International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 09 | Sep 2021 www.irjet.net p-ISSN: 2395-0072

Web-Application to Predict Diabetes using Machine Learning Classification Algorithms

Veer Kumar[1]

[1]*Student, Dept. of Electronics and Telecommunications and Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra, India*

Abstract – *Diabetes, also known as the silent killer, is one of the most deadly diseases of the 21st century that causes a surge in blood sugar. It plays a major role in the development of conditions such as kidney failure, blindness, limb amputations, and heart diseases. Just in the period of 16 years i.e 2000-2016, there was a 5% uptick in premature mortalities being directly attributed to diabetes. In 2019, as many as 1.5 million deaths were directly caused by diabetes earning it the tag of being called a silent killer. Under normal circumstances, a patient would want to visit a healthcare center and have a doctor decide the line of treatment to be followed. However, the experimental analysis that I have implemented seek to help patients who are at remote locations, with no immediate healthcare resources at their disposal. This web application aims to be a user-friendly interface for any individual having valid medical inputs, to check the presence of diabetes. I have implemented 3 different classifiers with appropriate hyperparameter tuning to better the previously achieved accuracy levels.*

Key Words: Diabetes, Hyperparameter Tuning, Streamlit, KNN, SVM, Random Forest.

1. INTRODUCTION

Diabetes is a common health condition worldwide caused by inadequate production of insulin, due to which blood sugar throughout the body cannot be regulated. The pancreas is the organ responsible for insulin generation, which helps glucose from the food enter into our cells to produce energy. However, sometimes, the body does not produce enough of it, thus, leaving glucose to stay in the blood, not reaching cells.

- *Under type1 diabetes, the body does not make insulin and the immune system attacks the cells in the pancreas that produce insulin. This is more common among children and adolescents.*
- *Under type 2 diabetes, the human body doesn't make insulin at all or it does not use the available insulin in the required manner. This condition is more common amongst middle-aged and older people.*

1.1 DATASET

We have used the Pima Indian diabetes dataset, which was created by the national institute of Diabetes and Digestive and Kidney Diseases. It consists of several medical parameters(such as Pregnancies and Glucose). The dataset is focused on the female gender and comprises of 9 columns and 8 independent parameters, and we have 768 observations. Here, we have 268 testing positive for diabetes, which is indicated by '1' and 500 respondents turn out to be negative for it.

Column Features	Definition
Pregnancies	Number of times pregnant
Glucose	The Glucose tolerance test(Oral)
blood pressure	These are Diastolic readings of blood pressure values(which is the pressure being developed in the arteries during the resting phase of the heart in between beats)
skin thickness	Skinfold thickness at triceps, so that we can find out the total body fat in an individual.
Insulin	Insulin facilitates the movement of blood sugar(or glucose), into the cells from the bloodstream.
BMI	Body Mass Index, tells us whether the patient is at the risk of being over or underweight.
DiabetesPedigreeFunction	It provides information on the history of Diabetes mellitus in relatives and those who are genetically related to the patient.
Age	Patient age in years
Outcome	This is our target variable, and here a classification of 1 indicated that the person has diabetes while 0 indicates otherwise.

1.2 Literature Review

The primary objective of this research project is to create an easy-to-use Diabetes predictor web application that is accessible to the general public. The user interface was

developed keeping in mind those patients who aren't technically proficient, this user-friendly web application will allow users to enter various parameters required by our model to make predictions. While making a user-friendly application is the top priority, I did not want to compromise on the results, hence exhaustive research has been conducted by me wherein I studied several research papers published in the past, that have tackled the same problem statement(predicting diabetes) by using different approaches. After thoroughly analyzing similar research work on several other medical datasets where numerous machine learning techniques were employed to make predictions, we obtain plenty of results. These predictive models were created by other researchers who made use of varying techniques to mine the data and carry out the implementation of machine learning algorithms on them. K.VijiyaKumar et al. [11] presented a Random Forest algorithm for the Prediction of diabetes to develop a system that could detect the presence of diabetes in a patient in the early stages with higher accuracy by using the Random Forest algorithm in the machine learning algorithm. This model gives the most accurate results for diabetic prediction and the result showed that the prediction system is capable of predicting diabetes disease more instantly and effectively. N. Joshi et al. [12] proposed Diabetes Prediction Using Machine Learning Techniques to detect the presence of diabetes by three different supervised machine learning methods that included: SVM, Logistic regression, Artificial Neural Networks. This project puts forth an effective technique for the early detection of diabetes. Deeraj Shetty et al. [15] presented diabetes disease prediction using the Diabetes Disease Prediction System that gave an analysis of diabetes malady using the diabetes patients database. In this system, they made use of algorithms like Bayesian and KNN (K-Nearest Neighbor) and applied them to diabetes patients database, following which they analyzed them by taking various attributes responsible for diabetes into account, to detect diabetes. B.M. Patil, R.C. Joshi, and Hindu deity Toshniwal(2010) proposed HybridPredictionModel which implements s K-means clump algorithmic program, after which classification of the algorithmic program is done to the result of the clump algorithmic program. To create classifiers, the C4.5 Decision Tree algorithmic program was employed. d.[10] Mani Butwall and Shraddha Kumar (2015) put forth a model using Random Forest Classifier to forecast and predict diabetes behavior. r.[7] Nawaz Mohamudally and Dost Muhammad (2011) implemented the C4.5 decision tree algorithm, Neural Network, K-means clustering, and used data Visualization techniques to predict diabetes. .[8] Humar Kahramanli and Novruz Allahverdi (2008) made use of an Artificial neural network coupled with fuzzy logic to predict diabetes disease.

2. About Libraries:

While developing this diabetes predicting web application I made use of some fundamental libraries that aided by the project and helped me make the predictive model, they were as follows:

1. Streamlit: This is an open-source Python library that makes it convenient to create and deploy powerful custom-made web applications for machine learning and data science projects.

2. Pandas: It is a fast and efficient tool to carry out data analysis and data manipulation. In addition to that, it allows us to import data from varied file formats such as JSON, CSV's, SQL, and Microsoft EXCEL, enabling us to perform complex data cleaning and wrangling operations on these files to read, interpret and gather insights from raw data.

3. Numpy: It is an array processing package that allows us to perform several scientific computations on arrays, including several high-level mathematical functions.

4. Sklearn: This is the fundamental machine learning library in Python, It contains a wide variety of modules to aid machine learning and statistical modeling processes such as classification, regression, clustering, and dimensionality reduction.

5. PIL: Python imaging library, is an open-source additional library that provides features that allow opening, manipulation, and saving many different image file formats.

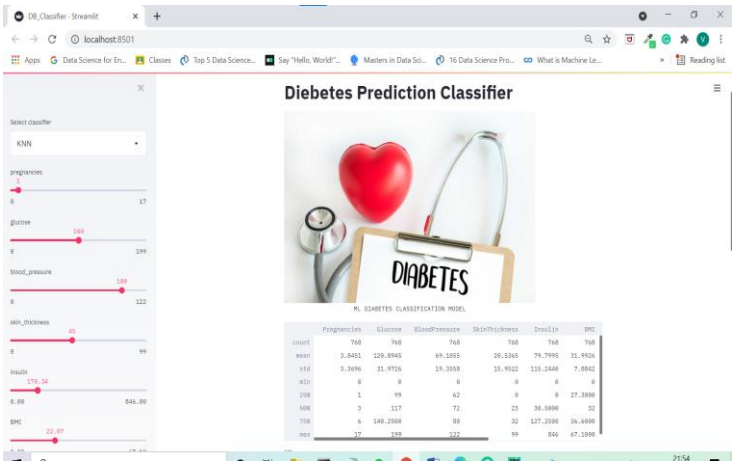

Fig 1. Diabetes Prediction Classifier Web-Application Interface

3. Algorithms:

3.1 K-Nearest Neighbours(KNN):

KNN is a supervised machine learning algorithm, it assumes the similarity between the new data points and available data points and puts the new datapoint into the category that is most similar to the available set of categories. It is a non-parametric algorithm, in other words, it will not make any assumption on the underlying data.

3.2 Support Vector Machine(SVM):

SVM is also a supervised machine learning algorithm, that can be used for classification and regression tasks. Here, we plot each data point in an n-dimensional space, and here 'n' is the number of features in our dataset. Basically, in SVM the algorithm will create a line or hyperplane due to which the data points get split into two distinct classes.

3.3 Random Forest:

Random Forest is another supervised learning algorithm, where the forest is said to be an ensemble of decision trees. In simple terms, the random forest classifier will divide the dataset into different subsets and these subsets are fed into every tree of the random forest algorithm. Every decision tree will produce its specific output.

4. Methodology:

4.1 Dataset collection:

We feed the Pima Indian Diabetes dataset as a CSV file and read it into pandas, converting it into a data frame, which enables us to perform various numerical operations on the dataset to transform it.

4.2 Dataset Pre-processing steps:

In this step, we are primarily concerned with separating the feature columns from our target column, which in our case is the 'Outcome' feature, which shall predict whether or not a particular patient is positive for diabetes.

4.3 Feature Scaling:

As we can infer from the above dataset, all of our features have different units of measurement, thus, there is a wide variation in the range of feature values between these columns. In order to account for this difference, we need to standardize our features by re-scaling them, we do this by applying a feature scaling technique known as StandardScaler() function.

4.4 Train-Test Split and Model Fitting:

Now, we divide our dataset into training and testing data. Our objective for doing this split is to assess the performance of our model on unseen data and to determine how well our model has generalized on training data. This is followed by a model fitting which is an essential step in the model building process, as without proper model fitting our predicted outcome will not be suitable to implement under practical scenarios.

4.5 Hyperparameter Optimization:

This is a subset of Feature engineering, we use Hyperparameter Optimization techniques to fine-tune our machine learning models. Since fitting the model with arbitrary parameters may not necessarily yield the best accuracies hence we run hyperparameter tuning methods to find the optimal combination of hyperparameters that would reduce the pre-defined loss function and give higher test accuracies. In my model, I have implemented hyperparameter tuning using the GridSearchCV module.

4.6 Model Evaluation and Predictions:

This is the final step, in which we assess how well our model has performed on testing data using certain scoring metrics, I have used 'accuracy_score' to evaluate my model.

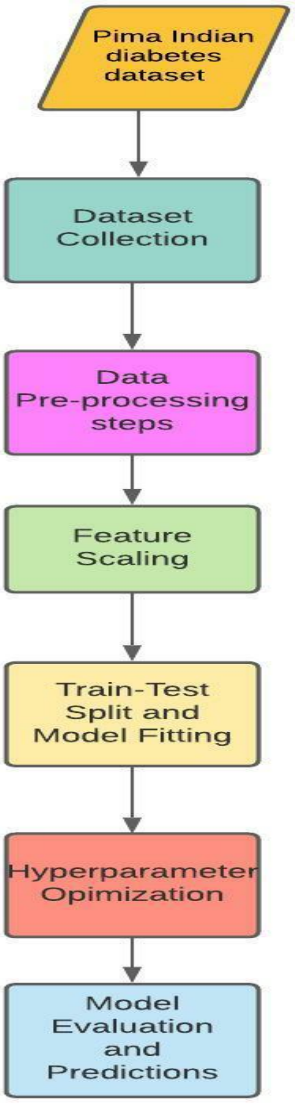

Fig 2. Methodology Flowgraph

5.0 RESULTS AND INFERNECES:

Algorithms	Test Accuracy
KNN	75.55%
SVM	77%
Logistic Regression	76.41%
Random Forest	80.42%

As we can see from the above test accuracies, for a given configuration of input parameters the Random Forest algorithm does perform marginally better than the other algorithms under consideration. Although, we must not that this testing accuracy has been noted for a specific configuration only, and varying the input parameters can result in slightly different results, although, this was found to be the general trend across many inputs that I have considered in my research.

6.0 OUTPUT:

ML DIABETES CLASSIFICATION MODEL

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	
count	768	768	768	768	768	
mean	3.8451	120.8945	69.1055	20.5365	79.7995	31
std	3.3696	31.9726	19.3558	15.9522	115.2440	7
min	0	0	0	0	0	
25%	1	99	62	0	0	27
50%	3	117	72	23	30.5000	
75%	6	140.2500	80	32	127.2500	36
max	17	199	122	99	846	67

Fig 3. Summary statistics of various features in the PIMA Indian diabetes dataset.

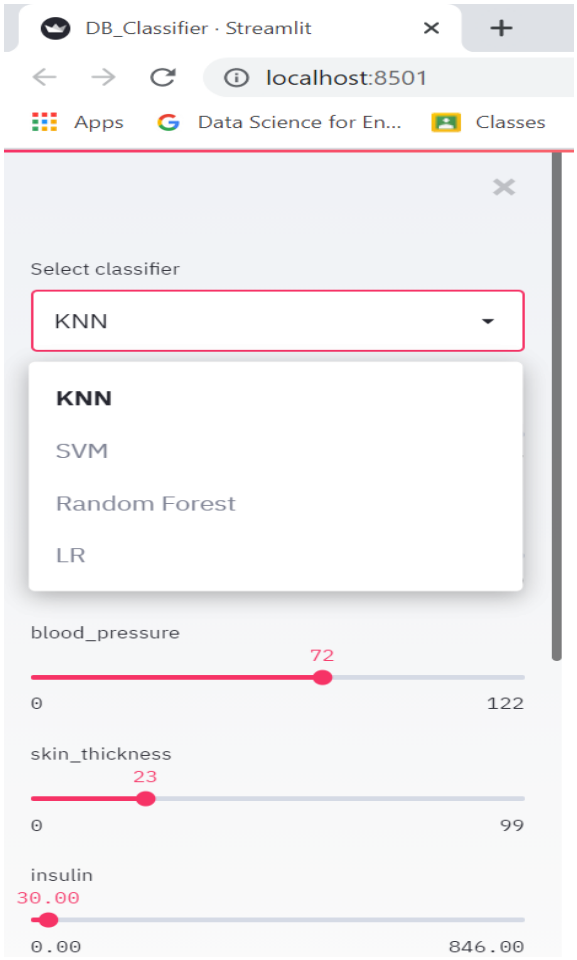

Fig 4. Sidebar to select the type of classifier and input parameters.

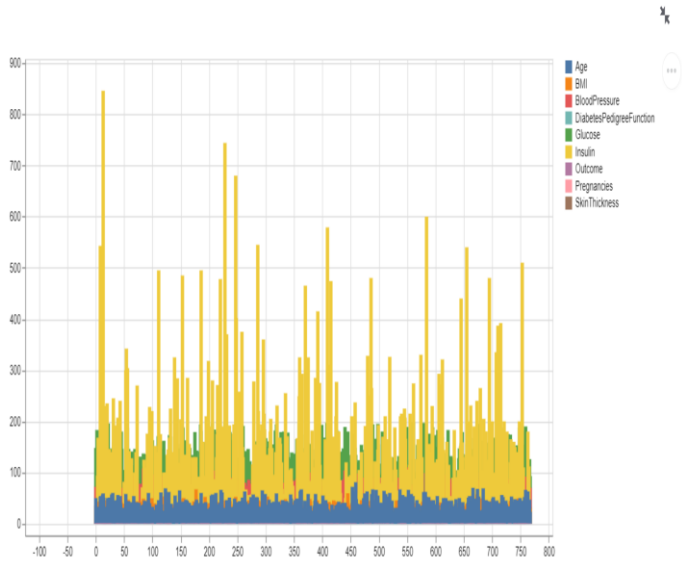

Fig 5. The bar graph plots column features along with corresponding values.

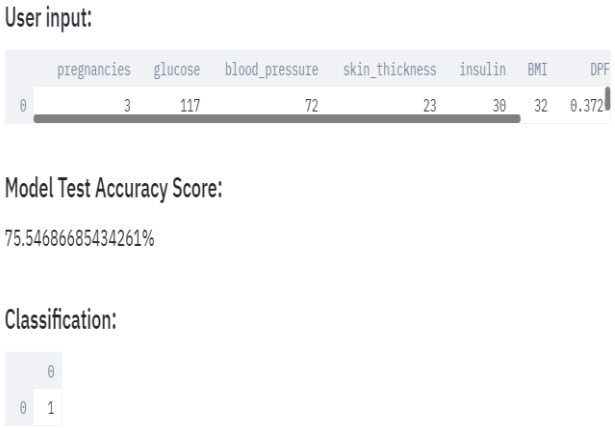

Fig 5. Displaying user input, Model test accuracy score obtained and the prediction made by classification model.

3. CONCLUSION

In the research I have conducted, We can infer that although all classifiers perform decently well, the Random forest classifier has performed marginally better as compared to the others by providing us a test accuracy of 80.42 on average, In addition to this, I observed that implementing hyperparameter tuning techniques like GridSearchCV help in improving model test accuracy by obtaining the most optimal parameters required to make accurate predictions on unseen data. Apart from this, the user interface I have built using streamlit broadens the scope of the utility for this application, as with some more fine-tuning and by incorporating additional parameters we can further boost the test accuracy, following which we can deploy this application for large-scale use. One drawback of the above model is that as we increase the number of cross-validations in hyperparameter tuning, the execution time of our code also increases linearly, one way to tackle this could be to reduce the dimensionality of our dataset by applying principle component analysis(PCA).

REFERENCES

[1] Dr. Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd.
[2] B.M. Patil, R.C. Joshi, and Durga Toshniwal, "Association Rule for Classification of Type-2 Diabetic Patients", ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.
[3] Mani Butwall and Shraddha Kumar, "AData Mining Approach for the diagnosis of diabetes MellitususingRandomForest classifiers".
[4] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm ", Journal Of Computing, Volume 3, Issue 12, December 2011.
[5] Mitchell T. Machine learning. McGraw Hill0-07-042807-7; 1997 2.
[6] Nai-Arun, N., Moungmai, R., 2015. Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science 69, 132–142. doi:10.1016/j.procs.2015.10.014.
[7] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.
[8] Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015
[9] Diabetes, World Health Organization (WHO): 30 Oct 2018.
[10]Changsheng Zhu, Christian Uwa Idemudia, Wenfang Feng. "Improved logistic regression model for diabetes

prediction by integrating PCA and K-means techniques" SCIENCE DIRECT: 4 April 2019, Vol. 17, p1.
[11]S., Hina, A., Shaikh, and S., Abul Sattar, "Analyzing DiabetesDatasets using Data Mining," Journal ofBasic& Applied Sciences, vol. 13, pp. 466–471, 2017.
[12] C. D. Mathers and D. Loncar, "Projections of Global Mortality and Burden of Disease from 2002 to 2030," PLoS Medicine, vol.3, no.11,p.e442, Nov.2006. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

BIOGRAPHIES

Mr. Veer Kumar is a student, currently studying at Rajiv Gandhi Institute of Technology, Mumbai. He is in his final year of Electronics and telecommunications engineering. Veer is passionate about the field of Analytics and looks to explore opportunities, as he aspires to become a Data Scientist in the near future.