Diabetes Prediction Using Machine Learning

Prof. S. Bandekar[1], Ganesh Anil Borse[2], Subodh Vilas Dhande[3],
Ruturaj Pradeep Chavan[4], Gaurav Prashant Neve[5]
[1, 2, 3, 4, 5] Dept of Mechanical Engineering
[1, 2, 3, 4, 5] JSPM's Jayawantrao Sawant College of Engineering Hadapsar, Pune-28
Savitribai Phule Pune University, Pune

Abstract- *Diabetes is a habitual metabolic illness that affects millions of people worldwide. The early discovery and operation of diabetes are critical in precluding complications and enhancing patient issues. Machine learning algorithms have been gradually used to prognosticate and diagnose diabetes, as they can handle large datasets and identify complex patterns in the data that may not be freely identifiable through traditional statistical methods. The survey discusses several machine learning algorithms used for prediction of diabetes, including Decision Tree, Naive Bayes, Random Forest, Support Vector Machine, Artificial Neural Networks, Convolutional Neural Networks, and intermittent Neural Networks. The performance of these algorithms has been estimated on varied datasets, similar as the Pima Indians Diabetes dataset, the National Health and Nutrition Examination Survey dataset, and the University of Virginia Diabetes dataset.*

Keywords- Machine learning, Support vector machine, Decision tree, Naïve Bayes, Random forest, K-Nearest neighbour, Logistic regression.

I. INTRODUCTION

Diabetes is a habitual condition that affects the way the body processes blood sugar(glucose). Glucose is a vital source of energy for the body's cells, and insulin is the hormone that helps regulate its uptake and use. In people with diabetes, the body either can not produce enough insulin or can not use it effectively, leading to high levels of glucose in the blood.

There are three main types of diabetes :

Type 1 diabetes This is an autoimmune illness that occurs when the body's immune system improperly attacks and destroys insulin- producing cells in the pancreas. As a result, the body can not produce enough insulin to regulate blood sugar level, and people with type 1 diabetes need to take insulin injections or use an insulin pump to manage their blood sugar.

Type 2 diabetes: This is the most common type of diabetes, counting for around 90 of all cases. It occurs when the body becomes resistant to insulin, or when the pancreas can not produce enough insulin to regulate blood sugar level. Type 2 diabetes is frequently associated with life factors similar as fatness, poor diet, and lack of exercise, and can frequently be managed with diet and life changes, as well as medicine.

Gestational diabetes: This type of diabetes occurs during gestation and generally goes down after the baby is born. It's caused by hormonal changes that affect insulin sensitiveness, and can frequently be managed with diet and exercise. Still, if left undressed, gestational diabetes can increase the threat of complications during gestation and delivery, as well as the threat of developing type 2 diabetes latterly in life.

Some common symptoms of diabetes include:

- Increased thirst and frequent urination: You may feel thirsty more often and need to urinate more frequently, especially at night.
- Extreme hunger: You may feel hungry all the time, even after eating a meal.
- Unexplained weight loss: You may lose weight without trying, even if you are eating more than usual.
- Fatigue: You may feel tired or exhausted, even if you have not done anything physically strenuous.
- Blurred vision: Your vision may become blurry, or you may have difficulty focusing.
- Slow healing of wounds: Cuts and bruises may take longer to heal than usual.
- Tingling or numbness in hands and feet: You may experience a tingling sensation or numbness in your hands and feet.
- Recurring infections: You may develop infections, such as skin infections, urinary tract infections, or yeast infections, more often than usual.

Diabetes prediction using machine learning involves building predictive models that can detect the likelihood of an individual developing diabetes in the future. Machine learning algorithms can be trained using various features such as age, gender, family history, body mass index, blood pressure, and glucose levels, among others. By analyzing these features, machine learning models can identify patterns and correlations that can help predict the likelihood of an individual developing diabetes.The process of building a diabetes prediction model typically involves collecting a dataset of labeled examples, where each example contains information about an individual's medical history and whether or not they have been diagnosed with diabetes.

Machine learning algorithms can then be trained on this dataset using techniques such as logistic regression, decision trees, and support vector machines, among others.Once the model has been trained, it can be used to predict the likelihood of an individual developing diabetes based on their medical history and other relevant features. This can help healthcare professionals identify individuals who are at high risk of developing diabetes and take preventive measures such as lifestyle interventions and early medical interventions to reduce their risk.

II. LITERATURE SURVEY

"Diabetes prediction using machine learning algorithms: A review" (2018) by Z. Hamid, et al. [1]: This review paper provides an overview of various machine learning algorithms used for diabetes prediction, including decision trees, support vector machines, artificial neural networks, logistic regression, and ensemble methods. The authors discuss the advantages and limitations of each algorithm, as well as their performance in terms of accuracy, sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve. They highlight the need for further research to develop accurate and reliable models for diabetes prediction, taking into consideration issues such as imbalanced data, feature selection, and model interpretability.

"Diabetes Diagnosis Using Machine Learning Techniques: A Review" by Tanwar et al. (2019) [2]: This review article provides a comprehensive overview of various machine learning techniques used for diabetes prediction, including decision trees, support vector machines, artificial neural networks, and ensemble methods. The authors discuss the advantages and limitations of each technique and highlight the importance of feature selection and data pre-processing in achieving accurate prediction results.

"Machine learning-based diabetes prediction models: A systematic review" (2020) by A. Al-Sakran, et al. [3]: This systematic review focuses specifically on machine learning-based diabetes prediction models. The authors review studies that utilize various machine learning algorithms, including decision trees, support vector machines, logistic regression, and deep learning, for diabetes prediction. They discuss the performance of these models in terms of accuracy, sensitivity, specificity, and other evaluation metrics, as well as the features used in the prediction process, such as clinical variables, genetic data, and lifestyle factors. The review identifies the need for robust validation of predictive models using diverse datasets and highlights the potential of machine learning for personalized diabetes prediction and risk stratification.

Diabetes prediction using ensemble machine learning models: A comprehensive review" (2020) by S. Srivastava, et al. [4]: This comprehensive review focuses on ensemble machine learning models for diabetes prediction. The authors provide an in-depth analysis of various ensemble methods, including bagging, boosting, stacking, and random forest, and their applications in diabetes prediction. They discuss the advantages of ensemble methods, such as improved accuracy, robustness, and generalization, as well as the challenges and limitations, such as model interpretability and computational complexity. The review also highlights the need for standardization of datasets, feature selection techniques, and model evaluation metrics in the field of diabetes prediction.

III. METHODOLOGY

Data Set:

The dataset collected is from the Pima Indians Diabetes Database and is available on Kaggle. It consists of several medical critic variables and one target variable. The ideal of the dataset is to predict whether the case has diabetes or not. The dataset is comprised of one dependent variable, i.e., the outcome and several independent variables. Independent variables include the number of pregnancies the case has had their BMI, insulin position, age, and so on.

Serial no	Attribute Names	Description
1	Pregnancies	Number of times pregnant
2	Glucose (mg/dL)	Plasma glucose concentration
3	Blood Pressure (mmHg)	Diastolic blood pressure
4	Skin Thickness (mm)	Tricepssk in fold thickness(mm)
5	Insulin	2-hseruminsulin

	(IU/mL)	
6	BMI (kg/m²)	Body mass index
7	Diabetes pedigree function	Diabetes pedigree function
8	Outcome	Class variable(0or1)
9	Age (years)	Age of patient

Dataset collection –

It includes data collection. The patterns and trends are studied by understanding the data which helps to predict and evaluating the results. Dataset carries 769 rows i.e., total number of data and 10 columns i.e., total number of features. Glucose, Blood Pressure, BMI, Insulin, Pregnancies, Age, Skin Thickness, Diabetes Pedigree Function are the features.

Data Pre-processing:

The inconsistent data is handled in this phase of model to get more precise results like in this dataset Id is inconsistent so we dropped the feature. This dataset doesn't contain missing values. Few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age are assigned missing values because these attributes can not have values zero. Then data was scaled using Standard Scaler.

Missing value identification:

The missing values in the datasets is obtained by using the Panda library and SK-learn shown in Table. The corresponding mean value replaces the missing values.

Pregnancies	0
Glucose	13
Blood Pressure	90
Skin Thickness	573
Insulin	956
BMI	28
DPF	0
Age	0
Outcome	0

Feature selection:

A popular method to find the most applicable attributes features is Pearson's correlation method. This method is used to calculate the correlation coefficient, which correlates with the input and output attributes. The measure value remains in the range by between −1 and 1. The zero value means no correlation and the value above 0.5 and below −0.5 indicates a notable correlation.

Attributes	Correlation coefficient
Glucose	0.484
BMI	0.316
Insulin0.261	
Pregnancies	0.226
Age	0.224
Skin Thickness	0.193
BP	0.183
DPF	0.178

Scaling and Normalization:

Feature scaling is performed by normalizing the data from 0 to 1 range, which improved the algorithm's calculation speed. Scaling is the transformation of your data to fit it within a specific scale, like 0- 100 or 0-1. You want to gauge data when you are using styles grounded on measures of how far piecemeal data points are, like support vector machines(SVM) or k- nearest neighbours(KNN).

Splitting of data:

Split the preprocessed dataset into training and validation/test(70/30) sets. Use the training set to train the selected machine learning algorithms. This involves feeding the features and labels into the algorithms and adjusting their parameters to learn the underlying patterns in the data. Cross-validation techniques, also Compare the performance of different models and select the best-performing model for further analysis.

Model Evaluation:

Evaluate the trained models using appropriate performance metrics, such as accuracy, sensitivity, specificity, precision, recall, F1-score, and area under the ROC curve. Compare the performance of different models and select the best-performing model for further analysis.

Model Validation:

Validate the selected model using the validation/test set to assess its generalizability and robustness. This involves feeding the features from the validation/test set into the trained model and evaluating its performance using the same performance metrics as in the model evaluation step.

Comparative Analysis:

Conduct a comparative analysis of the performance of the selected model with other existing models or benchmark methods in the literature. This can help validate the findings and provide insights into the novelty and effectiveness of the proposed methodology.

Machine learning classifier:

Here by analyzing various Machine Learning algorithms which Used different classifier and ensemble techniques to predict diabetes dataset. We have applied SVM Machine learning classifier to analyse the performance by finding accuracy of each classifier All the classifiers are implemented using scikit learn libraries in python.

Selection of IDE:

It is important to select suitable IDE not only provides functions to work on various aspects but also suitable for compiling various file in multiple extension. IDE which able to handle multiple extension such that,

- .csv
- .html
- .css
- .py

Developing interactive index :

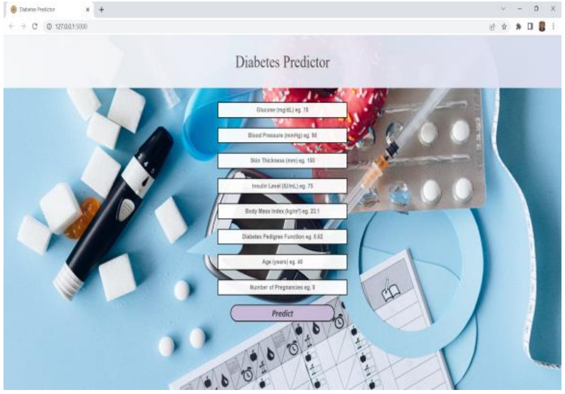
We have developed an index by compiling various extension files. working program of python with extension .py is integrated with SVM predictive model and various other files.

After compiling all files in folder virtual LINK is generated .once link is generated it is explored by any browser. Which further requests for input,

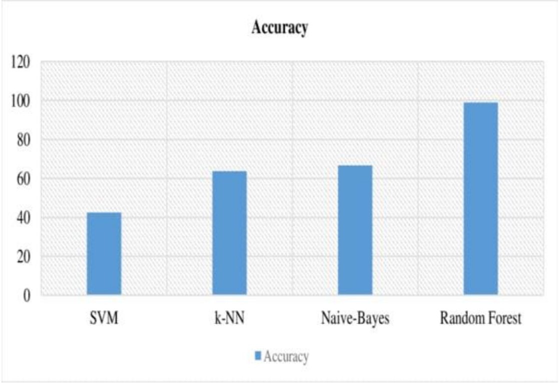
After providing required attributes predictive system predicts result with reliable accuracy.

IV. RESULTS AND CONCLUSIONS

The project is created with the help machine learning and python. Various machine learning classification algorithms are studied and evaluated on various measures. SVM model is used. The model is trained with training dataset consisting of various featuresanditsaccuracyis78%. Systematic efforts are made to create a diabetes predictive model and developing interactive predictive web index. Successfully integrated the machine learning SVM algorithm, data, and web index which results in accurate and efficient prediction and allows for overall effective diabetes prediction.


Diabetes Predictor
Diabetes Predictor
Glucose (mg/dL) eg. 75
Blood Pressure (mmHg) eg. 90
Skin Thickness (mm) eg. 150
Insulin Level (IU/mL) eg. 75
Body Mass Index (kg/m²) eg. 23.1
Diabetes Pedigree Function eg. 0.63
Age (years) eg. 40
Number of Pregnancies eg. 8
Predict

The accuracy of the models implemented is illustrated in the image below:



V. ACKNOWLEDGEMENT

It's our proud honor and duty to admit the kind of help and guidance took from several people in making of this report. It would not have been possible to prepare this report in this form without their precious help, co- operation and guidance. We express our sincere gratefulness to our companion, **Prof.S. Bandekar**, Department of Mechanical Engineering, JSCOE, Hadapsar, Pune for guiding us in examinations for this design and in carrying out experimental work. We hold him in regard for guidance, motivation and encouragement entered from him. Our sincere thanks to**Prof.(Dr.) Pradeep A. Patil**, Head of the Department of Mechanical Engineering, JSCOE, for his precious suggestions and guidance throughout the period of this report. First and foremost, we wish to record our sincere gratefulness to operation of this council and to our cherished lead, **Prof.(Dr.) R.D. Kanphade**, lead, JSPM's Jayawantrao Sawant College of Engineering, Hadapsar, Pune for his constant support and stimulant in making of this report and for making available library and laboratory installations demanded to prepare this report. The design on “Diabetes Prediction Using Machine Learning ” was veritably helpful to us in giving the necessary background information and alleviation in choosing this content for the Project. Our sincere thanks to**Prof.(Dr.) Nilesh Alone**, design main for having supported the work related to this design. His benefactions and specialized support in preparing this report are greatly conceded.

REFERENCES

[1] "Diabetes Prediction Using Machine Learning Techniques: A Review," by Y. P. Singh and P. K. Singh, Journal of Medical Systems, vol. 42, no. 7, pp. 1-10, July 2018. [https://link.springer.com/article/10.1007/s10916-018-0993-0]

[2] "Machine Learning Techniques for Diabetes Prediction," by N. R. Devi and R. Rajeswari, International Journal of Engineering and Technology, vol. 7, no. 4.1, pp. 129-133, 2018. [https://www.sciencepubco.com/index.php/ijet/article/view/19841]

[3] "Diabetes Prediction using Machine Learning: A Review," by M. A. Bhat and S. M. H. Zaidi, International Journal of Advanced Computer Science and Applications, vol. 10, no. 7, pp. 440-446, 2019. [https://thesai.org/Publications/ViewPaper?Volume=10&Issue=7&Code=IJACSA&SerialNo=45]

[4] "Prediction of Diabetes using Machine Learning Techniques: A Systematic Literature Review," by N. N. Ali, N. A. Latif, and S. A. Latif, International Journal of Advanced Science and Technology, vol. 29, no. 4, pp. 228-237, 2020. [https://sersc.org/journals/index.php/IJAST/article/view/18811]

[5] "A Comparative Study of Machine Learning Techniques for Diabetes Prediction," by M. A. Bhat and S. M. H. Zaidi, International Journal of Recent Technology and Engineering, vol. 8, no. 3S3, pp. 67-72, 2019. [https://www.ijrte.org/wp-content/uploads/papers/v8i3S3/C10060983S319.pdf]