Diabetes Prediction in Healthcare

Leveraging Machine Learning for Early Detection

Team members:

- *Apurva Ajit Deshpande*
- *Rushikesh Bhavesh Gholap*
- *Sushmitha Rajeswari Muppa*
- *Vedavarshita Nunna*

Importance of early detection for effective management and patient care

Timely Intervention:

- Proactive Treatment: Allows swift intervention at the earliest signs of a health issue.
- Preventive Measures: Enables implementation of preventive measures to mitigate the progression of conditions.

Optimized Treatment Plans:

- Personalized Care: Facilitates tailored treatment plans based on individual patient needs.
- Higher Treatment Success Rates: Increases success rates with responsive interventions in early stages.

Reduced Health Complications:

- Minimized Disease Progression: Early detection prevents or minimizes disease progression.
- Improved Quality of Life: Addresses health conditions early, enhancing quality of life.

Cost-Efficient Healthcare:

- Lower Treatment Costs: Results in more cost-efficient healthcare.

Data Source - Where Our Insights Come From

Sources and Collection Methodology:

Electronic Health Records (EHRs) serve as the primary data source for the Diabetes Prediction dataset, offering a comprehensive digital repository of patient health records. Compiled by healthcare providers, these records include medical history, diagnoses, treatments, and outcomes. By aggregating EHRs from diverse sources, we ensured dataset integrity through thorough cleaning and preprocessing.

Leveraging EHRs for the dataset provides distinct advantages, given their inclusive nature encompassing demographic and clinical details. This facilitates accurate machine learning model development, and the longitudinal view of a patient's health aids in identifying patterns. The dataset's relevance to real-world healthcare settings is accentuated by the widespread use of EHRs in clinical practice.

The collection methodology involves acquiring medical and demographic data from patients diagnosed with or at risk of diabetes. This includes age, gender, BMI, hypertension, heart disease, smoking history, HbA1c level, and blood glucose level, gathered through surveys, medical records, and laboratory tests.

We found a dataset ready using this procedure in **Kaggle**

Data Source - What our data looks like

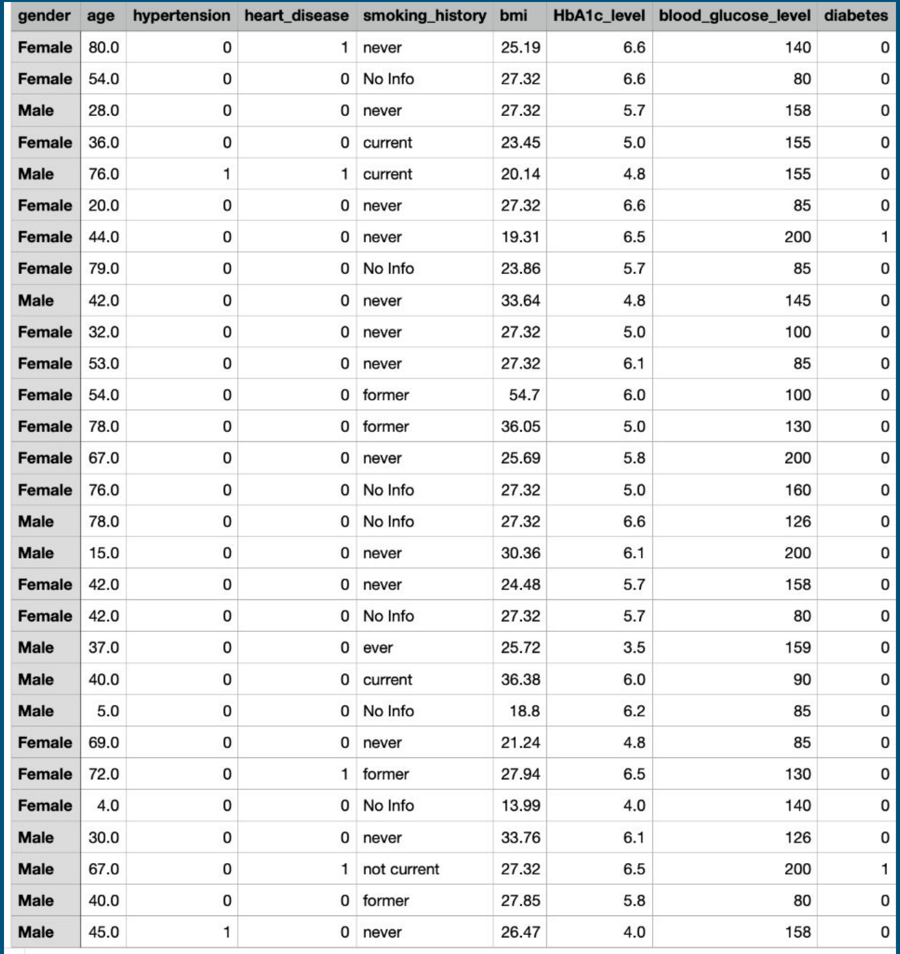
- **Gender**: Gender denotes an individual's biological sex, influencing susceptibility to diabetes. The categories include male, female, and other.
- **Age**: Age is a crucial factor, with diabetes more prevalent in older adults. Our dataset covers ages ranging from 0 to 80.
- **Hypertension**: Hypertension signifies persistently elevated blood pressure. It is represented by values 0 (absence) or 1 (presence).
- **Heart Disease**: Heart disease is linked to an increased diabetes risk, denoted by values 0 (absence) or 1 (presence).
- **Smoking History**: Smoking history is a diabetes risk factor. Our dataset includes categories: not current, former, no info, current, never, and ever.
- **BMI** (Body Mass Index): BMI, indicating body fat based on weight and height, correlates with diabetes risk. Ranging from 10.16 to 71.55 in our dataset, BMI categories include underweight, normal, overweight, and obese.
- **HbA1c Level**: HbA1c level measures average blood sugar over 2-3 months. Levels above 6.5% suggest a higher diabetes risk.
- **Blood Glucose Level**: Blood glucose level, indicating current glucose in the bloodstream, is a key diabetes indicator.
- **Diabetes**: The **target variable**, diabetes, is predicted with values 1 (presence) or 0 (absence).

Data Source - What our data looks like

The dataset contains nearly 100 thousand samples with the following distribution of diabetic and non-diabetic data.

Diabetic: 8,500

Non-diabetic: 91,500



gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Female	80.0	0	1	never	25.19	6.6	140	0
Female	54.0	0	0	No Info	27.32	6.6	80	0
Male	28.0	0	0	never	27.32	5.7	158	0
Female	36.0	0	0	current	23.45	5.0	155	0
Male	76.0	1	1	current	20.14	4.8	155	0
Female	20.0	0	0	never	27.32	6.6	85	0
Female	44.0	0	0	never	19.31	6.5	200	1
Female	79.0	0	0	No Info	23.86	5.7	85	0
Male	42.0	0	0	never	33.64	4.8	145	0
Female	32.0	0	0	never	27.32	5.0	100	0
Female	53.0	0	0	never	27.32	6.1	85	0
Female	54.0	0	0	former	54.7	6.0	100	0
Female	78.0	0	0	former	36.05	5.0	130	0
Female	67.0	0	0	never	25.69	5.8	200	0
Female	76.0	0	0	No Info	27.32	5.0	160	0
Male	78.0	0	0	No Info	27.32	6.6	126	0
Male	15.0	0	0	never	30.36	6.1	200	0
Female	42.0	0	0	never	24.48	5.7	158	0
Female	42.0	0	0	No Info	27.32	5.7	80	0
Male	37.0	0	0	ever	25.72	3.5	159	0
Male	40.0	0	0	current	36.38	6.0	90	0
Male	5.0	0	0	No Info	18.8	6.2	85	0
Female	69.0	0	0	never	21.24	4.8	85	0
Female	72.0	0	1	former	27.94	6.5	130	0
Female	4.0	0	0	No Info	13.99	4.0	140	0
Male	30.0	0	0	never	33.76	6.1	126	0
Male	67.0	0	1	not current	27.32	6.5	200	1
Male	40.0	0	0	former	27.85	5.8	80	0
Male	45.0	1	0	never	26.47	4.0	158	0

Prior related work

Gradient Boosting - "*Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine*" (reference) - Explores the use of Gradient Boosting for diabetes prediction, combining weak prediction models, typically decision trees, in a stage-by-stage process similar to other boosting approaches. Subsequently, the model is simplified through an arbitrary differentiable loss function optimization.

K-Nearest Neighbor - "*Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method*" (reference) explores KNN for Regression and primarily classification, where the algorithm assigns a new case to the most similar category based on existing case similarities.

Random Forest Classifier - "*Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach*" (reference) Explores another forecasting and behavior analysis model—random forest. Comprising numerous decision trees, each representing a unique instance, this algorithm classifies input data. The technique assesses each instance independently and outputs the most commonly predicted result.

Key studies and methodologies that inspire our approach

Decision Tree - Several papers (including *Machine Learning based Diabetes Prediction using Decision Tree* reference) indicate that Decision Trees give better results due to their ability to capture and interpret complex relationships within data, offering transparency and ease of understanding, which is crucial for effective clinical decision-making.

Naive Bayes - Want to explore Naive Bayes classifier as it is efficient in handling high-dimensional data, and ability to provide probabilistic predictions considering independence of features. Several papers have tried out this approach and it gave them good results.

This positive outcome demonstrates Naive Bayes' robust capability in handling complex patterns, making it a compelling choice for precise and reliable diabetes onset.

Approach - Navigating the Diabetes Prediction Landscape

Following are the algorithms we are exploring for diabetes detection:

1. Decision Tree
2. Logistic Regression
3. Naive Bayes classifier
4. K-Nearest Neighbors
5. Linear Discriminant Analysis
6. Ensemble of a few classifiers

Why we chose these classifiers?

- The positive outcomes from the prior work done using Decision Trees and Naive Bayes encourages us to explore them.
- Logistic Regression is worth exploring due to its simplicity, interpretability, and ability to model the probability of binary outcomes, making it a practical choice for medical decision-making.
- Want to explore Linear Discriminant Analysis as it effectively separates classes in high-dimensional data, aiding accurate diabetes prediction with statistical rigor and simplicity.

Addressing Imbalanced Dataset

1. Over-sampling:

- Description: Over-sampling involves increasing the number of instances in the minority class to balance class distribution.
- Methods:
  - Random Over-sampling: Replicating instances randomly from the minority class.
  - SMOTE (Synthetic Minority Over-sampling Technique): Creating synthetic instances by interpolating between existing minority class instances.

2. Under-sampling:

- Description: Under-sampling reduces the number of instances in the majority class to achieve a balanced dataset.
- Methods:
  - Random Under-sampling: Randomly removing instances from the majority class.
  - Cluster-based Under-sampling: Removing instances from clusters in the majority class to preserve important patterns.

Dataset Under-sampling

It turned out that random under-sampling was preferable for diabetes datasets as it reduced the computational complexity, mitigated model overfitting, and preserved existing patterns.

smaller dataset:

- Interpretability
- Noise Filter
- Computationally efficient
- Simpler Models

Optimizing Model Inputs: Feature Selection in Dataset Preprocessing

With respect to features, removing smoking history from training data has helped improve the results as it is irrelevant when looked at the correlation with diabetes data.

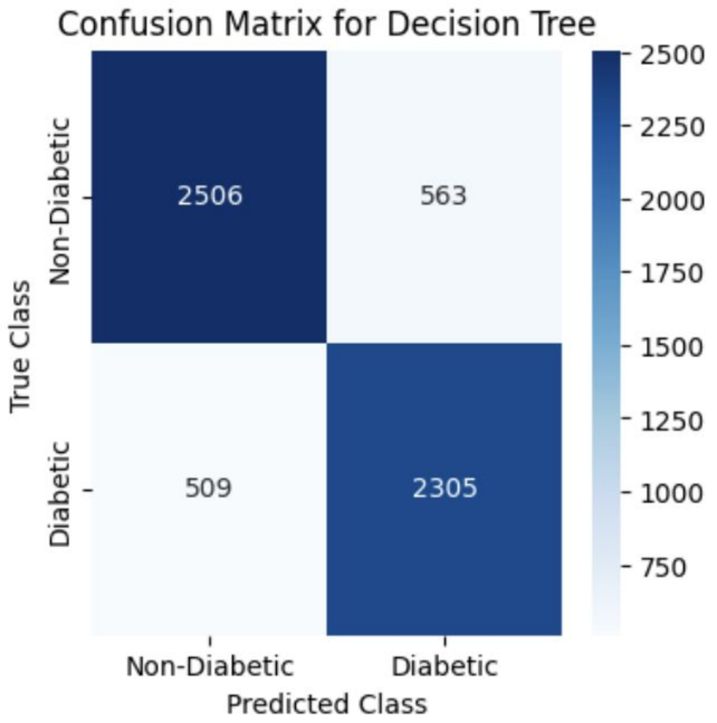
To build a model for predicting diabetes, smoking history is not significantly contributing to the prediction and it introduces biases due to limited data on smokers. So, it has been excluded to simplify the model and avoid potential confounding effects, leading to a more focused and effective diabetes prediction model.

Decision Trees

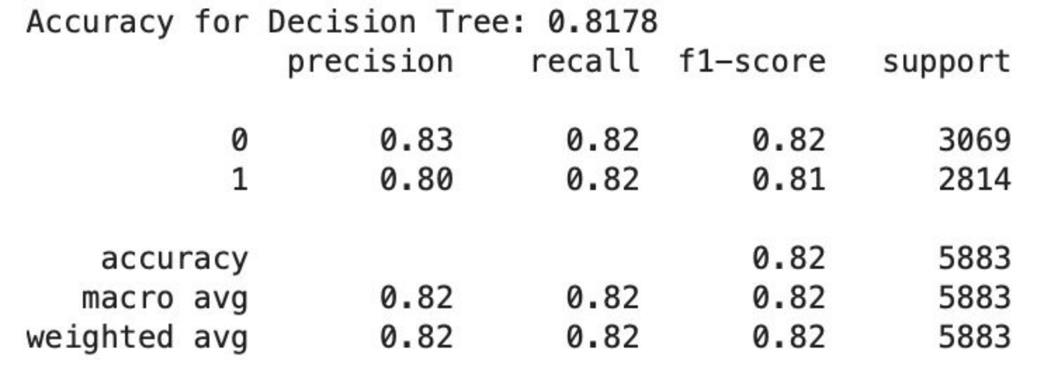
Decision trees perform well in diabetes prediction due to their non-linear handling, robustness to irrelevant features, ability to handle mixed data types, and effectiveness in handling imbalanced datasets—making them ideal for capturing complex relationships in healthcare data.

- *Strengths:* Decision trees are interpretable, handle both numerical and categorical data, and can capture non-linear relationships.
- *Weaknesses:* They can be sensitive to noisy data and may overfit the training data.

Decision Trees Model Evaluation



Max False positives



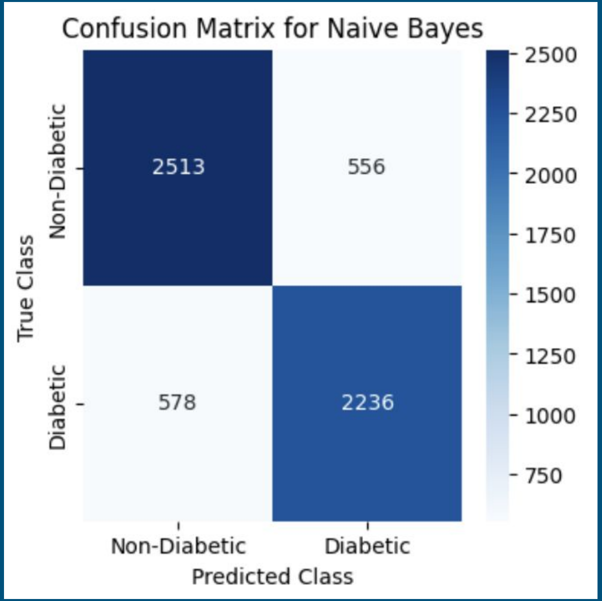
```
Accuracy for Decision Tree: 0.8178
              precision    recall  f1-score   support

           0       0.83      0.82      0.82      3069
           1       0.80      0.82      0.81      2814

    accuracy                           0.82      5883
   macro avg       0.82      0.82      0.82      5883
weighted avg       0.82      0.82      0.82      5883
```

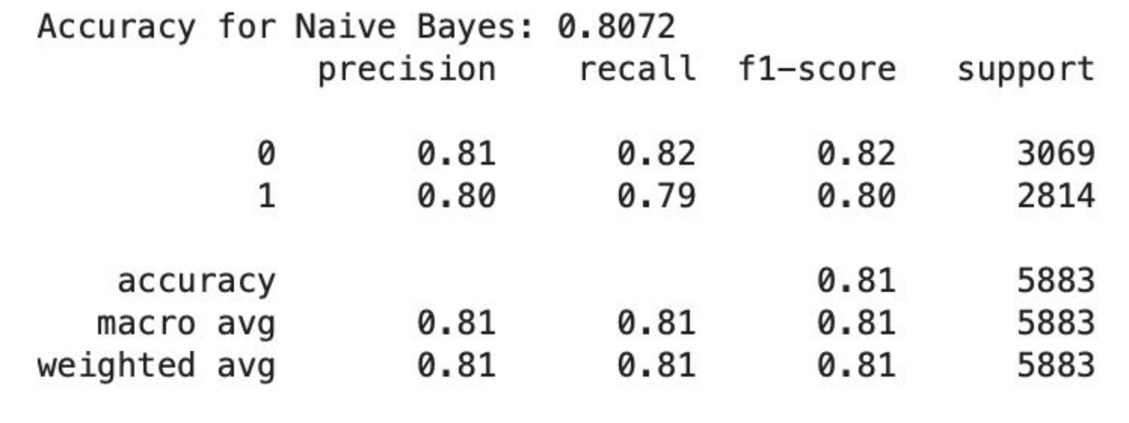
Naive Bayes

Naive Bayes emerges as a descent algorithm in the prediction of diabetes, showcasing its strength in healthcare analytics. This algorithm's efficacy with high-dimensional data, and computationally efficient even though it’s high dimensional. This simplifies the complexity of modeling intricate relationships within the diabetes prediction dataset.

- *Strengths:* Naive Bayes is simple, computationally efficient, and works well with high-dimensional data.
- *Weaknesses:* It assumes independence between features, which might not be true in all cases. It may not capture complex relationships in the data.

Naive Bayes Model Evaluation





```
Accuracy for Naive Bayes: 0.8072
              precision    recall  f1-score   support

           0       0.81      0.82      0.82      3069
           1       0.80      0.79      0.80      2814

    accuracy                           0.81      5883
   macro avg       0.81      0.81      0.81      5883
weighted avg       0.81      0.81      0.81      5883
```

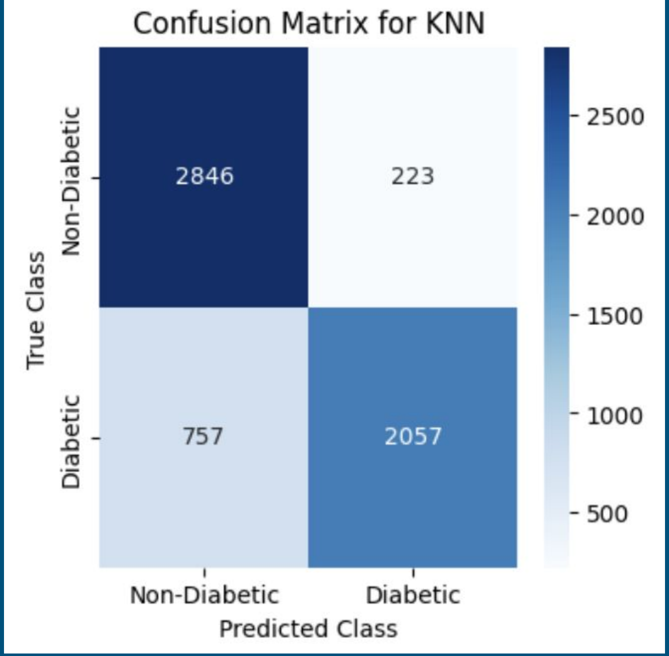
K-Nearest Neighbors

KNN, a non-parametric and instance-based learning algorithm, relies on the distance of data points in feature space for classification. It has highest results among other classifiers majorly because its adaptability to complex patterns within the diabetic prediction dataset, making minimal assumptions about the underlying data distribution.

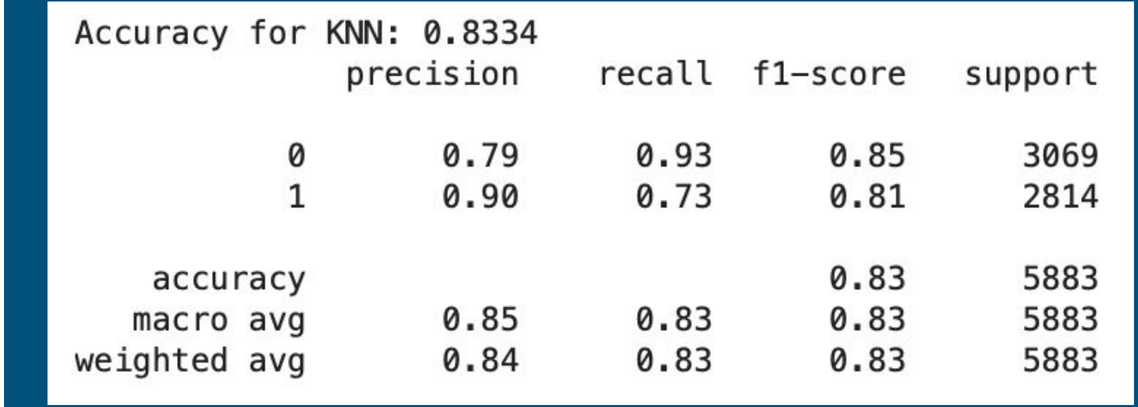
KNN excels in capturing local relationships, which is crucial in the context of diabetes prediction where the influence of certain features may vary across the data.

- *Strengths:* KNN is non-parametric and can capture complex decision boundaries. It's easy to understand and implement.
- *Weaknesses:* It can be computationally expensive, especially with large datasets. The prediction performance degrades in high-dimensional spaces.

K-Nearest Neighbors Model Evaluation



Min false positives



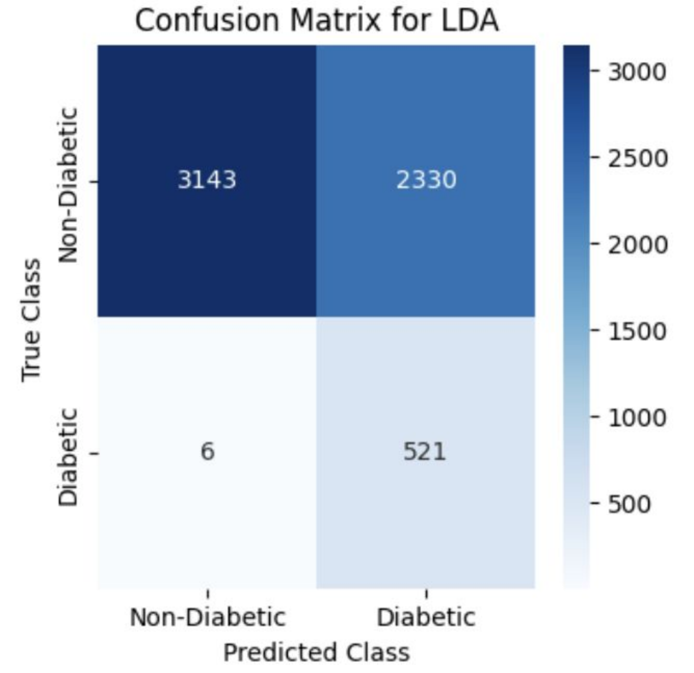
```
Accuracy for KNN: 0.8334
              precision    recall  f1-score   support

           0       0.79      0.93      0.85      3069
           1       0.90      0.73      0.81      2814

    accuracy                           0.83      5883
   macro avg       0.85      0.83      0.83      5883
weighted avg       0.84      0.83      0.83      5883
```

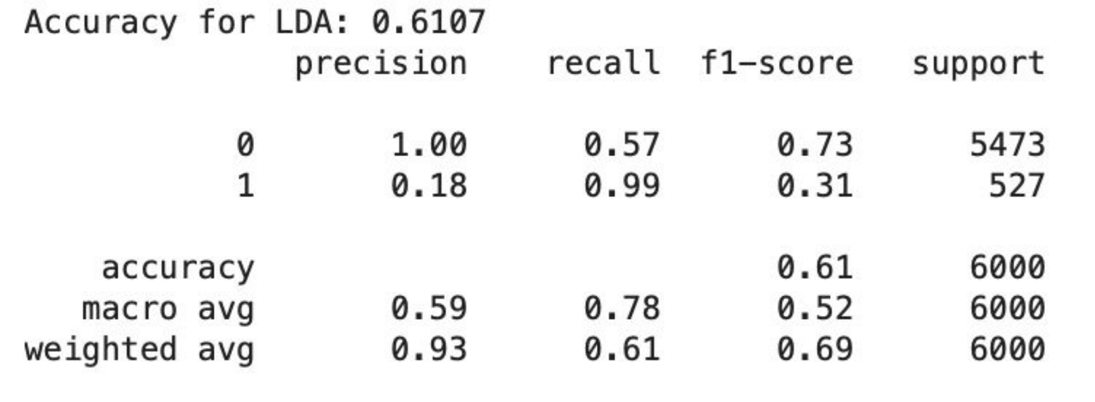
Linear discriminant analysis

Linear Discriminant Analysis (LDA) is a powerful statistical method used for linear classification by linearly combining the features that characterizes or separates two classes i.e., diabetes and non-diabetes and it also can be used as dimensionality reduction technique.

Strength: LDA is less prone to overfitting, especially when the number of samples per class is small.
performs dimensionality reduction by finding a linear combination of features
Weakness: LDA assumes that the features within each class are normally distributed so it's sensitive to outliers

Linear discriminant analysis Model Evaluation





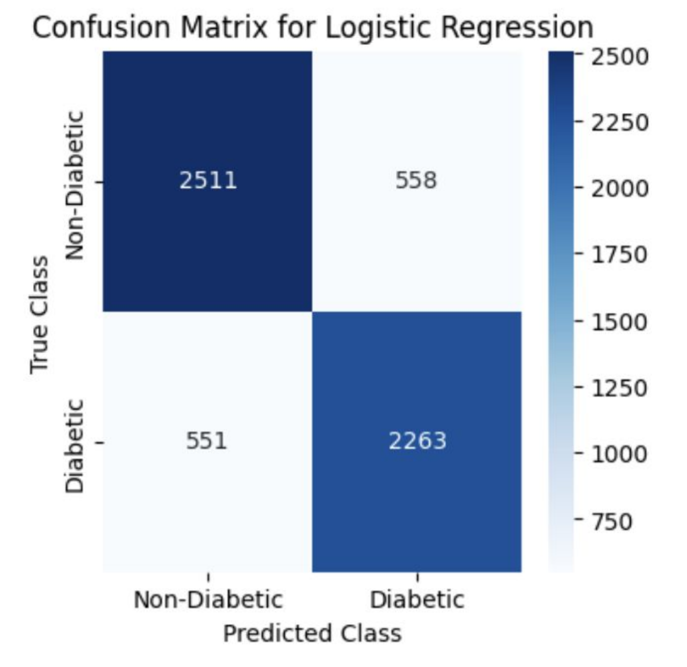
```
Accuracy for LDA: 0.6107
              precision    recall  f1-score   support

           0       1.00      0.57      0.73      5473
           1       0.18      0.99      0.31       527

    accuracy                           0.61      6000
   macro avg       0.59      0.78      0.52      6000
weighted avg       0.93      0.61      0.69      6000
```

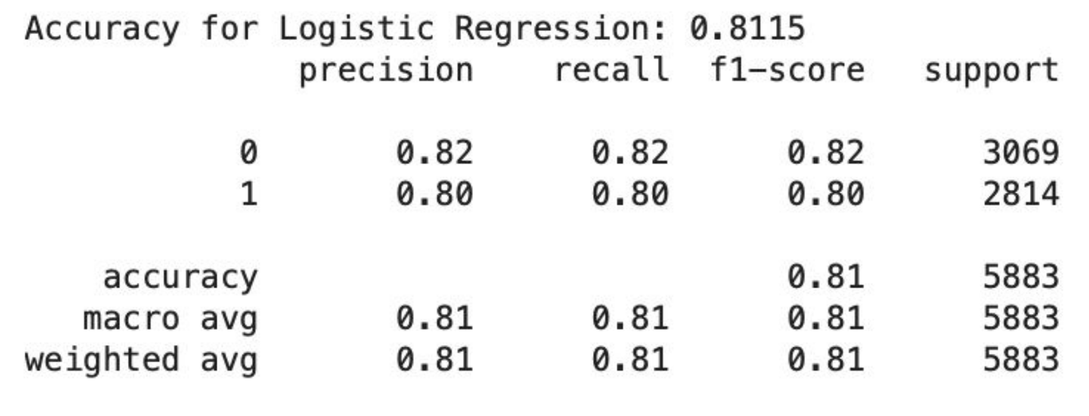
Logistic Regression

Logistic Regression employs a probabilistic approach, modeling the likelihood of an instance belonging to a particular class, making it adept at capturing intricate relationships within the diabetes prediction dataset.

- *Strengths:* Logistic regression is simple, and well-suited for binary classification problems. It provides probabilities for predictions.
- *Weaknesses:* It assumes a linear relationship between features and the log-odds of the response. May not perform well when the true relationship is non-linear.

Logistic Regression Model Evaluation


Confusion Matrix for Logistic Regression


```
Accuracy for Logistic Regression: 0.8115
              precision    recall  f1-score   support

           0       0.82      0.82      0.82      3069
           1       0.80      0.80      0.80      2814

    accuracy                           0.81      5883
   macro avg       0.81      0.81      0.81      5883
weighted avg       0.81      0.81      0.81      5883
```

Ensemble: Voting

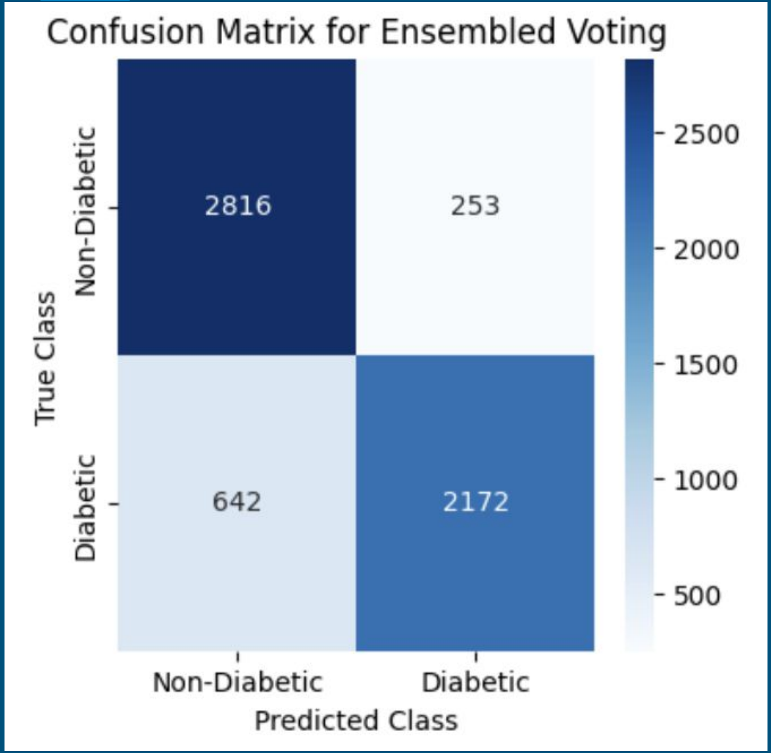
Ensemble Methods:

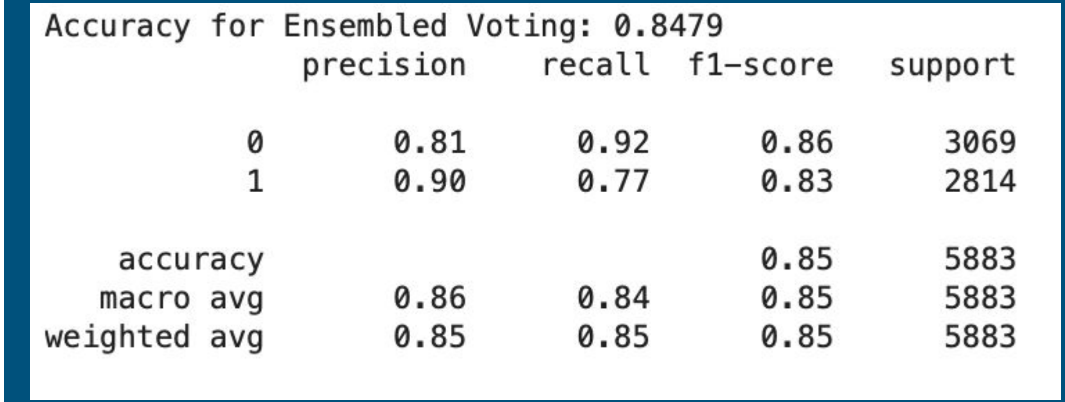
1. Bagging
2. **Voting**
3. Boosting
4. Stacking

The ensemble voting methodology amalgamates these individual predictions from the above models, harnessing the collective intelligence embedded within the diverse model architectures.

This fusion of distinct predictive perspectives not only serves to mitigate the impact of individual model idiosyncrasies but also capitalizes on the inherent diversity to achieve a more robust and generalized predictive outcome

Ensemble Voting Model Evaluation





```
Accuracy for Ensembled Voting: 0.8479
              precision    recall  f1-score   support

           0       0.81      0.92      0.86      3069
           1       0.90      0.77      0.83      2814

    accuracy                           0.85      5883
   macro avg       0.86      0.84      0.85      5883
weighted avg       0.85      0.85      0.85      5883
```

Future Work

1. Use more robust algorithms like Random Forest, Artificial Neural Networks etc for diabetes prediction
2. Exploring a better way to make use of over-sampling to achieve better accuracies.
3. Dynamic Exploration, were we can analyze predictive model performance over time, it is crucial for understanding its evolving efficacy in diabetes prediction.

Thank You

References

"Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine" (reference)

"Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method" (reference)

"Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach" (reference)

"Nadeem et al. [7]"

"Machine Learning based Diabetes Prediction using Decision Tree" reference

Dataset (reference)