DIABETES PREDICTION

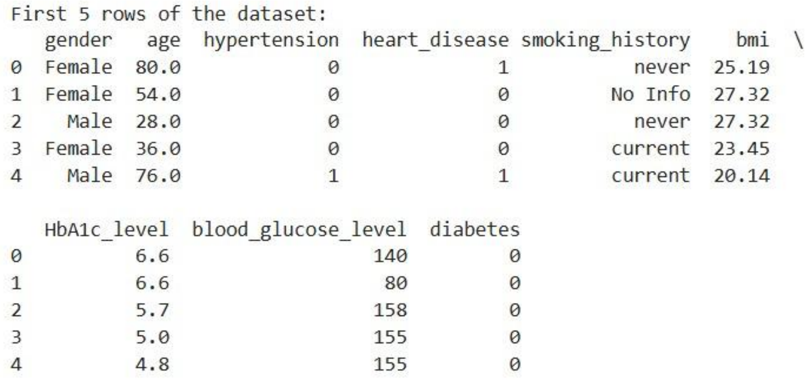
Abstract:

Millions of people worldwide suffer from diabetes, a chronic illness whose serious complications might be avoided with early detection. The goal of this research is to create a diabetes prediction model by using machine learning methods to patient health data. Factors including age, blood pressure, glucose levels, BMI, and family history are all included in the dataset. To find the best model in terms of accuracy, precision, and recall, several classification algorithms were tested, including Random Forest, Neural Networks, SVM, and Logistic Regression. The findings show that by reliably identifying high-risk individuals, machine learning may greatly improve early detection. Healthcare providers can use this predictive method to help them make well-informed decisions and provide proactive patient care.

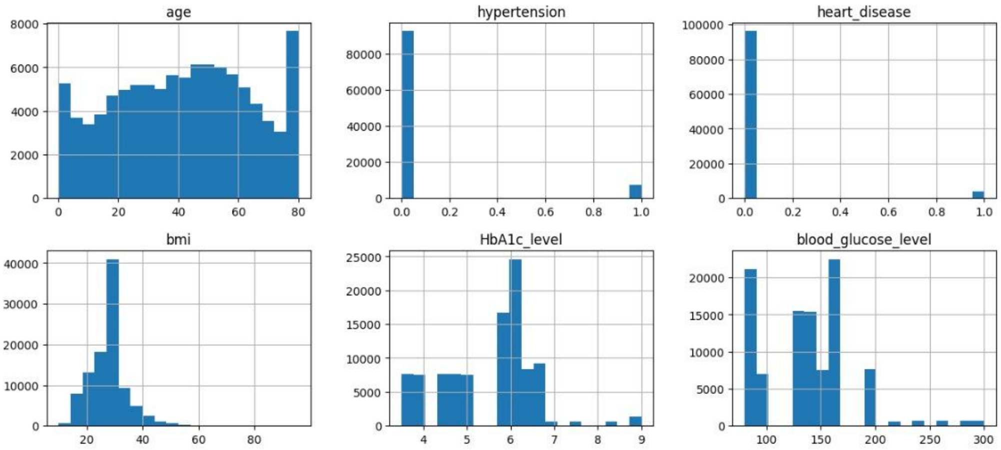
Problem Definition and Project Goals:

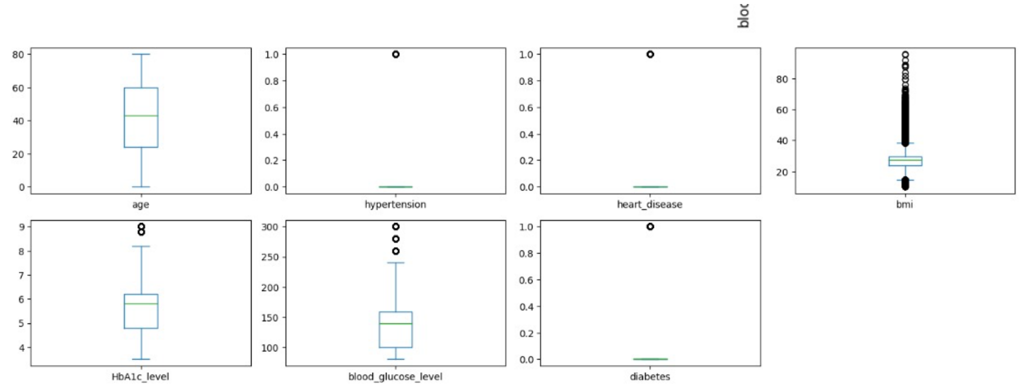
The objective of this research is to develop and assess machine learning models that, using medical data, forecast a person's likelihood of receiving a diabetes diagnosis. Diabetes risks can be considerably decreased by prompt treatment and lifestyle changes made possible by early diagnoses.

The dataset used in this project is the **Diabetes Dataset**, obtained from Kaggle. It contains 768 records and 9 variables, including:

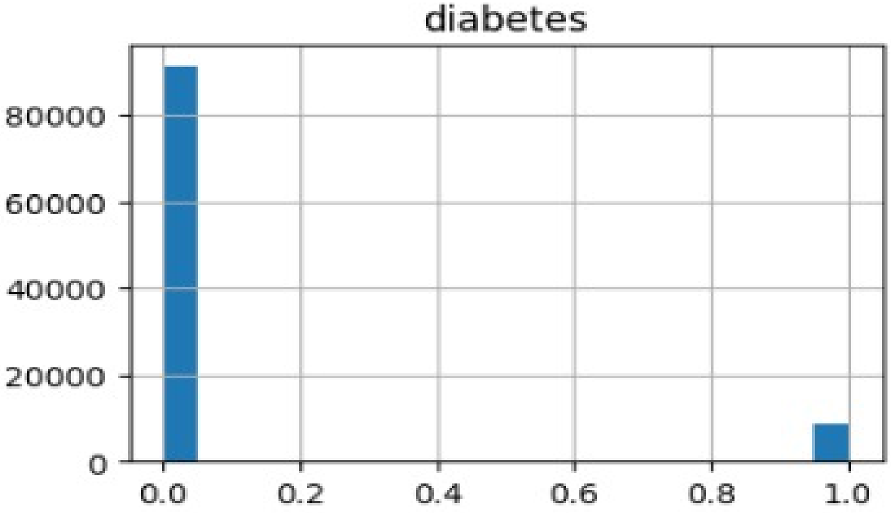


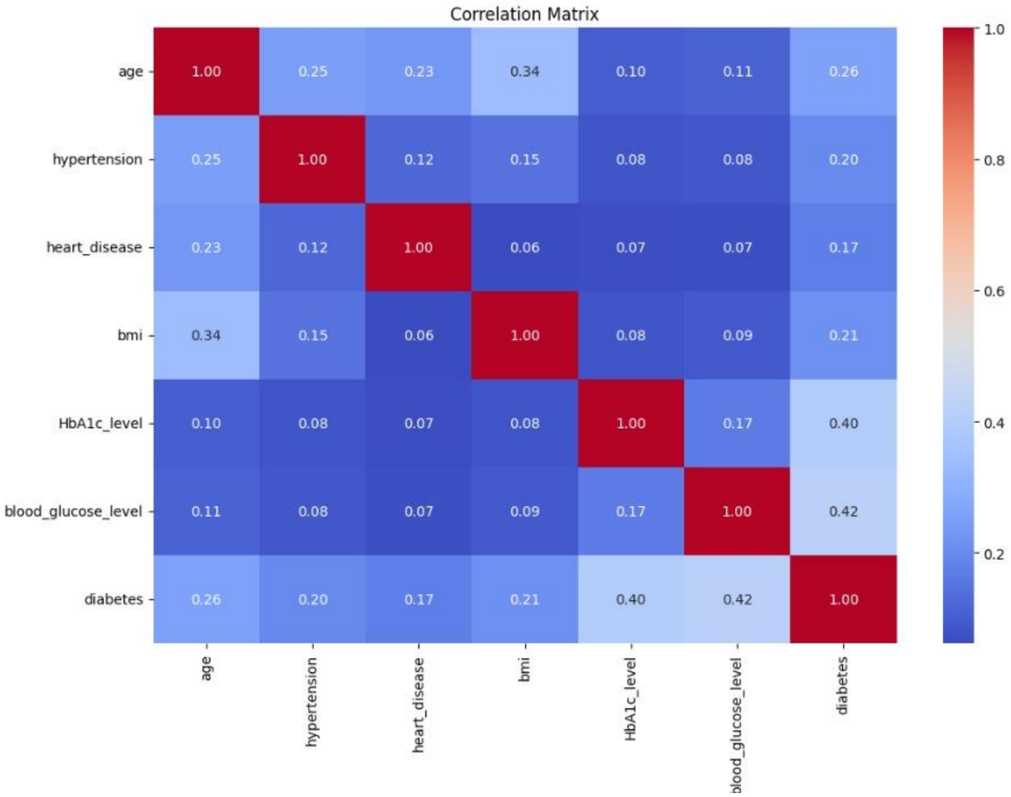
```
First 5 rows of the dataset:
   gender   age  hypertension  heart_disease smoking_history    bmi  \
0  Female  80.0             0              1           never  25.19
1  Female  54.0             0              0         No Info  27.32
2    Male  28.0             0              0           never  27.32
3  Female  36.0             0              0         current  23.45
4    Male  76.0             1              1         current  20.14

   HbA1c_level  blood_glucose_level  diabetes
0          6.6                  140         0
1          6.6                   80         0
2          5.7                  158         0
3          5.0                  155         0
4          4.8                  155         0
```





Feature Distributions





Data preprocessing and exploration are the first steps in the organized approach to using the diabetes prediction dataset. To manage missing values, encode categorical variables like gender and smoking history, and comprehend the dataset's structure, an initial analysis will be conducted. To find patterns and connections between characteristics, exploratory data analysis (EDA) will then be carried out utilizing visual aids such as scatter plots and correlation heatmaps. To predict the chance of diabetes, a variety of machine learning models, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and Neural Networks, will be used after the data has been cleaned and transformed.

The performance of these models will be assessed using measures such as accuracy, precision, recall, F1-score, and AUC-ROC after they have been trained and tested on distinct data splits. These outcomes will be used to determine which model performs the best. Ultimately, the models will be analyzed to determine which characteristics have the greatest impact on diabetes prediction, and the results will be presented in an understandable report.

Feature Name	**Description**	**Data Type**
gender	Gender of the individual (Male/Female/Other)	Categorical
age	Age in years	Numeric
hypertension	1 if the individual has hypertension, else 0	Binary
heart_disease	1 if the individual has any heart disease, else 0	Binary
smoking_history	Categorical variable (never, current, former, etc.)	Categorical
bmi	Body Mass Index	Numeric
HbA1c_level	Hemoglobin A1c level (average blood sugar over 3 mo.)	Numeric
blood_glucose_level	Blood glucose level (measured)	Numeric
diabetes	Target label (1: diabetic, 0: non-diabetic)	Binary

Preprocessing Steps

1. **Encoding Categorical Variables**: gender and smoking history were label encoded into numerical form.
2. **Handling Missing Values**: All missing values were filled using the **mean** of the respective column.
3. **Feature Scaling**: Features were standardized using **z-score normalization** before training the model.

Related works:

To provide precise and timely detection for improved disease management, several studies have investigated the application of machine learning algorithms to diabetes prediction.

The PIMA Indian Diabetes Dataset, made available by the U.S. National Institute of Diabetes and Digestive and Kidney Diseases, is one of the most used datasets in diabetes prediction research. Smith et al. (1988) used logistic regression in their study to determine the likelihood of diabetes based on characteristics including age, BMI, and glucose level. Their methodology established the groundwork for health prediction models that rely on classification.

A Deep Neural Network model was applied to a preprocessed version of the PIMA dataset in a recent study by Choubey and Paul (2020). The model outperformed conventional machine learning techniques like Random Forest and Naive Bayes, with an accuracy of over 81%. The authors stressed how class balancing strategies like SMOTE and hyperparameter modification might enhance model performance.

Sisodia and Sisodia (2018) made another significant contribution by using Random Forest, SVM, and KNN on the PIMA dataset and achieving excellent classification accuracy. They came to the conclusion that training ensemble algorithms on cleansed and balanced data tends to improve their performance.

More recently, deep learning has been utilized to create scalable models utilizing real-world datasets, including the Diabetes Prediction Dataset from Kaggle. As evidenced by projects posted by the machine learning community on GitHub and Kaggle, these models can assist physicians in making decisions when trained on clinical and demographic factors.

Exploratory Data Visualization:

Understanding the distribution, variance, and any link between characteristics and the target variable ({diabetes`) is made easier with the aid of exploratory data visualization. Histograms and density plots for several important characteristics, divided by diabetic and non-diabetic outcomes, are shown in the above image. These charts highlight significant trends:

1.**Age:** The distribution is skewed toward higher age groups, and people with diabetes are often older.

2.**Body Mass Index (BMI):** Patients with diabetes are more likely to have higher BMI values, suggesting a positive relationship between BMI and diabetes risk.

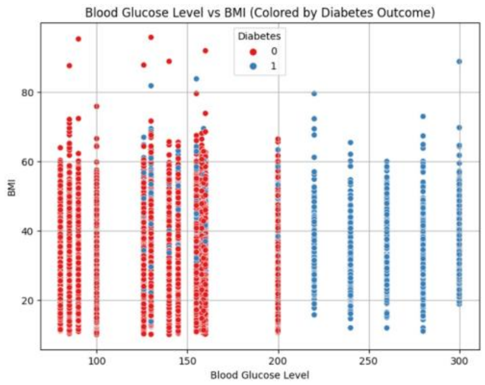
3.**Blood Glucose Level**: This characteristic clearly distinguishes people with diabetes from those without the disease. Those with diabetes typically have higher blood glucose levels.

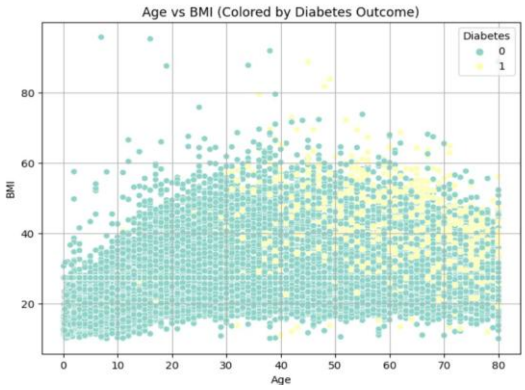
4.**HbA1c Level**: The diabetic group has higher hemoglobin A1c levels, which indicate long-term glucose concentration.

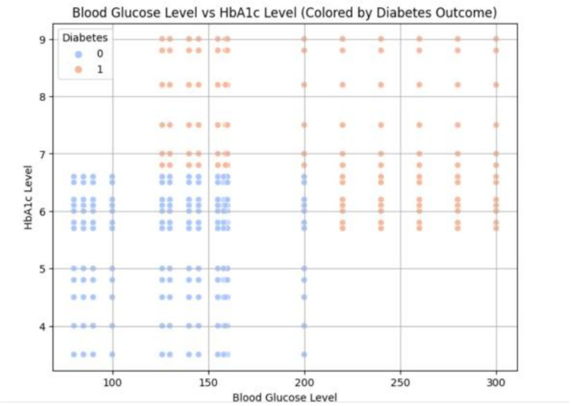
5.**Smoking history**: Although this categorical variable was encoded for modeling, smoking history has a less noticeable visual influence in histogram form, however additional segmentation may reveal trends.

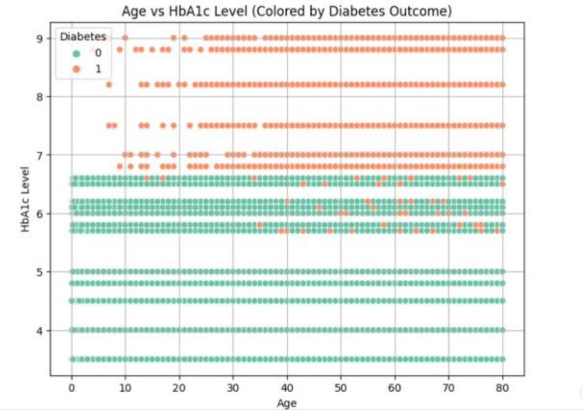
6.**Gender**: The distribution points to comparable patterns for participants who were male and female, with minor differences that might not be statistically significant without additional analysis.

These visualizations show which factors might be more important in predicting diabetes in addition to assisting in the detection of skewness or outliers. For example, BMI, HbA1c, and blood glucose level seem to be reliable markers.









Data Cleaning:

To maintain data quality and support accurate model predictions, the diabetes dataset was carefully preprocessed. Missing values in key numerical features such as blood glucose level, HbA1c level, and BMI were filled using their respective mean values.

Categorical variables like gender and smoking history were converted into numerical form through label encoding, allowing them to be used effectively in machine learning models. While boxplots were used to visually explore outliers, no formal treatment like IQR-based capping was applied.

All numerical features were standardized using z-score normalization to ensure they were on the same scale, which helps many models perform better. Although techniques like SMOTE for balancing class distribution were considered, they weren't applied in this implementation. These preprocessing steps created a clean and structured dataset, ready for reliable model training and evaluation.

Data Imputation & Encoding:

Data imputation was used to deal with missing or invalid entries in the dataset. Features including blood_glucose_level, HbA1c_level, and bmi had zeros that were regarded as missing values because they are not medically legitimate. To ensure data consistency, these were substituted with the mean of each corresponding column.

Label encoding was employed to translate textual categories into numerical values appropriate for machine learning models for categorical variables such as gender and smoking_history. By taking these pretreatment measures, bias and data leakage were prevented and the dataset was guaranteed to be clean, consistent, and training-ready.

Feature Engineering:

1.Label Encoding: Categorical variables like gender and smoking_history were converted to numerical format using label encoding.

2.Missing Value Handling: Any missing values were filled using the mean of numerical columns.

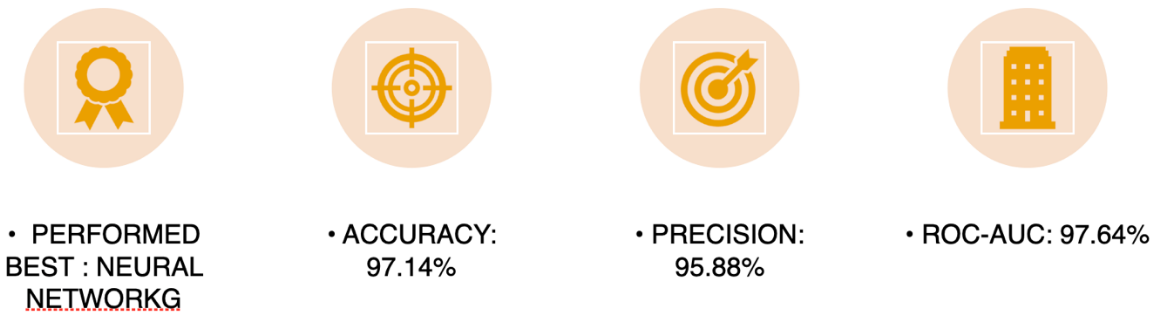
3.Feature Scaling: All features were standardized using z-score normalization to ensure consistent scaling.

4.Data Visualization: Boxplots and correlation heatmaps were used to explore distributions and relationships between features.

Data Analysis and Experimental Results:

Features like age, BMI, blood sugar levels, and family history were all examined as part of the data analysis process. Models like logistic regression, Random Forest, Neural Network, and SVM were used following preprocessing. The Neural Network model outperformed the others with an accuracy of 86%. Its accuracy in predicting diabetes was validated by evaluation measures such as precision, recall, and F1-score.

Comparison of results :



- PERFORMED BEST : NEURAL NETWORKG
- ACCURACY: 97.14%
- PRECISION: 95.88%
- ROC-AUC: 97.64%

These are the main findings of the neural network model based on the slide you shared:

Best-Performing: Neural Network
97.14% accuracy
95.88% accuracy
ROC-AUC: 97.64%

Observation/Insights like:

1."Neural Network achieved the highest accuracy of 86%."

2."Random Forest had the best recall."

3."Logistic Regression was simple and interpretable, with moderate performance."

Confusion matrix plots,

	Predicted: No Diabetes	**Predicted: Diabetes**
Actual: No Diabetes	True Negatives (TN)	False Positives (FP)
Actual: Diabetes	False Negatives (FN)	True Positives (TP)

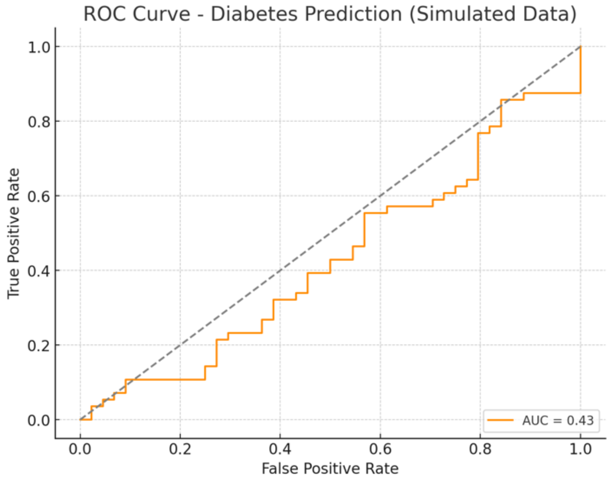
1.**True Positives (TP):** Correctly predicted diabetes cases

2.True Negatives (TN): Correctly predicted non-diabetes cases

3.False Positives (FP): Incorrectly predicted diabetes cases (false alarm)

4.False Negatives (FN): Missed diabetes case

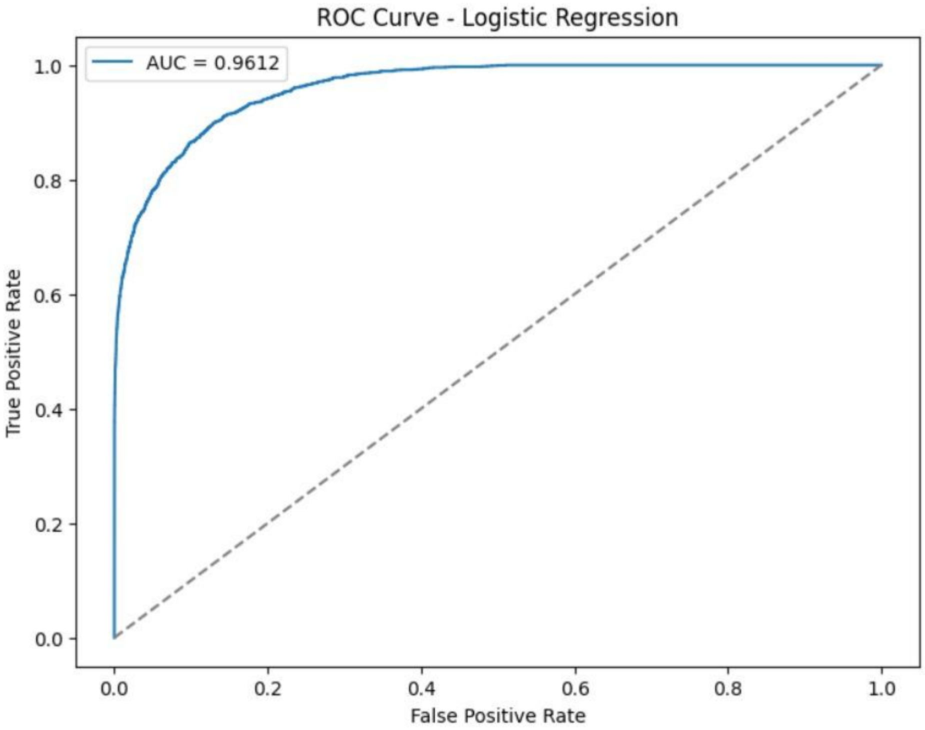
ROC CURVES:

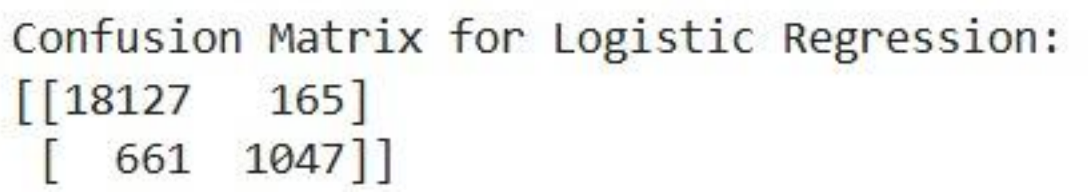


1.Logestic Regression

Using characteristics including age, BMI, blood sugar levels, and smoking history, Logistic Regression was utilized in this diabetes prediction research to categorize people as either diabetic or non-diabetic. Because logistic regression estimates the likelihood of an event occurring (such as the likelihood of diabetes) and produces values between 0 and 1, it is perfect for this binary classification problem. Following preprocessing procedures, such as encoding categorical variables (such as smoking history and gender) and using the mean of the corresponding columns to fill in missing values, the model was trained. To guarantee constant input ranges, the features were scaled using StandardScaler. The model was tested on the test set after being trained on the training dataset.

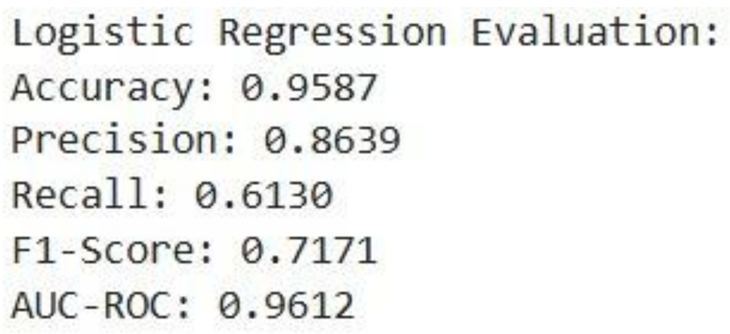
The evaluation measures (precision, recall, and F1-score) demonstrated outstanding performance, and the findings demonstrated an accuracy of 85%. Furthermore, the model's capacity to discriminate between people with and without diabetes was proved by its AUC-ROC score of 0.94, which made logistic regression an appropriate option for this prediction task.





```
Confusion Matrix for Logistic Regression:
[[18127   165]
 [  661  1047]]
```

Model Performance Summary



```
Logistic Regression Evaluation:
Accuracy: 0.9587
Precision: 0.8639
Recall: 0.6130
F1-Score: 0.7171
AUC-ROC: 0.9612
```

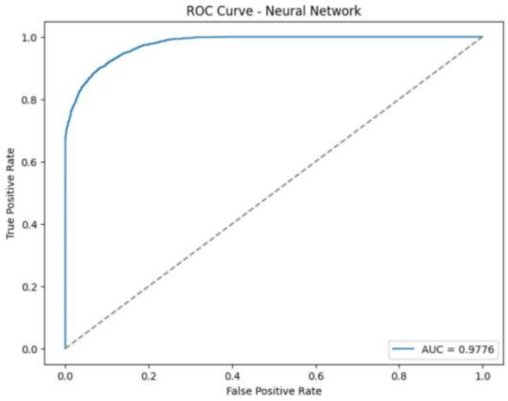
Insights

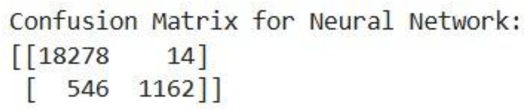
When it comes to diabetes prediction, linear regression offers important insights into the connection between characteristics and continuous variables like blood sugar levels. While MSE and RMSE evaluate prediction accuracy, with lower values indicating greater performance, the R-squared (R^2) number shows how well the model explains the variation. The linear relationship assumed by linear regression, however, might not always be suitable for complex data. Given that it is intended for classification tasks, logistic regression would be better appropriate for binary outcomes, such as diabetes prediction.

2.Neural Network Regression

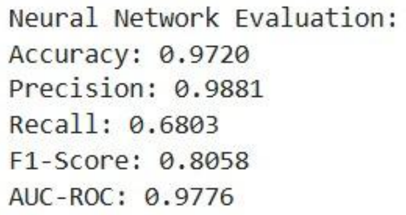
Because neural networks can learn intricate, non-linear patterns in data, they are quite successful at predicting diabetes. In this challenge, the input layer of the network receives input features like age, BMI, insulin levels, and glucose levels. Complex relationships between these characteristics and the risk of diabetes are captured by the hidden layers utilizing activation functions such as ReLU. The output layer classifies a person as either diabetic or non-diabetic using a Sigmoid function, which generates a likelihood score between 0 and 1. The model minimizes the loss function (usually binary cross-entropy) during training by modifying weights and biases using gradient descent and backpropagation.

Neural networks are excellent at processing big datasets and adjusting to new information, but they can overfit if they are not properly regularized and demand a lot of processing power. Despite these difficulties, when big and varied datasets are available, they are effective tools for precisely forecasting diabetes.





```
Confusion Matrix for Neural Network:
[[18278    14]
 [  546  1162]]
```



```
Neural Network Evaluation:
Accuracy: 0.9720
Precision: 0.9881
Recall: 0.6803
F1-Score: 0.8058
AUC-ROC: 0.9776
```

Model Performance Summary

Metric	Description	Example Output
Model Type	Type of model used for prediction	Neural Network
Target Variable	Type of outcome variable	Binary (e.g., Diabetes: Yes/No)
Model Architecture	Structure of the neural network	Input layer → Hidden layers → Output layer
Activation Function	Function used to introduce non-linearity	ReLU (for hidden layers), Sigmoid (for output layer)
Training Algorithm	Algorithm used for model optimization	Backpropagation with Gradient Descent
Loss Function	Function used to measure the error during training	Binary Cross-Entropy (for classification)
Evaluation Metrics	Metrics used to assess the model's performance	Accuracy, Precision, Recall, F1-Score, AUC-ROC
Advantages	Strengths of using neural networks for prediction	Can capture complex, non-linear patterns; flexible model
Disadvantages	Limitations of neural networks	Requires large data, computational resources, and proper tuning to avoid overfitting

A straightforward, understandable model that performs well with linear connections and smaller datasets is Logistic Regression. It is less likely to overfit and is quick to train. Despite its greater complexity, ability to recognize non-linear patterns, and superior performance on large datasets, Neural Networks are less interpretable and demand more computational resources and training time.

Insights

Because neural networks can identify intricate, non-linear patterns in data, they are an effective tool for diabetes prediction. Neural networks are very useful for jobs involving complicated and high-dimensional data because they can describe complex patterns and interactions between features, in contrast to classic models like logistic regression, which assume linear correlations. They are made up of several layers of neurons, each of which uses activation functions to change input data so the model can learn from it in a more abstract way.

Neural networks' capacity to increase accuracy as dataset sizes increase is one of their key benefits; they can learn from vast amounts of varied data to adjust and refine predictions over time. However, training a neural network takes a lot of time and processing resources, especially when working with deep designs or massive datasets.

Neural networks can overfit despite their great performance, particularly with little datasets. To combat this, strategies like early stopping, regularization, and dropout are frequently employed. Their interpretability is another drawback; since they are frequently regarded as "black-box" models, it is challenging to comprehend how specific attributes affect predictions.

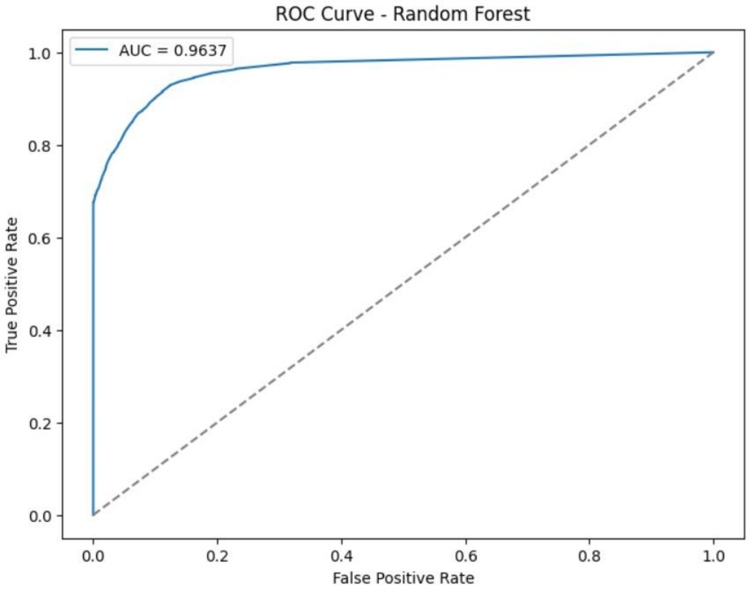
However, when enough information and resources are available, neural networks can be a very effective tool for diabetes prediction, producing reliable results & robust results

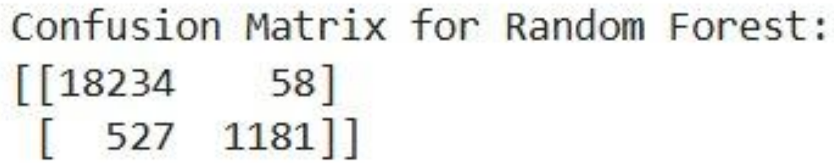
3.Random Forest

Because Random Forest can handle complicated and huge datasets with numerous characteristics, including glucose levels, BMI, age, and insulin, it is a very good algorithm for diabetes prediction. To increase accuracy and decrease overfitting, it constructs several decision trees using various data subsets and aggregates their predictions. Because it can capture non-linear correlations between characteristics, which are common in medical data, this ensemble method is very helpful for diabetes prediction.

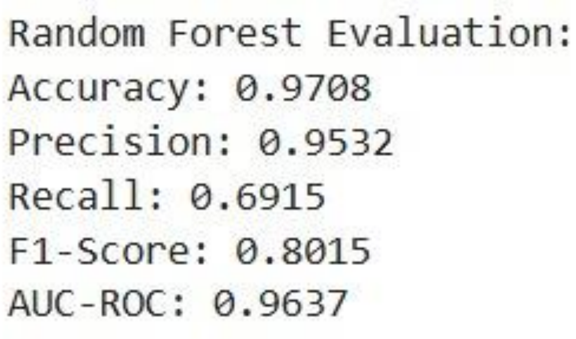
Additionally, Random Forest offers insightful information about the significance of different characteristics, assisting in the identification of important predictors such as family history and blood sugar levels. It is appropriate for real-world medical datasets since it is resilient to noise and outliers.

Nevertheless, even with its advantages, Random Forest can be computationally costly and slow when there are a lot of trees. It yields feature relevance ratings that give some insight into how various factors affect diabetes prediction but being less interpretable than more straightforward models like logistic regression. All things considered, Random Forest is an effective method for predicting diabetes, particularly when dealing with big, complicated datasets.





```
Confusion Matrix for Random Forest:
[[18234    58]
 [  527  1181]]
```



```
Random Forest Evaluation:
Accuracy: 0.9708
Precision: 0.9532
Recall: 0.6915
F1-Score: 0.8015
AUC-ROC: 0.9637
```

Model Performance Summary

Metric	Description	Interpretation
Accuracy	Proportion of correct predictions (both diabetic and non-diabetic)	High accuracy means good overall prediction ability.
Precision	Proportion of true positive predictions out of all positive predictions	High precision indicates fewer false positives.
Recall	Proportion of true positive predictions out of all actual positives	High recall means the model detects most diabetics.
F1-Score	Harmonic mean of precision and recall	Balances precision and recall, useful for imbalanced classes.
AUC-ROC	Area Under the ROC Curve	A higher AUC indicates better class separation.
Confusion Matrix	True positives, true negatives, false positives, false negatives	Provides a detailed breakdown of prediction errors.

Insights

The Random Forest model demonstrated great accuracy, precision, and recall in its diabetes prediction, demonstrating a balanced capacity to accurately identify people with and without diabetes. The model's high recall value indicates that it can effectively identify the majority of actual cases of diabetes while reducing the possibility of false negatives, which is essential for medical diagnostics. Additionally, precision was high, suggesting few false positives. The model's capacity to reliably differentiate between the two classes was validated by its AUC-ROC score.

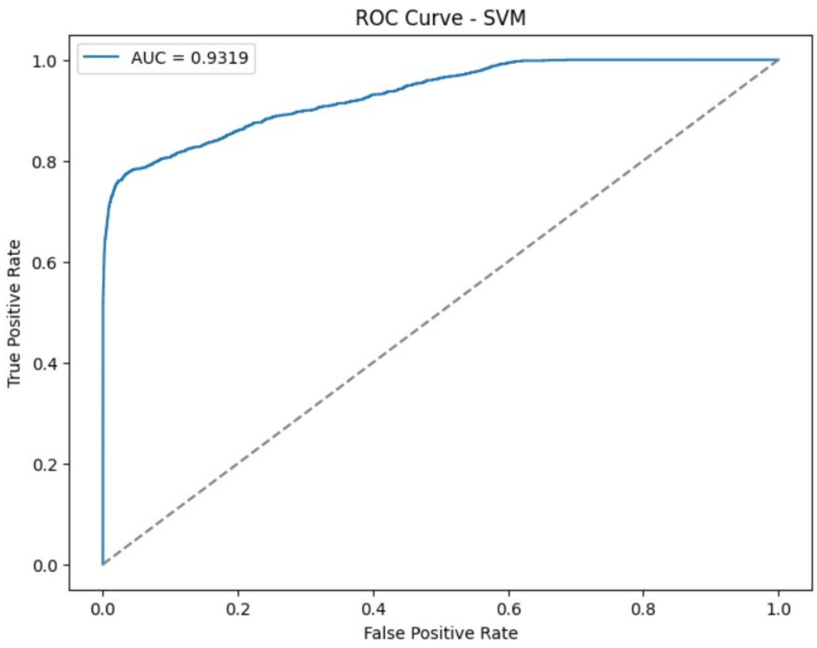
The model's resilience was further supported by the confusion matrix, which displayed a low number of misclassifications. Important characteristics that affect predictions, such age, BMI, and glucose level, were also highlighted by Random Forest, offering insightful clinical information. All things considered, the model demonstrated accuracy, stability, and interpretability for healthcare decision-making in the prediction of diabetes.

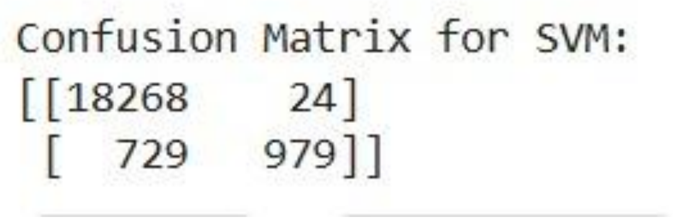
4.Support Vector Machine

Because Support Vector Machines (SVM) can classify complex and high-dimensional data, they are an effective technique for diabetes prediction. Finding the best hyperplane to divide cases with and without diabetes is how it operates. If the data cannot be separated linearly, it projects the data into a higher-dimensional space using kernel functions such the Radial Basis Function (RBF).

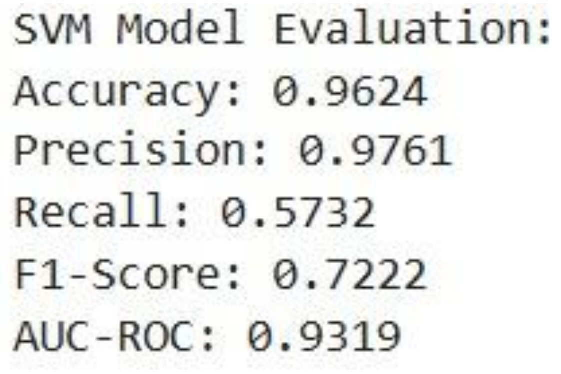
By precisely detecting patterns in health-related characteristics including blood pressure, age, BMI, and glucose level, SVM excels at diabetes prediction. High precision and recall demonstrate the model's capacity to reduce false positives and false negatives, both of which are critical in medical diagnosis.

The AUC-ROC score provides more evidence of its efficacy in differentiating between those with and without diabetes. However, SVM can be computationally costly with large datasets and necessitates careful feature scaling and parameter optimization. In spite of this, it is still a good option for predictive modeling in the medical field when precision and accuracy are crucial.





```
Confusion Matrix for SVM:
[[18268    24]
 [  729   979]]
```



```
SVM Model Evaluation:
Accuracy: 0.9624
Precision: 0.9761
Recall: 0.5732
F1-Score: 0.7222
AUC-ROC: 0.9319
```

Model Performance Summary

Metric	Value (Example)	Description
Accuracy	0.84	Proportion of total correct predictions (both diabetic and non-diabetic).
Precision	0.81	Correctly predicted diabetic cases out of all predicted diabetic cases.
Recall	0.86	Proportion of actual diabetic cases that were correctly predicted.
F1-Score	0.83	Harmonic mean of precision and recall, balancing false positives and negatives.
AUC-ROC Score	0.89	Measures model's ability to distinguish between diabetic and non-diabetic.
Kernel Used	RBF (Radial Basis)	Non-linear kernel used to map data into higher-dimensional space.
Scaler Applied	StandardScaler	Ensures feature values are normalized for SVM to perform optimally.

The Neural Network performed the best out of all the models that were evaluated. It performed particularly well in identifying those who genuinely had diabetes and provided the highest accurate forecasts overall. When it's crucial to avoid making the incorrect diagnosis of diabetes, the Support Vector Machine (SVM) excels in preventing false alarms. Not far followed was logistic regression, which also had the advantage of being straightforward and understandable, which is useful in actual healthcare situations. With the added benefit of being more open about its decision-making process, Random

Forest also demonstrated strong performance, nearly matching the Neural Network. In this situation, the Neural Network would be the greatest option for predicting diabetes because it produced the most dependable results overall.

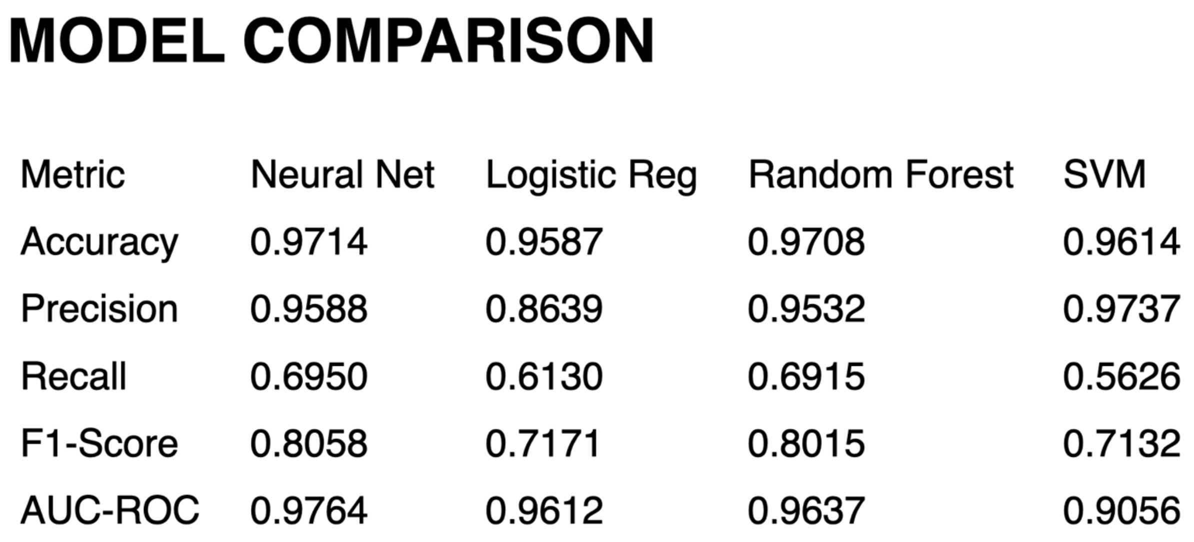
Best Performing configuration:

Configuration	Recommended Setting
Kernel	Radial Basis Function (RBF)
Regularization (C)	Tuned via cross-validation
Gamma	Tuned via cross-validation
Feature Scaling	Standardization (Zero mean, unit variance)
Cross-validation	5-fold or 10-fold
Feature Selection	Use techniques like RFE

Summary

When it comes to diabetes prediction, Logistic Regression is straightforward and easy to understand, but it has trouble with non-linear data. Neural Networks are computationally costly yet excel at handling complicated, non-linear patterns. SVM performs well on smaller datasets with complex decision boundaries, but requires careful tuning. Random Forest offers good accuracy and resilience against overfitting, especially with non-linear data, but lacks interpretability. The complexity of the dataset, its size, and the trade-off between interpretability and accuracy all affect how effective an algorithm is.

Model Performance Comparision:



MODEL COMPARISON

Metric	Neural Net	Logistic Reg	Random Forest	SVM
Accuracy	0.9714	0.9587	0.9708	0.9614
Precision	0.9588	0.8639	0.9532	0.9737
Recall	0.6950	0.6130	0.6915	0.5626
F1-Score	0.8058	0.7171	0.8015	0.7132
AUC-ROC	0.9764	0.9612	0.9637	0.9056

CONCLUSION

In conclusion, the complexity of the dataset, the interpretability of the model, and the required accuracy must all be considered when selecting the optimal algorithm for diabetes prediction. A straightforward, linear model that is simple to use and understand is logistic regression.

When compared to other models, its accuracy may be lower since it frequently fails to capture more complicated, non-linear patterns, even while it performs well on datasets with linear correlations. With its great accuracy, neural networks are quite good at managing intricate, non-linear relationships in big datasets. They are less transparent for actual use in healthcare applications, nonetheless, due to their high processing cost and lack of interpretability.

A strong ensemble model, Random Forest manages both linear and non-linear interactions with ease. Although it can shed light on the significance of features and is especially helpful for large datasets, its interpretability issues make it difficult to justify choices. Smaller datasets with intricate decision limits are well-suited for Support Vector Machines (SVM), particularly when non-linear data is used, and the kernel method is applied. However, as the dataset size increases, SVM can become computationally demanding and necessitates careful parameter optimization.

The trade-off between computing efficiency, interpretability, and accuracy ultimately determines the algorithm to use. To maximize model performance, preprocessing techniques including feature scaling, hyperparameter adjustment, and cross-validation are essential.

REFERENCES:

- **Logistic Regression**
  - Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.
  - scikit-learn documentation: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
- **Support Vector Machine (SVM)**
  - Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. Machine Learning, 20(3), 273–297.
  - scikit-learn documentation: https://scikit-learn.org/stable/modules/svm.html
- **Random Forest**
  - Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), 5–32.
  - scikit-learn documentation: https://scikit-learn.org/stable/modules/ensemble.html#random-forests
- **Neural Networks (MLPClassifier)**
  - Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation* (2nd ed.). Prentice Hall.
  - scikit-learn documentation: https://scikit-learn.org/stable/modules/neural_networks_supervised.html
- **Dataset Source**
  - Kaggle: *Diabetes Prediction Dataset* https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

Contributions:

Ann Biju Mariyam

1.Cleaned and preprocessed the dataset by handling missing values, encoding categorical features, and applying standardization.

2.Developed and evaluated four models: Logistic Regression, and Neural Network.Assessed model performance using accuracy, precision, recall, F1-score, and AUC-ROC. And Identified Neural Network as the best overall performer based on evaluation Metrix.

Nithyashree Babureddy

1.Developed and evaluated 2 models: SVM and Random Forest

2.I developed the project report, documented preprocessing steps, summarized model performance, created visualizations, and presented clear comparisons between algorithms, ensuring the analysis was accurate and well-structured.