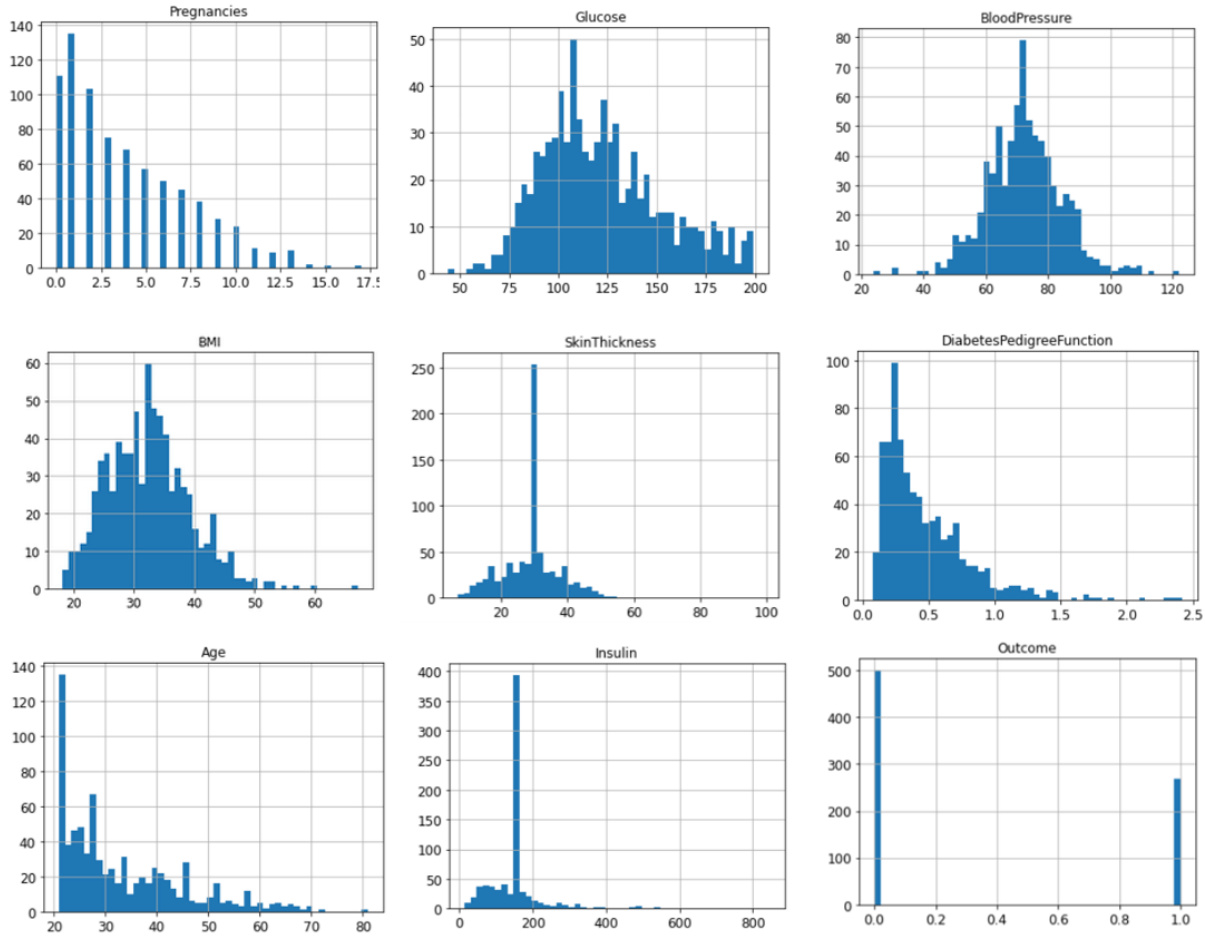

Figure 2: Histograms of the different attributes

It can be observed that glucose has the highest positive correlation with the outcome variable, followed by Diabetes Pedigree Function.

Moreover, histograms were generated to have a better visual interpretation of the data, shown in Figure 2. In addition to the better visualization histograms provide, the figures can make it easier to detect possible outliers that may negatively affect the proposed model.

B. Data Preprocessing

The quality of the data used to train the model significantly affects the results, especially when exposed to new data. Real-world data can contain errors or missing values, as well as outliers. Preprocessing of data helps minimize the effect of such errors, increasing the success rate of the project at hand.

In the Pima Indian Dataset, multiple values were missing from a couple of instances. Having zero blood pressure, for example, does not make any sense. Since the number of instances present (768) was quite low, instead of dropping instances with zeros, the values were filled with the mean. Please note that Figure 1 was generated after the missing values were filled in.

Also, the dataset had different scales, so it had to be standardized. Skipping this step could lead to the contribution of a feature more than the other to the target, whereas when the range of all features is normalized each feature contributes approximately proportionately to the final decision. The dataset was standardized using a standard scaler.

IV. MACHINE LEARNING ALGORITHM

After analyzing the data and filling in all the missing values in attributes such as blood pressure, skin thickness, and BMI, the data was split into two parts: test set and training set. The training set will be used to test the model, while the test set will be used to validate the ability of the model to generalize to new data. The classifier models that have been tested are:

A. Logistic Regression Classifier

The first model that was used is the Logistic Regression Classifier, it is similar to the linear regression model that computes a weighted sum of the input features, but instead of outputting the result as the Linear Regression does, it outputs the logistic of the result. [15] It models the chance of a certain outcome based on individual characteristics.

B. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis projects the features in higher dimension space onto a lower-dimensional space. [16] A series of steps are performed starting by calculating the between-class variance, followed by the within-class variance, and finally constructing a lower-dimensional

space that minimizes the within-class variance calculated and maximizes the between-class variance.

C. Linear Support Vector Machine (SVC)

Linear SVC is one of the algorithms that is commonly used when the data is likely to be linearly separable. According to [17], "the objective of a Linear SVC (Support Vector Classifier) is to fit the data you provide, returning a "best fit" hyperplane that divides, or categorizes your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is."

D. Polynomial Kernel with SVC

The polynomial kernel method with SVC is similar to Linear SVC mentioned above, but it allows the learning of non-linear models instead of only linear. The kernel simply adds more features to the data by making combinations of the features already present, and since increasing the number of features increases the possibility of the data being linearly separable it may result in a higher accuracy compared to a model that only uses linear SVC.

E. Random Forest Classifier

This method is one of the simplest and most diverse algorithms used for both classification and regression tasks, it uses multiple individual decision trees to operate as a single one. Each tree classifies the class to which an instance belongs, and the class with the highest votes is the predicted class.

F. Voting Classifier

A voting classifier makes predictions using multiple classifiers and predicts the output based on the predicted probabilities for each classifier, since the model uses soft voting the outcome chosen depends on the calculated weighted probabilities. The three models I used in this classifier are LDA, Logistic Regression, and Random Forest Classifier.

V. RESULTS

Beginning with the logistic regression classifier, the accuracy of the trained model on the unseen dataset was 80%. Figure 3 shows the confusion matrix of this model.

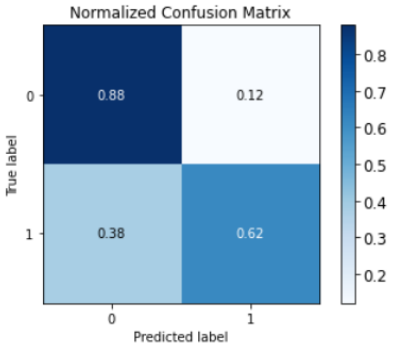

Figure 3: Logistic Regression Confusion Matrix

The second model was trained using the Linear Discriminant Analysis classifier. The accuracy of the model, when run on the unseen test set, was 79%. Figure 4 shows the confusion matrix obtained from this model.

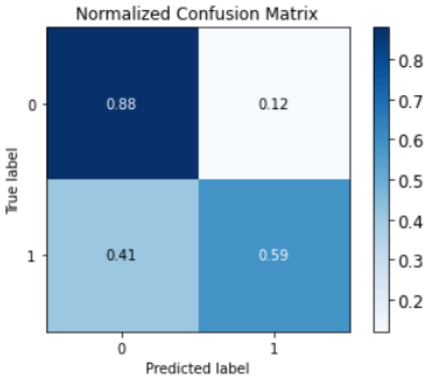

Figure 4: Linear Discriminant Analysis Confusion Matrix

Next, the model was trained using the Linear Support Vector Machine. The parameters were set as C=1, kernel = 'Linear' as the nonlinear one will be tested next. The accuracy obtained from this model was also 79%. Figure 5 shows its confusion matrix.

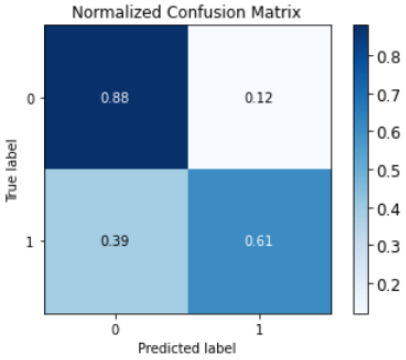

Figure 5: Linear SVM Confusion Matrix

Moving on to the fourth model, it was trained using polynomial SVM of degree 2, and C was set to 5. This model had an accuracy of 79%. Figure 6 shows the confusion matrix of its results.

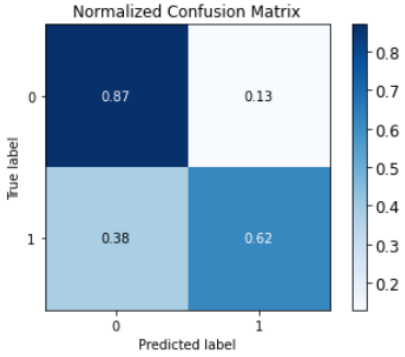

Figure 6: Polynomial SVM Confusion Matrix

Reaching the highest-scoring classifier, the model was trained using the Random forest technique obtaining the highest accuracy among the mentioned classifiers. The accuracy, when run on unseen data, was 82%. Figure 7 shows the confusion matrix of its results.

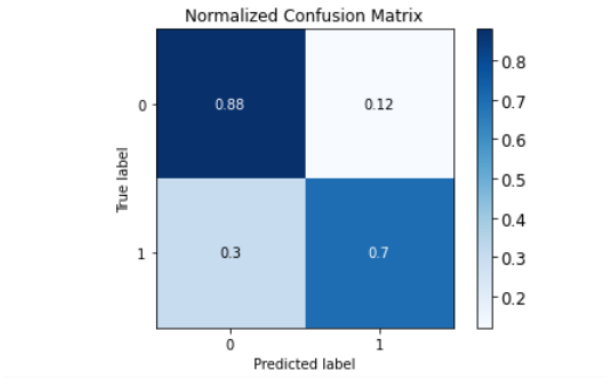

Figure 7: Random Forest Classifier Confusion Matrix

Reaching the final classifier, the model was trained using the voting classifier. The three models LDA, Logistic Regression, and Random Forest Classifier had accuracies of 79%, 80%, and 82% respectively. The voting classifier that combined all of them had an accuracy of 80%. Figure 8 below shows the confusion matrix of this classifier.

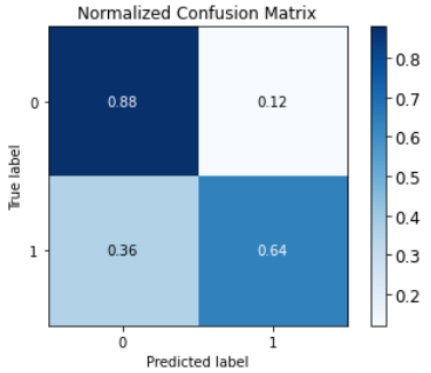

Figure 8: Voting Classifier Confusion Matrix

Table 3 shows the accuracy scores of the 6 models trained.

Table 3: Accuracy of Trained Models

Model Name	**Accuracy**
Logistic Regression	80%
LDA	79%
Linear SVC	79%
Polynomial kernel SVC	79%
Random Forest Classifier	82%
Voting Classifier	80%

VI. CONCLUSION

In a nutshell, this project's main goal was to build a model using supervised learning methods that could help assist doctors in the early detection of diabetes to improve the quality of patient's lives. The paper presented multiple techniques that were used to train multiple models, Random Forest Classifier achieved the highest accuracy of 82%.

ACKNOWLEDGMENTS

The authors would like to thank PSUT for supporting the publication of this research, which is based on a machine learning class project.

REFERENCES

[1] "Diabetes Fast Facts," [Online]. Available: https://www.cdc.gov/diabetes/basics/quick-facts.html.

[2] "Statistics and facts about type 2 diabetes," Medical news today, [Online]. Available: https://www.medicalnewstoday.com/articles/318472.

[3] "Diabetes and Obesity," The global diabetes community, [Online]. Available: https://www.diabetes.co.uk/diabetes-and-obesity.html.

[4] "Obesity and overweight," World Health Organization, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[5] "Obesity Rising: Can We Do Anything to Reverse This Deadly Trend?," healthline, [Online]. Available: https://www.healthline.com/health-news/obesity-rising-can-we-reverse-this-deadly-trend#Complex-factors-behind-obesity.

[6] L. O. Schulz, P. H. Bennett, E. Ravussin, J. R. Kidd, K. K. Kidd, J. Esparza and M. E. Valencia, "Effects of Traditional and Western Environments on Prevalence of Type 2 Diabetes in Pima Indians in Mexico and the U.S.," *Diabetes Care,* vol. 29, no. 8, pp. 1866--1871, 2006.

[7] Z. Bitar and A. A. Al-Mousa, "Prediction of Graduate Admission using Multiple Supervised Machine Learning Models," in *IEEE SoutheastCon*, Raleigh, 2020.

[8] S. Khalifeh and A. A. Al-Mousa, "A Book Recommender System Using Collaborative Filtering," in *Data'21*, Petra, 2021.

[9] L. Ahmad and A. A. Al-Mousa, "Identification of Donald Trump's Tweets Using Machine Learning," in *Multi-Conference on Systems, Signals & Devices*, Monastir, 2021.

[10] R. Atallah and A. A. Al-Mousa, "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method," in *2nd International Conference on new Trends in Computing Sciences (ICTCS)*, Amman, 2019.

[11] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Front Genet,* vol. 9, no. 515, 2018.

[12] H. Deshmukh, "Pima Indians Diabetes - Prediction & KNN Visualization," towards data science, [Online]. Available: https://towardsdatascience.com/pima-indians-diabetes-prediction-knn-visualization-5527c154afff.

[13] H. Lai, H. Huang, K. Keshavjee, A. Guergachi and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," *BMC Endocr Disord,* vol. 19, no. 101, 2019.

[14] M. N. H. R. L. Asmita Singh, "Impact of Different Data Types on Classifier," Melbourne, 2017.

[15] J. Zheng, "Resume of Logistic & Softmax Regression," 16 April 2019. [Online]. Available: https://jingwen-z.github.io/resume-of-logistic-softmax-regression/.

[16] J. Brownlee, "Linear Discriminant Analysis for Dimensionality Reduction in Python," Machine Learning Mastery, 2020. [Online]. Available: https://machinelearningmastery.com/linear-discriminant-analysis-for-dimensionality-reduction-in-python/.

[17] "Linear SVC Machine learning SVM example with Python," PythonProgramming, [Online]. Available: https://pythonprogramming.net/linear-svc-example-scikit-learn-svm-python/.