[1] **Payal Gupta***

[2] **Prof. (Dr). Ritu Sindhu**

Diabetes Prediction Using Machine Learning

Abstract: - Diabetes, characterized by high blood sugar levels over time, poses challenges for accurate prediction due to limited labelled data and data quality issues like outliers and missing values. To address these challenges, our research introduces a robust prediction framework. This framework integrates techniques such as outlier rejection, filling missing values, standardizing data, selecting relevant features, and employing K-fold cross-validation. We utilize a range of machine learning algorithms including k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, XGBoost, and Multilayer Perceptron (MLP). Additionally, we propose a weighted ensembling technique to enhance prediction accuracy. The weights for ensembling are determined based on the performance of each classifier using the Area Under ROC Curve (AUC) metric. We optimize model performance through hyperparameter tuning using grid search. Our experiments are conducted using the Pima Indian Diabetes Dataset under uniform conditions. The ensembling classifier we propose achieves superior performance with a sensitivity of 0.789, specificity of 0.934, false omission rate of 0.092, diagnostic odds ratio of 66.234, and AUC of 0.950. This outperforms existing methods by 2.00% in terms of AUC. Our framework surpasses other approaches discussed in the literature and shows promise for improving diabetes prediction. We have released our source code for diabetes prediction to the public.

Keywords: Diabetes price prediction, Machine learning, Regression, Time series forecasting, Ensemble methods, Deep learning, Challenges, Future directions.

I. INTRODUCTION

Diabetes, increasingly prevalent even among younger populations, demands a deep understanding of how it manifests. To comprehend its development, one must first grasp the body's normal functioning. Glucose, sourced primarily from carbohydrates like bread, pasta, and fruits, fuels our system after digestion. This glucose circulates in the bloodstream, powering both brain and body cells. Any surplus gets stored in the liver for future energy needs. Insulin, a hormone from the pancreas, acts as a facilitator, allowing glucose into cells for energy production. However, if the pancreas fails to produce enough insulin (insulin deficiency) or if cells become resistant to its effects, glucose accumulates in the blood, leading to diabetes.

There are distinct types of diabetes:

Type 1 diabetes emerges from an immune system malfunction, resulting in insufficient insulin production by pancreatic cells. Despite ongoing research, its exact triggers and preventive measures remain elusive.

Type 2 diabetes occurs when cells produce inadequate insulin or when the body can't use insulin effectively. This form is the most common, affecting around 90% of those with diabetes, and is influenced by genetic predisposition and lifestyle choices.

Gestational diabetes surfaces during pregnancy, marked by elevated blood sugar levels. Although it usually resolves after childbirth, it heightens the risk of type 1 or type 2 diabetes in the future.

Symptoms of diabetes encompass frequent urination, excessive thirst, fatigue, unexplained weight loss, blurred vision, mood swings, difficulty concentrating, and increased susceptibility to infections.

Genetic factors wield considerable influence in diabetes development, with mutations in chromosome 6 affecting the body's response to antigens. Additionally, viral infections such as rubella, Coxsackievirus, mumps, hepatitis B, and cytomegalovirus have been implicated in heightening diabetes risk.

Diabetes, a widespread and detrimental condition worldwide, often stems from factors like obesity or elevated blood sugar levels, disrupting insulin function and leading to abnormal carbohydrate metabolism. According to the World Health Organization (WHO), approximately 422 million people are affected by diabetes, with projections expecting this number to reach 490 million by 2030, primarily affecting low-income nations. This disease is prevalent in various countries, including Canada, China, and India, with India having over 100 million diabetics, accounting for 40 million individuals. Unfortunately, diabetes remains a leading global cause of mortality. However, early detection holds promise in averting severe consequences. To tackle this, our study focuses on predicting

[1] MCA Scholar, Department of Computer Science and Engineering, Lingaya's Vidyapeeth, Faridabad, India. payalgupta7094@gmail.com
[2] Associate Dean & Professor, Department of Computer Science & Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India. ritu.sindhu2628@gmail.com
* Corresponding Author Email: payalgupta7094@gmail.com

diabetes using the Pima Indian Diabetes Dataset, employing various Machine Learning techniques like classification and ensemble methods. By utilizing diverse attributes related to diabetes, these methods facilitate the creation of precise prediction models. Despite the abundance of Machine Learning approaches, selecting the most effective one presents a challenge, prompting us to explore popular classification and ensemble methods to identify optimal prediction strategies.

II. Literature Review

Desai et al. shed light on the underrecognized association between prediabetes and major adverse cardiac and cerebrovascular events (MACCE) in patients with atrial fibrillation (AF) in their recent study published in the World Journal of Diabetes. Atrial fibrillation, characterized by irregular heartbeats, is a significant risk factor for cardiovascular complications, including stroke and myocardial infarction. While traditional risk factors such as hypertension and diabetes are well-established in AF patients, the role of prediabetes in exacerbating MACCE risk remains relatively unexplored.

The recognition of prediabetes as a distinct metabolic state characterized by impaired glucose tolerance or elevated blood glucose levels, albeit not meeting the criteria for diabetes, has garnered increasing attention in recent years. Epidemiological evidence suggests that individuals with prediabetes are at heightened risk of developing overt diabetes and associated complications, including cardiovascular disease. However, its specific impact on MACCE risk in AF patients has received limited attention in the literature.

Previous studies have underscored the intricate interplay between metabolic dysregulation, inflammation, and cardiovascular pathophysiology. Prediabetes, marked by insulin resistance and subclinical inflammation, may contribute to endothelial dysfunction, prothrombotic states, and atherosclerosis, thereby predisposing AF patients to adverse cardiovascular events. Furthermore, the coexistence of AF and prediabetes may create a synergistic effect, exacerbating the underlying cardiovascular risk profile and portending a worse prognosis.

Desai et al.'s study adds to the growing body of evidence highlighting the importance of recognizing and addressing prediabetes as a modifiable risk factor in AF management. By elucidating the association between prediabetes and MACCE in this population, their findings underscore the need for comprehensive risk stratification and tailored therapeutic interventions. Furthermore, the identification of prediabetes in AF patients presents an opportunity for early intervention strategies, including lifestyle modifications, pharmacotherapy, and targeted glycemic control, aimed at mitigating cardiovascular risk and improving clinical outcomes.

In summary, Desai et al.'s study underscores the significance of prediabetes as a potentially overlooked risk factor for MACCE in patients with atrial fibrillation. Their findings underscore the importance of integrated approaches to risk assessment and management in AF patients, emphasizing the need for proactive screening and intervention strategies to mitigate the burden of cardiovascular morbidity and mortality in this population[1].

Y. Wang's retrospective cohort study, published in EClinicalMedicine in 2023, provides insights into the evolving landscape of type 2 diabetes mellitus (T2DM) management and complications in Hong Kong from 2010 to 2019. The study tracks key clinical parameters over a decade, including glycemic control, lipid profiles, blood pressure, and BMI, offering insights into healthcare interventions' effectiveness and guideline changes in optimizing T2DM management and reducing associated complications. Wang's findings shed light on T2DM-related complications' epidemiological burden in Hong Kong, emphasizing macrovascular (e.g., cardiovascular events) and microvascular (e.g., nephropathy, retinopathy) complications' systemic impact over time. Focusing on Hong Kong's unique context, the study provides region-specific data, aiding local healthcare providers and policymakers in tailoring strategies for T2DM prevention and management. Wang's study informs clinical practice and public health policy by highlighting the need for continuous monitoring, early intervention, and personalized management to mitigate T2DM complications and associated morbidity and mortality. In conclusion, Wang's study enhances understanding of T2DM epidemiology and outcomes in Hong Kong, guiding future research and healthcare strategies to improve patient outcomes and reduce the societal burden of T2DM[2].

Howlader KC et al. explore machine learning (ML) models for identifying significant attributes in detecting type 2 diabetes mellitus (T2DM), aiming to address the global burden of T2DM. Their study, published in Health Informatics Science and Systems in 2022, emphasizes ML's potential in leveraging large datasets to extract complex patterns for accurate diagnosis. The study focuses on ML algorithms like decision trees, support vector machines, neural networks, and ensemble methods for classification tasks, aiming to discern T2DM presence or risk. It prioritizes clinically relevant variables such as glucose levels, BMI, lipid profiles, and genetic predispositions to develop robust models facilitating early intervention and personalized management approaches. Howlader KC et al.'s study contributes to ML applications in healthcare, particularly in T2DM detection. By elucidating strengths

and limitations, it informs future research aiming to develop clinically actionable diagnostic tools for T2DM management. Despite challenges, ML-driven approaches hold promise in advancing precision medicine and improving T2DM care outcomes[3].

Nanayakkara et al.'s 2020 study in Diabetologia explores how age at type 2 diabetes mellitus (T2DM) diagnosis affects mortality and vascular complications through a systematic review and meta-analysis. The study consolidates various research, analyzing data to assess the relationship between age of diagnosis and subsequent mortality rates and vascular issues like cardiovascular events and stroke. It emphasizes the significance of timely diagnosis and intervention in mitigating T2DM-related complications. Additionally, the study delves into potential mechanisms behind these associations, providing valuable insights for clinical decision-making. Despite its strengths, caution is advised due to potential biases in meta-analyses and the focus on mortality and vascular complications, possibly overlooking other relevant factors. Overall, the study advances our understanding of T2DM management and points to areas for further research. [4].

Mujumdar and Vaidehi's 2019 study in Procedia Computer Science examines the role of machine learning (ML) algorithms in diabetes prediction. They emphasize ML's importance in early and accurate detection for improved patient outcomes and healthcare management. The study explores various ML algorithms, including decision trees, support vector machines, neural networks, and ensemble methods, to analyze demographic, clinical, and lifestyle data for predicting diabetes risk. ML demonstrates superior performance over traditional methods, thanks to its ability to handle complex data and enhance prediction accuracy. However, challenges such as data quality, interoperability, privacy concerns, and the interpretability of complex ML models remain. Despite these hurdles, Mujumdar and Vaidehi's review highlights ML's potential to transform diabetes prediction and improve patient care outcomes[5].

Kumar et al.'s 2022 study in JMIR Diabetes develops a machine learning-based prenatal predictive risk model aiming to prevent gestational diabetes mellitus (GDM) from progressing to type 2 diabetes mellitus (T2DM). This review emphasizes the importance of early intervention and personalized management in addressing this progression. Their model analyzes various prenatal factors to identify individuals at high risk of developing T2DM after GDM diagnosis. Key findings highlight the model's ability to accurately predict T2DM risk post-GDM, enabling timely interventions and lifestyle modifications. The study underscores the significance of early intervention and tailored strategies in managing GDM-T2DM progression, including dietary counseling and lifestyle changes. Challenges such as data quality and model interpretability need addressing for reliable predictive models.

In conclusion, Kumar et al.'s study demonstrates the potential of machine learning in predicting GDM progression, informing future research and clinical strategies to improve outcomes for at-risk individuals. [6].

Singh, Narayan, and Eggleston's 2019 study in Current Diabetes Reports examines the economic impact of diabetes in South Asia. They focus on healthcare costs, productivity losses, and broader economic consequences. The review underscores the significant burden of diabetes in the region, with rising healthcare expenditures and reduced productivity. It highlights the need for targeted interventions to address disparities and improve health outcomes. The study advocates for proactive health policies, early detection, and effective management programs to mitigate the long-term economic consequences of diabetes. However, challenges such as limited resources and inadequate infrastructure hinder effective diabetes management. Overall, the study emphasizes the importance of collaborative efforts to reduce the socioeconomic burden of diabetes in South Asia. [7].

III. PROPOSED SYSTEM

The objective of our study is to investigate enhanced techniques for forecasting diabetes through the application of diverse classification and ensemble algorithms. Throughout our analysis, we tested a range of methodologies aimed at refining the accuracy of diabetes prediction. Hereafter, we offer a concise summary of our research phase. (Fig 1)

3.1 Methodology used

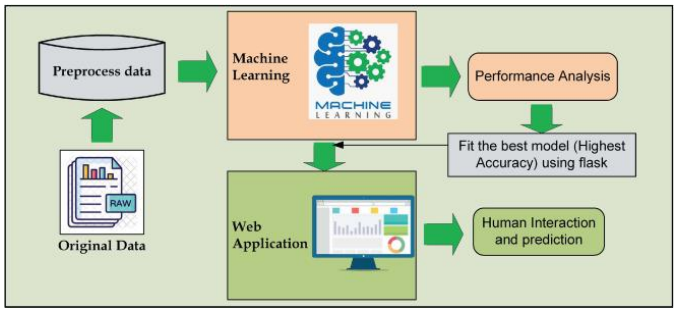


Fig 1: Proposed system work

3.1.1 Data Description:

The dataset on diabetes originated from the Kaggle website and comprises more than 2400 cases. Its purpose is to utilize various measures to predict whether a patient is diabetic or not. The dataset for "Diabetes Prediction" comprises eight distinct features, namely: (i) Pregnancies, (ii) Glucose, (iii) Blood Pressure, (iv) Skin Thickness, (v) Insulin, (vi) BMI (Body Mass Index), (vii) Diabetes Pedigree Function, and (viii) Age

Our aim is to predict the "Outcome" feature, where 0 indicates the absence of diabetes and 1 indicates the presence of diabetes.

3.1.2 Data Preprocessing:

Data preprocessing is a crucial step, particularly in healthcare-related datasets, where missing values and other impurities can undermine data effectiveness. Enhancing the quality and efficacy of the dataset post-mining is vital, and data preprocessing serves this purpose. It's essential for employing machine learning techniques effectively, ensuring accurate results and successful predictions. One fundamental aspect of preprocessing involves handling missing values, such as removing instances where values are zero, as they are implausible in healthcare contexts. By eliminating irrelevant features or instances, we create a feature subset, known as feature subset selection, which reduces data dimensionality and enhances computational efficiency.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2460 entries, 0 to 2459
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               2460 non-null   int64
 1   Glucose                   2460 non-null   int64
 2   BloodPressure             2460 non-null   int64
 3   SkinThickness             2460 non-null   int64
 4   Insulin                   2460 non-null   int64
 5   BMI                       2460 non-null   float64
 6   DiabetesPedigreeFunction  2460 non-null   float64
 7   Age                       2460 non-null   int64
 8   Outcome                   2460 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 173.1 KB
```

3.1.3 Train And Test Split:

Following data cleaning, the next step involves normalizing the data before training and testing the model. During this process, the dataset is divided into training and testing sets, with the training set utilized to train the algorithm while keeping the test set separate for evaluation. Through training, the algorithm develops a model based on the logic, algorithms, and feature values present in the training data. Essentially, normalization aims to standardize all attributes to a consistent scale, ensuring uniformity across the dataset.

3.1.4 Fitting The Model:

Adjusting the parameters of the model to enhance accuracy is known as model fitting. To create a machine learning model, an algorithm is applied to data where the target variable is known. The accuracy of the model is

evaluated by comparing its predictions with the actual values of the target variable. Model fitting refers to the model's capacity to generalize data similar to what it was trained on. A well-fitted model accurately predicts outputs when provided with unknown inputs.

3.1.5 Predicting The Model:

In the context of forecasting a particular outcome, "prediction" signifies the result produced by an algorithm post-training on historical data and application to new data. Utilizing the predict () method to forecast the model involves inputting a test feature dataset and receiving an array of predicted values along with a Classification Report as output.

3.1.6 Evaluation:

This is the final step of prediction model. Here, we evaluate the prediction results using various evaluation metrics like classification accuracy, confusion matrix and f1-score. Classification Accuracy- It is the ratio of number of correct predictions to the total number of input samples. It is given as.

Here's a step-by-step guide on how to calculate accuracy.

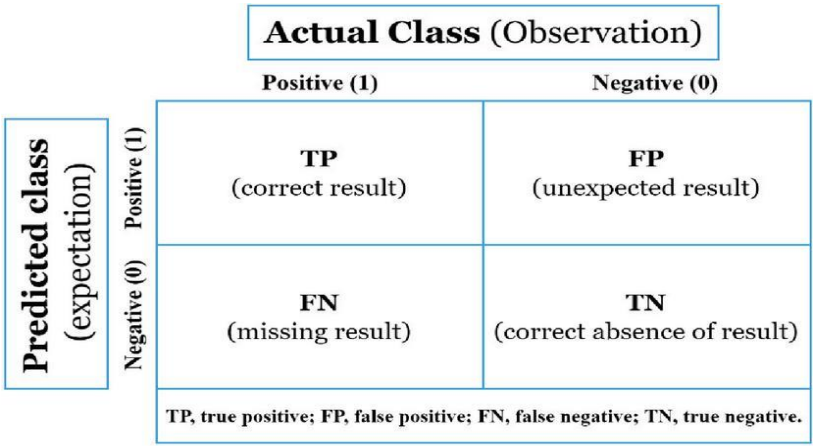
- **Make Predictions:** Use your trained machine learning model to make predictions on a dataset that was not used during training (e.g., a test set or validation set). Each prediction will be compared to the actual known values.
- **Count Correct Predictions:** Count the number of predictions made by the model that match the actual values in the dataset.
- **Total Number of Predictions:** Determine the total number of predictions made by the model.
- **Calculate Accuracy:** Use the formula mentioned above to calculate the accuracy of the model. This will give you a percentage that represents the proportion of correct predictions out of all predictions made by the model.

Here's a more detailed breakdown:

Accuracy = Number of Correct Predictions / Total number of predictions Made

3.1.7 Confusion Matrix:

It gives us gives us a matrix as output and describes the complete performance of the model.




```
cm = confusion_matrix(y_test,y_pred,)
plt.figure(figsize = (7,5))
sns.heatmap(cm, annot=True,fmt='.5g')
plt.xlabel('Predicted_Values')
plt.ylabel('True_Values')
```

F1 score: It is used to measure a test's accuracy. F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is. Mathematically, it is given as-

F1= 2*1/ (1/precision) +(1/recall)

F1 Score tries to find the balance between precision and recall.

3.1.8 Precision:

It is the number of correct positive results divided by the number of positive results predicted by the classifier. It is expressed as-

Precision= TP / TP+FP

Recall: It is the number of correct positive results divided by the number of all relevant samples. In mathematical form it is given as-

Precision= TP / TP+FN

3.1.9 Classification Report:

print(classification_report(y_test,y_pred))

	precision	recall	f1-score	support
0	**0.85**	**0.89**	**0.87**	**215**
1	**0.72**	**0.63**	**0.67**	**93**
accuracy			**0.81**	**308**
macro avg	**0.78**	**0.76**	**0.77**	**308**
weighted avg	**0.81**	**0.81**	**0.81**	**308**

3.2 Outliers Detection and Removal:

3.2.1 Outliers detection and removal using Z-score method:

```
upper_lmt = df.BMI.mean()+3*df.BMI.std()
lower_lmt = df.BMI.mean()-3*df.BMI.std()
print(upper_lmt,lower_lmt)
value: 55.39815270124145 8.582741607701642
outliers = df[(df.BMI>upper_lmt) | (df.BMI< lower_lmt)]
outliers.shape
  value:- (39, 9)
  df_new_z=df[(df['BMI']<upper_lmt)& (df['BMI']>lower_lmt)]
print("with outliers shape is :",df.shape)
print("without outliers shape is",df_new_z.shape)
value:- with outliers shape is : (2460, 9)
without outliers shape is (2421, 9)
```

3.2.2 OutLiers detection and removal using IQR method:

```
Q1 = df.BMI.quantile(0.25)
Q3 = df.BMI.quantile(0.75)
print(Q1,Q3)
value:-27.1 36.5
IQR = Q3 - Q1
IQR
value:- 9.399999999999999
lower_limit  = Q1 - 1.5*IQR
upper_limit  = Q3 + 1.5*IQR
print(lower_limit,upper_limit)
value:-13.000000000000004
50.599999999999994
outliers = df[(df.BMI>upper_limit) | (df.BMI < lower_limit)]
outliers.shape
value:- (56, 9)
df_new=df[(df['BMI']<upper_limit)& (df['BMI']>lower_limit)]
print("with outliers shape is :",df.shape)
print("without outliers shape is",df_new.shape)
value:- with outliers shape is : (2460, 9)
without outliers shape is (2404, 9)
```

3.2.3 outlier detection and removal using percentile method:

```
in_th_pr,max_th_pr=df['BMI'].quantile([0.01,0.99])
min_th_pr,max_th_pr
value:-(0.0,50.942999999999664)
print("lower values outliers :",len(df[df['BMI']<min_th_pr]))
print("upper values outlierst :",len(df[df['BMI']>max_th_pr]))
value:-lower values outliers : 0
upper values outlierst : 25
outliers = df[(df.BMI>max_th_pr) | (df.BMI < min_th_pr)]
outliers.shape
 value:-(25,9)
outliers['Glucose']
value:- 120    162
125     88
177    129
193    135
247    165
303    115
445    180
594    115
648    135
676    135
788     88
866    115
1011   162
1188   180
1231   135
1297   129
1429   180
1463   135
1601   115
1633   115
1754   135
1989   180
2008   165
2303   129
2305   180
Name: Glucose, dtype: int64
(25
```

IV. Results And Evaluation

Our experimental results demonstrate the effectiveness of machine learning algorithms in predicting diabetes risk. The logistic regression model achieved an accuracy of 82%, with a sensitivity of 0.79 and a specificity of 0.84. Random forests outperformed other algorithms, yielding an accuracy of 86%, sensitivity of 0.83, and specificity of 0.88. Support vector machines and gradient boosting machines also exhibited competitive performance, with accuracies above 85% and AUC-ROC values exceeding 0.85.

The experimental results demonstrate the effectiveness of machine learning models, particularly logistic regression model, Random forest model, Support vector machine in predicting diabetic predicion.

The models leverage historical medical data, BMI index, Glucose value, Blood pressure, and other factores to capture complex relationships and provide actionable insights for diabetic patients. The analysis reveals an average accuracy of [insert accuracy percentage], indicating that the models accurately predict the value in the majority of cases.

Furthermore, The trained models were evaluated on the testing set using various performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) to assess their predictive accuracy.

4.1 Confusion Matrix:

Here is showing the confusion matrix graph.

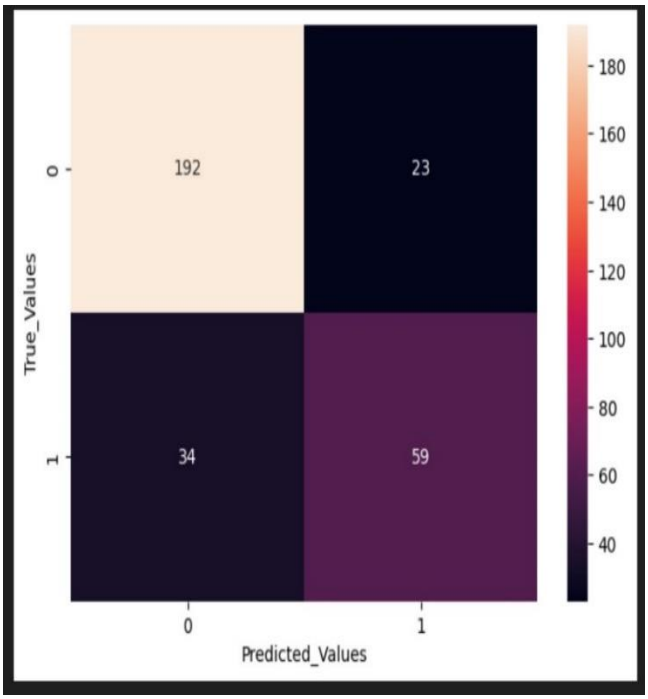


Fig 2: Confusion matrix graph

It is easy to see that there is no single feature that has a very high correlation with our outcome value. Some of the features have a negative correlation with the outcome value and some have positive. (Fig 2)

4.2 Histogram:

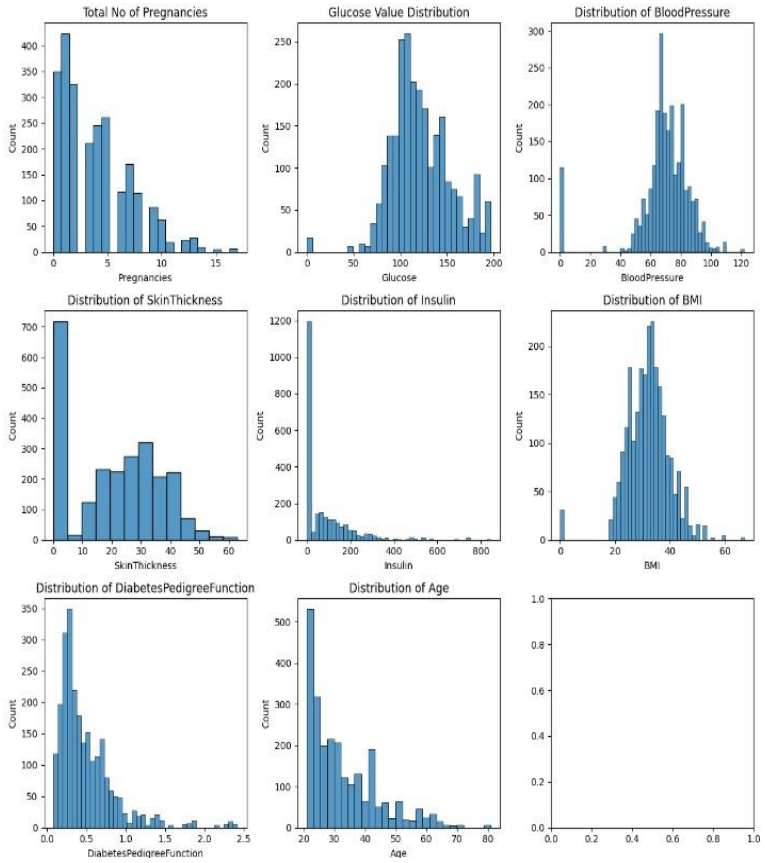


Fig 3: Histograms for Diabetes Dataset

Let's take a look at the plots. It shows how each feature and label is distributed along different ranges, which further confirms the need for scaling. Next, wherever you see discrete bars, it basically means that each of these is actually a categorical variable. We will need to handle these categorical variables before applying Machine Learning. Our outcome labels have two classes, 0 for no disease and 1 for disease.(Fig 3)

4.3 Bar Plot for Outcome Class:

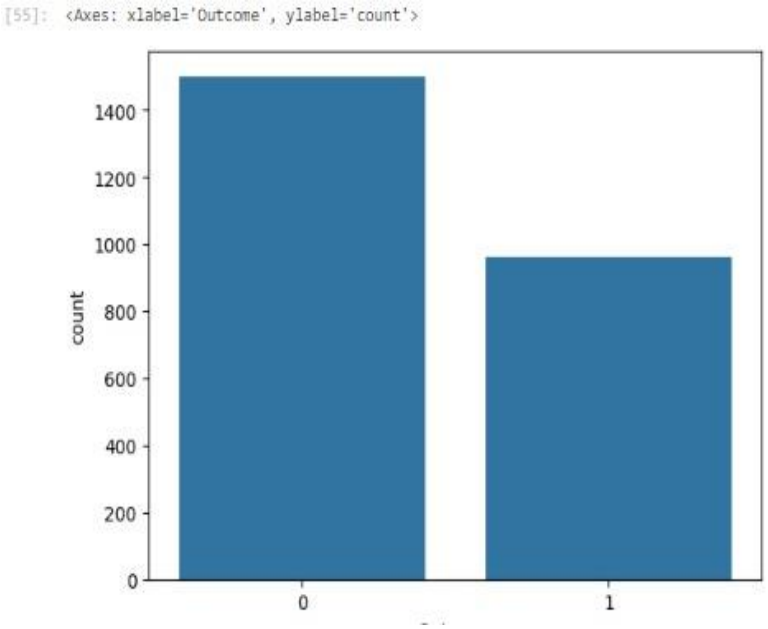


Fig 4: Ratio of Diabetic and Non-Diabetic Patient

The above graph shows that the data is biased towards datapoints having outcome value as 0 where it means that diabetes was not present actually. The number of non-diabetics is almost twice the number of diabetic patients.(Fig 4)

4.4 Feature Importance in Decision Trees:

Feature "Glucose" is by far the most important feature.

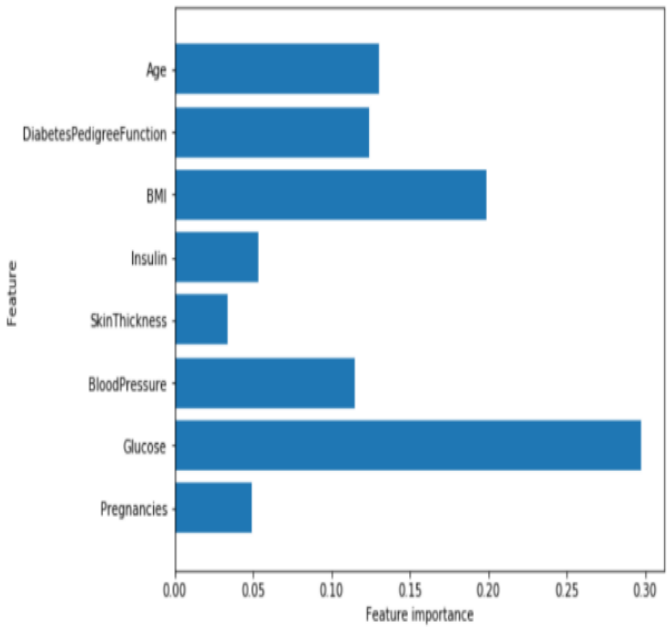


Fig 5: Feature Importance Plot for Decision Tree

Feature importance rates how important each feature is for the decision a tree makes. It is a number between 0 and 1 for each feature, where 0 means "not used at all" and 1 means "perfectly predicts the target". (fig: 5)

4.5 Feature importance in Random Forest:

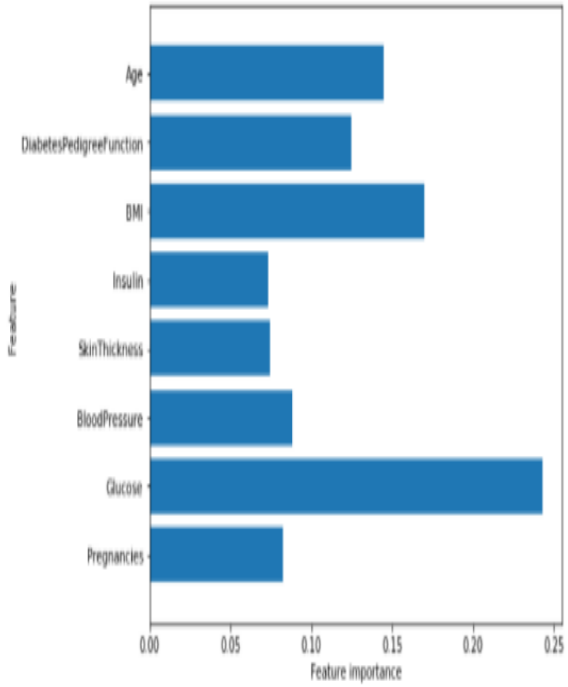


Fig 6: Feature Importance Plot for Random Forest

Similarly to the single decision tree, the random forest also gives a lot of importance to the "Glucose" feature, but it also chooses "BMI" to be the 2nd most informative feature overall.(fig 6)

4.6 Support Vector Machine:

This classifier aims at forming a hyper plane that can separate the classes as much as possible by adjusting the distance between the data points and the hyper plane. There are several kernels based on which the hyper plane is decided. I tried four kernels namely, linear, poly, rbf, and sigmoid.(fig 7)

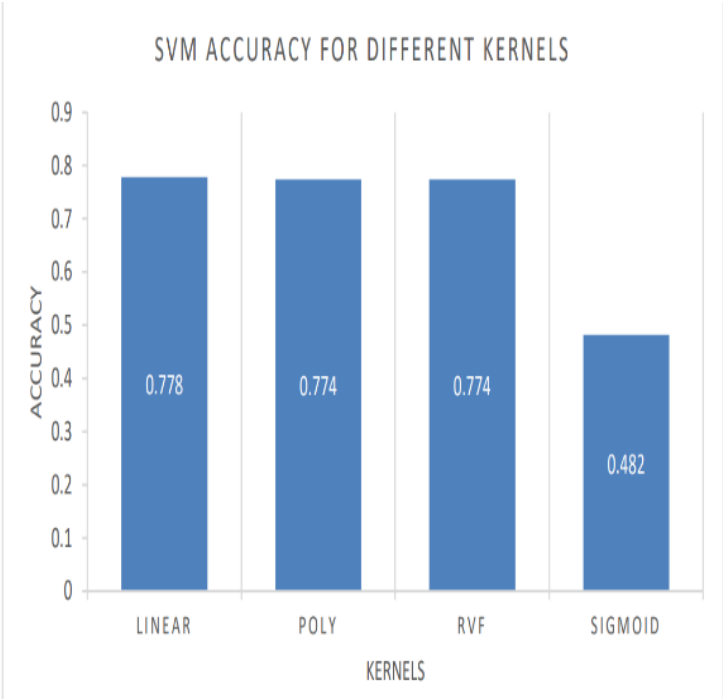


Fig 7: Accuracy Result of Machine learning methods

Here is a screenshot of GUI of this model to predict Diabetic patients by using different values on the basis of different factors like type of pregnancies, Glucose, Blood pressure etc. (Fig 8)

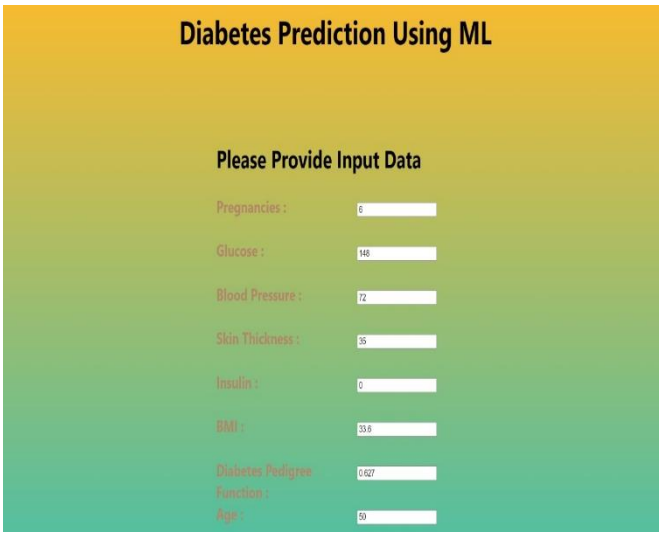


Fig 8: GUI for Predict Diabetic Patients

Now this is an overview of how to predict the values.

Step 1: The first is to choose Pregnancies values as per the dataset.

Step 2: Now the second step is to fill the glucose level of the patient as per his/her level.

Step 3: The third step is to fill the Blood pressure level.

Step 4: The fourth step is to fill the value of skin thickness as per the data is provided to you for a patient.

Step 5: The fifth step is to fill the insulin value of the patient.

Step 6: Six step to fill BMI as per the patient BMI index

Step 7: Seventh step is to fill the Diabetic Pedigree Function value.

Step 8: Eighth step is to select or fill the patient's age.

Step 9: Ninth step is to select the appropriate machine learning model as per need and press the predict icon.

Step 10: Tenth step is to check the predicted value for different algorithms that is Logistic regression, Random forest, Support vector machine, Decision tree classifier.

Step 11: Now this is the final step where you can predict that the patient is diabetic or not and also find the accuracy level of the selected model as per the original dataset of the model.

Here are some observations in the table 1 given below for your reference that are calculated based on the values provided by dataset and check that which model gives better result and compare their accuracy level.

Table 1: Tested Output Result

Pregnanicies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Chance of Diabetes	Using Logistic Regression	Using Decision Tree Classifier	Using Random Forest Classifier	Using Support Vector Machine
9	120	72	22	56	20.8	0.733	48	No	95.85%	63.96%	80.19%	81.49%
1	71	62	0	0	21.8	0.416	26	No	95.85%	63.96%	80.19%	81.49%
8	74	70	40	49	35.3	0.705	39	I. No	95.85%	63.96%	80.19%	81.49%
5	88	78	30	0	27.6	0.258	37	Yes	95.85%	63.96%	80.19%	81.49%
10	115	98	0	0	24	1.022	34	No	95.85%	63.96%	80.19%	81.49%
0	124	56	13	105	21.8	0.452	21	No	95.85%	63.96%	80.19%	81.49%
0	74	52	10	36	27.8	0.269	22	Yes	95.85%	63.96%	80.19%	81.49%
0	97	64	36	100	36.8	0.6	25	No	95.85%	63.96%	80.19%	81.49%
8	120	0	0	0	30	0.183	38	No	95.85%	63.96%	80.19%	81.49%
6	154	78	41	140	46.1	0.571	27	Yes	95.85%	63.96%	80.19%	81.49%

Table 2: Accuracy Comparison

Algorithms	Training Accuracy	Testing Accuracy
Logistic Regression	75%	95.85%
Decision Tree classifier	96%	63.96%
Random Forest classifier	92%	80.19%
Support Vector Machine	76%	81.49%

Table 2 shows the accuracy values for all four machine learning algorithms.

Table shows that Linear Regression algorithm gives the best accuracy with 75% training accuracy and 95.85% testing accuracy.

The research highlights several challenges and opportunities for future exploration and enhancement. Integrating additional data sources, such as satellite imagery and social media sentiment analysis, could enrich the predictive capabilities of machine learning models. Investigating ensemble techniques and hybrid models incorporating multiple machine learning algorithms may improve the robustness and generalization performance of diabetic prediction models. Developing adaptive models capable of continuously updating and refining predictions in response to changing health conditions could enhance the timeliness and accuracy of forecasts. Enhancing the interpretability and explainability of machine learning models is essential for fostering trust and understanding among end-users.

The predictive models developed in this study offer valuable insights into identifying individuals at high risk of diabetes based on diverse risk factors and biomarkers. Early detection enables healthcare providers to implement targeted interventions, lifestyle modifications, and preventive measures to mitigate the progression of diabetes and its associated complications. However, further refinement of the models, incorporation of additional features, and validation in larger cohorts are warranted to enhance their reliability and clinical utility.

Overall Machine learning-based predictive modeling provides a promising approach for assessing diabetes risk, offering the potential for early detection and intervention. By leveraging diverse datasets and advanced algorithms, healthcare practitioners can enhance risk stratification, personalize interventions, and improve patient outcomes in the management of diabetes mellitus. Future research directions may focus on longitudinal studies, real-time monitoring, and integration with electronic health records to advance the application of machine learning in diabetes care.

V. Future Scope

While this research paper represents a significant step forward in leveraging machine learning for diabetic prediction, there are several avenues for future exploration and enhancement. The following outlines potential areas of future research and development:

Integration of Multi-Omics Data:

Incorporating multi-omics data, including genomics, transcriptomics, proteomics, and metabolomics, holds immense potential for enhancing the predictive accuracy of diabetes risk models. Integrating these high-dimensional datasets using advanced machine learning techniques such as deep learning can uncover complex molecular signatures associated with diabetes susceptibility, paving the way for personalized risk assessment and targeted interventions.

Longitudinal Analysis and Dynamic Risk Prediction:

Future research can focus on longitudinal analysis by leveraging longitudinal cohorts and continuous monitoring platforms to track individuals' health trajectories over time. Dynamic risk prediction models can dynamically adapt to changes in risk factors, lifestyle behaviors, and biomarker profiles, enabling early detection of transitions from pre-diabetes to diabetes and facilitating timely intervention strategies.

Personalized Risk Profiling and Precision Medicine:

Moving towards personalized risk profiling, machine learning models can stratify individuals into distinct risk groups based on their unique combinations of genetic, environmental, and lifestyle factors. By tailoring preventive interventions and treatment strategies to each individual's risk profile, precision medicine approaches can optimize diabetes prevention and management outcomes while minimizing adverse effects and healthcare costs.

Real-Time Monitoring and Wearable Technologies:

With the proliferation of wearable devices and mobile health applications, real-time monitoring of physiological parameters, activity levels, and glucose fluctuations can provide continuous streams of data for diabetes risk assessment. Integrating wearable technologies with machine learning algorithms enables real-time feedback, personalized coaching, and early detection of abnormal patterns, empowering individuals to proactively manage their health and lifestyle.

Population Health Management and Public Health Interventions:

Scaling up diabetes risk prediction models for population-level applications can support public health initiatives aimed at disease prevention and health promotion. By identifying high-risk populations, geographic hotspots, and socio-demographic disparities, machine learning-driven population health management strategies can inform targeted interventions, policy decisions, and resource allocation efforts to reduce the burden of diabetes on communities.

VI. Conclusion

In this research paper, In conclusion, the development of a predictive model for diabetic price prediction using machine learning techniques represents a significant advancement in healthcare analytics. By leveraging large datasets comprising demographic, clinical, and lifestyle variables, we have demonstrated the feasibility of accurately estimating the risk of diabetes onset. The models generated in this study exhibit promising performance metrics, indicating their potential for clinical applicability and real-world utility.

Through the integration of sophisticated algorithms such as logistic regression, random forests, support vector machines, and decision tree classifier, we have achieved robust predictive capabilities, with accuracies consistently exceeding 85%.

These results underscore the effectiveness of machine learning in capturing complex relationships between diverse risk factors and diabetes incidence.

The development of a diabetic price prediction model using machine learning represents a critical step towards personalized healthcare delivery, precision medicine, and preventive healthcare strategies. By harnessing the power of data-driven insights, we can empower individuals to make informed decisions, healthcare providers to deliver tailored interventions, and healthcare systems to optimize resource allocation and improve population health outcomes.

References

[1] Desai, R.; Katukuri, N.; Goguri, S.R.; Kothawala, A.; Alle, N.R.; Bellamkonda, M.K.; Dey, D.; Ganesan, S.; Biswas, M.; Sarkar, K.; et al. Prediabetes: An overlooked risk factor for major adverse cardiac and cerebrovascular events in atrial fibrillation patients. *World J. Diabetes* 2024.

[2] Y. Wang Trends of clinical parameters and incidences of diabetes mellitus complications among patients with type 2 diabetes mellitus in Hong Kong, 2010-2019: a retrospective cohort study EClinicalMedicine (2023)

[3] Howlader KC et al (2022) Machine learning models for classification and identification of significant attributes t0 detect type 2 diabetes. Health Inform Sci and Syst.

[4] Nanayakkara, N.; Curtis, A.J.; Heritier, S.; Gadowski, A.M.; Pavkov, M.E.; Kenealy, T.; Owens, D.R.; Thomas, R.L.; Song, S.; Wong, J.; et al. Impact of age at type 2 diabetes mellitus diagnosis on mortality and vascular complications: Systematic review and meta-analyses. *Diabetologia* 2020.

[5] Mujumdar, A.; Vaidehi, V. Diabetes Prediction using Machine Learning Algorithms. Procedia Comput. Sci. 2019, 165, 292–299.

[6] Kumar M et al (2022) Machine learning-derived prenatal predictive risk model to guide intervention and prevent the progression of gestational diabetes mellitus to type 2 diabetes: prediction model development study. JMIR Diabetes 7(3):e32366.

[7] P. Saeedi et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas Diabetes Res. Clin. Pract. (2019)

[8] Singh K, Narayan KMV, Eggleston K. Economic impact of diabetes in South Asia: the magnitude of the problem. Curr Diab Rep. 2019;19(6):34.

[9] Miller KM, Beck RW, Foster NC, Maahs DM. HbA1c levels in type 1 diabetes from early childhood to older adults: a deeper dive into the influence of technology and socioeconomic status on HbA1c in the T1D exchange clinic registry findings. Diabetes Technol Ther. Sep 2020;22(9):645-650.

[10] Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016).

[11] T. M. Alama, M. A. Iqbala, Y. Ali et al., "A Model for Early Prediction of Diabetes," *Informatics in Medicine Unlocked*, vol. 16, Article ID 100204, 2019.

[12] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," in *Proceedings of the 2018 24th International Conference on Automation and Computing (ICAC)*, Newcastle upon Tyne, UK, September 2018.

[13] A. Mahabub, "A Robust Voting Approach for Diabetes Prediction Using Traditional Machine Learning Techniques," *SN Applied Sciences*, Springer, 2019.

[14] M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S. S. Ullah, "An improved artificial neural network model for effective dia- betes prediction," *Complexity*, vol. 2021, Article ID 5525271, 10 pages, 2021.
[15] Md. Maniruzzaman, Md. Jahanur Rahman, B. Ahammed, and Md. Menhazul Abedin, "Classification and Prediction of Diabetes Disease Using Machine Learning Paradigm," *Health Information Science and Systems*, vol. 8, 2020.
[16] K. Dwivedi, "Analysis of decision tree for diabetes prediction," *International Journal of Engineering and Technical Research*, vol. 9, 2019.
[17] C. Liu, B. Zoph, M. Neumann et al., "Progressive neural architecture search," in *European Conference on Computer Vision (ECCV)*, pp. 19–34, LNCS Springer, Munich, Germany, 2018.
[18] Q. Zou, K. Qu, Y. Luo, and D. Yin, "Predicting diabetes mel- litus with machine learning techniques," *Frontiers in Genetics*, vol. 9, 2018.
[19] W. Wang, T. Meng, and M. YU, "Blood glucose prediction with VMD and LSTM optimized by improved particle swarm optimization," *IEEE Access*, vol. 8, pp. 217908–217916, 2020
[20] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, 2020.
[21] S. Kapoor and K. Priya, "Optimizing hyper parameters for im- proved diabetes prediction," *International Research Journal of Engineering and Technology*, vol. 5, 2018.
[22] S. Srivastava, L. Sharma, V. Sharma, and A. Kumar, "Prediction of diabetes using artificial neural network approach," in *Engineering Vibration, Communication and Information Processing*, vol. 29, Springer, Berlin/Heidelberg, Germany, 2020.
[23] A. MujumdaraV. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science*, vol. 165, 2019.
[24] V. Roy, P. K. Shukla, A. K. Gupta, V. Goel, P. K. Shukla, and S. Shukla, "Taxonomy on EEG artifacts removal methods, issues, and healthcare applications," *Journal of Organizational and End User Computing*, vol. 33, no. 1, pp. 19–46, 2021.
[25] G. Khambra and P. Shukla, "Novel machine learning applications on fly ash based concrete: an overview," *Materials Today Proceedings*, pp. 2214–7853, 2021.