Diabetes Detection In Women

Prof. Ajay Talele
Assistant Professor
Vishwakarma Institute of Technology
Pune, India

Aanya Jain
Computer Engineering
Vishwakarma Institute of Technology
Pune, India

Aashi Mathur
Computer Engineering
Vishwakarma Institute of Technology
Pune, India

Abhijeet Bharate
Computer Engineering
Vishwakarma Institute of Technology
Pune, India

Adarsh Pathak
Computer Engineering
Vishwakarma Institute of Technology
Pune, India

Aditi Dharmadhikari
Computer Engineering
Vishwakarma Institute of Technology
Pune, India

***Abstract*—This paper discusses the development of a web application that uses a machine learning algorithm to predict the presence, or the absence of diabetes based on medical data of the user. The steps undertaken to understand the dataset, clean the dataset, so that it is ready to train a model on, selecting the right model, and developing a web application using principles of objective oriented programming are discussed in detail in this paper. Random Forest was chosen from all the tested classifiers as it provided the best accuracy of 84.07%. Streamlit, an open-source python library, was used to develop and host the web application. Concepts like Abstraction, Encapsulation, Inheritance, Polymorphism and Exception Handling were employed to make the application structured and efficient. The application provides various input parameters for common users as well as for medical staff. The prediction and the graphs are displayed on the application to help the user better understand the importance of the variables.**

Keywords—Predictive modeling, Machine learning, Diabetes, Diabetes in women, Web application.

I. Introduction

Number of people suffering from Diabetes has risen from 200 million in 1990 to 830 million in 2022. This addition in the number of people having diabetes is more common in low and middle-income nations than in high-income countries.

Diabetes causes high blood sugar levels, which damage the blood vessels. Diabetes is also known to cause heart attacks, stokes, blindness and kidney failures.

Diabetes occurs when the body is not able to produce enough insulin or when the body is not able to use the produced insulin.

A healthy diet accompanied by regular physical exercise and restrain from tobacco are essential to prevent or delay the onset of diabetes.

Diagnosing a patient with diabetes takes a few days as the samples have to be sent to a laboratory, and the results back to the sender. This is a resource intense process. Machine Learning helps by predicting which patient may have a higher chances of having diabetes. Medical professionals can accordingly decide what samples need not be sent to the laboratory or what samples must be sent.

This research paper goes through the steps our team followed to develop a web application that employs machines learning to predict the presence of diabetes.

II. Literature Review

In [5], N. Fazakis et al, focus on achieving high sensitivity and area under the curve (AUC) through supervised machine learning techniques. The paper discusses the use of Pearson correlation and LASSO for identifying significant features that contribute to T2DM prediction.

In line with the above-mentioned paper was the paper [4] of G.A. Colditz et al, in which a strong association was found between increased BMI and diabetes risk, with relative risks significantly rising for women with a BMI over 22 kg/m², particularly pronounced for those over 25 kg/m². The findings suggest that even women of average weight are at increased risk for diabetes, emphasizing the importance of current weight over historical weight.

In [2], various algorithms, including OPTICS and BIRCH, are utilized for data clustering and analysis, enhancing the accuracy of diabetes detection by T. Sharma et al. Deep learning models, particularly DNNs, show superior accuracy in detection in diabetes compared to traditional machine learning methods, with accuracy reaching up to 99.5%.

In [1], E. Afsaneh et al, discuss multiple datasets and algorithms, such as Decision Trees, Support Vector Machines, and Random Forests, applied to different diabetes and patient demographics. Significant findings include the development of predictive models with high accuracy for diagnosing diabetes and forecasting blood glucose levels, with some models achieving AUC values above 0.85.

In [3], A range of machine learning models, including Naive Bayes, SVM, and Random Forest, are evaluated for their predictive performance by E. Dritsas et al. Experiments are conducted on a high-performance computing system, utilizing 10-fold cross-validation and an 80:20 data split for model evaluation. Results indicate that KNN and Random Forest models achieve the highest accuracy, with KNN reaching 98.59% and Random Forest 99.22% under different validation methods.

C.-Y. Chou et al, evaluated model performance using metrics such as true positives, false negatives, accuracy, precision, recall, F1 score, and AUC values in [6]. The research successfully utilized eight characteristics to develop a predictive model, achieving high AUC scores of 0.976 and 0.991.

III. METHODOLOGY

This project was majorly divided into six steps. The first step being understanding the dataset. In this step, the entire team understood the meaning of each feature in the dataset. Since the dataset did not have many features, this step did not take long. The next step was preprocessing this data so that it can be used to train algorithms. This step included removal of misinputs, removal of contradicting inputs, normalization of the features to equalize their scale and deletion of outliers. The next step was selecting a suitable machine learning algorithm for the prediction and use in web application. The final step was to develop a web application that takes inputs from the users and displays the prediction made by the algorithm. All the mentioned steps are discussed in detail below.

A. Dataset

The dataset used in this project was sourced from Kaggle. This data is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the original dataset, to obtain a final dataset which comprised women of Indian heritage only. The final dataset contains 768 records and 9 features namely 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age' and 'Outcome'.

Columns in the dataset:

- Pregnancies: Number of times the subject has been pregnant.
- Glucose: Plasma glucose concentration in the bloodstream.
- BloodPressure: Diastolic blood pressure in mm Hg.
- SkinThickness: Thickness of the skin in mm when folded at the triceps.
- Insulin: Insulin level in the blood in micro-IU/ml.
- BMI: Body Mass Index of the subject.
- DiabetesPedigreeFunction: A function that evaluates the probability of diabetes based on family history.
- Age: Age of the subject in years.
- Outcome: Target variable. Tells whether the subject has diabetes or not (1 or 0).

B. Data Preprocessing

The raw dataset underwent cleaning to handle wrong values, inconsistencies, and ensure proper formatting.

The following preprocessing steps were undertaken:

- Initial Analysis: Basic information about the dataset, such as column data types and non-null counts, was reviewed to understand its structure. No NA values were observed throughout the dataset.
- Removing Misinputs: Records having Blood Pressure as zero, BMI as zero, Glucose as zero and Skin Thickness as zero were deleted.
- Removing Inconsistent Entries: Records having insulin level as zero but outcome as 0 were removed.
- Normalization: This dataset contains features of a varying scale, Diabetes Pedigree Function ranges from 0 to 2.24 whereas, Blood Pressure is in the 70s or 80s. To bring each feature to a common scale, z score normalization was applied.
- Removing Outliers: Outliers were removed from the normalized dataset by calculating the Inter Quartile Range for each feature.

C. EDA (Exploratory Data Analysis)

- Histogram: Histogram plots were used for plotting the distribution of Glucose, Blood Pressure, Skin Thickness, Insulin, BMI and Diabetes Pedigree Function to better understand the spread of the data.
- Count plot: This plot was utilized to visualize the class imbalance present in Outcome.
- Pair plot: In our project, the pair plot provided insights into how various numerical features correlate.
- Heatmap: The heatmap helped us understand relationships between numerical variables. The heatmap is colour coded based on the magnitude of the correlation coefficient of the two selected variables.

D. Feature Selection

Dimensional Reduction was not required as the dataset only had 9 features. However, for convenience of the user of the web application, the most easily accessible features were selected to create a separate data-frame. Skin Thickness and Diabetes Pedigree Function were removed as they cannot be accurately measured independently. This data-frame along with the pre-processed data-frame would be used to train and test various classification algorithms.

E. Model Selection

This stage focused on testing various Classifiers available and selecting the one with the best accuracy. The data was divided into training and testing segments. 70% of the data was reserved for training the algorithm and the remaining 30% to test the algorithm. We had observed an imbalance in the data earlier. This imbalance was fixed by resampling the training data. The testing data was not resampled as resampled testing data hinders the performance of the model and may give false good results. The training data was resampled using SMOTE (Synthetic Minority Oversampling Technique). SMOTE was chosen over Random Oversampling as SMOTE does not generate duplicate rows like Random Oversampling which leads the algorithm to become biased. These steps were applied to both the original preprocessed data-frame and the data-frame created for user convenience.

The resampled data was then used to train various classifiers like Logistic Regression, Decision Trees, Random Forest, k-NN, Naïve Bayes etc. The accuracy of all these algorithms was noted on both the data-frames.

At the end Random Forest was chosen to be used in the web application as it provided the best accuracy of 82.30% with the original data-frame and an accuracy of 84.07% with the data-frame created for the users.

F. *Web Application Development*

Streamlit was chosen to develop an interactive and user-friendly web application. The Home page contains information about Diabetes, statistics sourced from the World Health Organization, and a link to the Predict page. The application features a dedicated Predict page where users can provide inputs to get a prediction about their chance of being diabetic. Users can choose the input parameters, User' or 'Professional'. 'User' section contains only easily accessible parameters whereas 'Professional' contains all the parameters that were present in the original dataset. Once the inputs are filled and submitted, a message is displayed according to what the algorithm predicts. If the prediction is that the user is diabetic, graphs showing the distribution of Glucose, Blood Pressure, Insulin and BMI are displayed with information about where the user might be plotted based on the data used to train the algorithm. A disclaimer message stating that these graphs are generated based on the training data is also displayed to inform the user of the math and logic used to calculate the graph. A Working page informs the users of the algorithms used, their accuracies, and plots of normalized features so that the users can better understand the relation of each input with the probability of having diabetes. A link to a detailed report is also provided in case the user wants to know the reason behind the choice of the algorithm.

G. *Application of Principles of Object Oriented Programming*

In the application development, the principles of Object Oriented Programming were used to organize and simplify the structure of the application.

Encapsulation ensures that all the data needed to normalize the user inputs stays in a single class and is hidden from the user.

Abstraction is utilized to hide unnecessary information like how the data is being normalized and fed into the algorithm.

Polymorphism in the form of method overloading is used. The two forms (User and Professional sections of the Predict page) call the same function but provide different arguments to normalize the data.

Inheritance is used to display output to the user. The information about what model is loaded is stored in a class which acts a parent class, and the child class uses that information to display the prediction made by that model.

Exception Handling is employed to ensure that if any errors occur while trying to load the model file, they are properly communicated to the user.

IV. Results and Discussions

The accuracies of various models tested are listed below.

With all features:

Algorithm	Accuracy (%)
Logistic Regression	79.646
Decision Trees	76.1062
Random Forest	82.3001
kNN	72.5664
Naïve Bayes	78.7611
SVM	77.8761
Gradient Boosting	81.4159

With selected features for user convenience:

Algorithm	Accuracy (%)
Logistic Regression	81.4159
Decision Trees	75.2212
Random Forest	84.0707
kNN	75.2212
Naïve Bayes	76.9911
SVM	79.6461
Gradient Boosting	79.6461

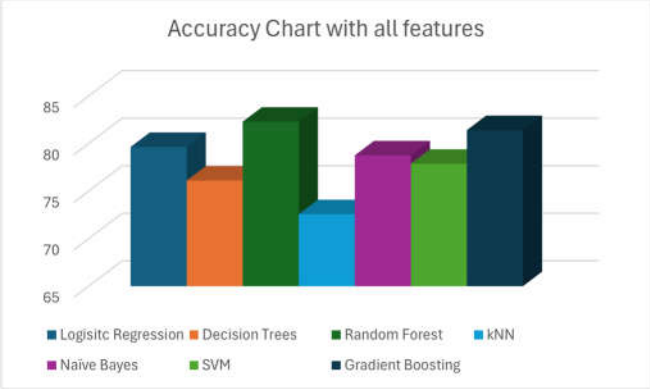

Figure 1. Accuracies on dataset having all the features.

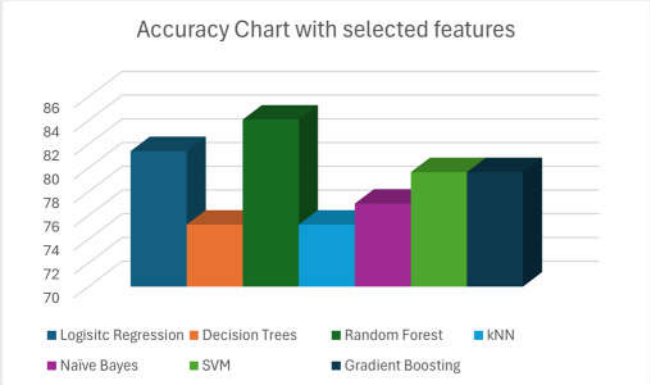

Figure 2. Accuracies on dataset having only selected features for user.

Figure 3. Screenshot of the Users section in the web application.

Figure 4. Screenshot of the prediction displayed by the application.

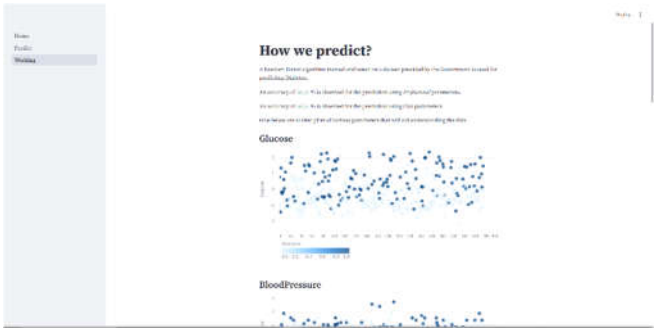

Figure 5. Screenshot of the Working page.

V. Conclusion

In conclusion, during this project, the task of choosing a model and using it in a web application to predict the presence of diabetes was successfully accomplished using Machine Learning and principles of Object Oriented Programming.

VI. Future Scope

The future scope of this predictive machine learning application includes addition of more features like integration with a database, where users can store and keep track of their medical data. Using multithreading to make the application more efficient. Training more algorithms on different data and including that in the application to improve the accuracy of the prediction. Developing algorithms not only for Diabetes in women but for the general population. In the future, we can branch out to other diseases which are harder to detect like cancer.

Acknowledgment

The authors would like to express their sincere gratitude to Vishwakarma Institute of Technology for providing the opportunity and resources to pursue this project. In addition, the author would like to thank their guide, Prof. (Dr.) Ajay Keshav Talele, and all faculty members of the DOME (Department of Multidisciplinary Engineering) department for their support, guidance, and invaluable insights throughout the project.

References

[1] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, "Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review," Diabetology & Metabolic Syndrome, vol. 14, no. 196, pp. 1-11, 2022, doi: 10.1186/s13098-022-00969-9

[2] 2. T. Sharma and M. Shah, "A comprehensive review of machine learning techniques on diabetes detection," Visual Computing for Industry, Biomedicine, and Art, vol. 4, no. 30, pp. 1-12, 2021, doi: 10.1186/s42492-021-00097-7

[3] 3. E. Dritsas and M. Trigka, "Data-Driven Machine-Learning Methods for Diabetes Risk Prediction," Sensors, vol. 22, no. 5304, pp. 1-21, 2022, doi: 10.3390/s22145304

[4] G. A. Colditz, W. C. Willett, M. J. Stampfer, J. E. Manson, C. H. Hennekens, R. A. Arky, and F. E. Speizer, "Weight as a risk factor for clinical diabetes in women," American Journal of Epidemiology, vol. 132, no. 3, pp. 501–513, 1990

[5] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, "Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction," IEEE Access, vol. 9, pp. 114769–114780, 2021

[6] C.-Y. Chou, D.-Y. Hsu, and C.-H. Chou, "Predicting the Onset of Diabetes with Machine Learning Methods," Journal of Personalized Medicine, vol. 11, no. 8, pp. 1–12, 2021

[7] Ghimire, Devndra. "Comparative study on Python web frameworks: Flask and Django." (2020).

[8] Laakso, Markku, and Kalevi Pyörälä. "Age of onset and type of diabetes." Diabetes care 8.2 (1985): 114-117.

[9] Colditz, Graham A., et al. "Weight as a risk factor for clinical diabetes in women." American journal of epidemiology 132.3 (1990): 501-513.

[10] Colditz, Graham A., et al. "Diet and risk of clinical diabetes in women." The American journal of clinical nutrition 55.5 (1992): 1018-1023.

[11] American Diabetes Association. "Management of diabetes in pregnancy." Obstetrical & Gynecological Survey 72.5 (2017): 264-266.

[12] Hunt, Kelly J., and Kelly L. Schuller. "The increasing prevalence of diabetes in pregnancy." Obstetrics and gynecology clinics of North America 34.2 (2007): 173-199.

[13] Chou, Chun-Yang, Ding-Yang Hsu, and Chun-Hung Chou. "Predicting the onset of diabetes with machine learning methods." Journal of Personalized Medicine 13.3 (2023): 406.

[14] Malik, Sumbal, Saad Harous, and Hesham El-Sayed. "Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women." International Symposium on Modelling and Implementation of Complex Systems. Cham: Springer International Publishing, 2020.

[15] Wei, Sidong, Xuejiao Zhao, and Chunyan Miao. "A comprehensive exploration to the machine learning techniques for diabetes identification." 2018 IEEE 4th World Forum on Internet of Things (WF-IoT). IEEE, 2018.