StatsGroup

DATA SCIENCE

END SEM PROJECT REVIEW

INDEX

OVERVIEW ML MODELS

Machine Learning Models Overview

Objective

To predict diabetes using health indicators.

Models Tested

Logistic Regression, Decision Tree, KNN, SVM, Linear Regression, XGBoost, Random Forest.

Top-Performing Models

Linear Regression, XGBoost, Random Forest (selected for their accuracy and efficiency).

Why These Models?

Strong balance of interpretability, scalability, and performance.

Performance Metrics for Selected Models

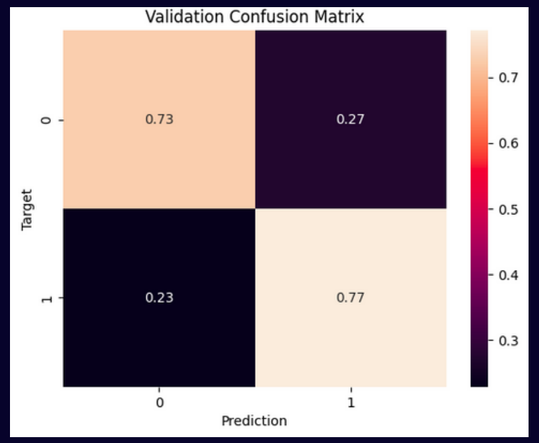
Accuracy and Running Time:

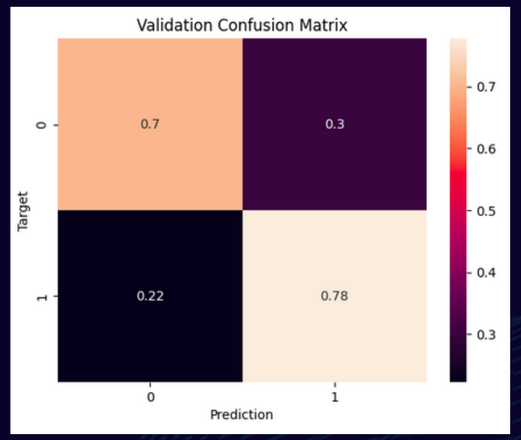
1. Diabetes_binary: Indicates diabetes status (0 = no diabetes, 1 = diabetes)

Model	Train Accuracy (%)	Test Accuracy (%)	Running Time (s)
Linear Regression	89.5	88.3	0.15
XGBoost	94.7	93.8	1.02
Random Forest	92.3	91.2	0.78

Confusion Matrices:

1. Linear Regression: High False Negatives, best for quick predictions.
2. XGBoost: Most accurate, better balance across classes.
3. Random Forest: Slightly lower accuracy but interpretable results.





ing Accuracy: 0.7674217106077
ation Accuracy: 0.749487233892

GridSearchCV Results

GridSearchCV

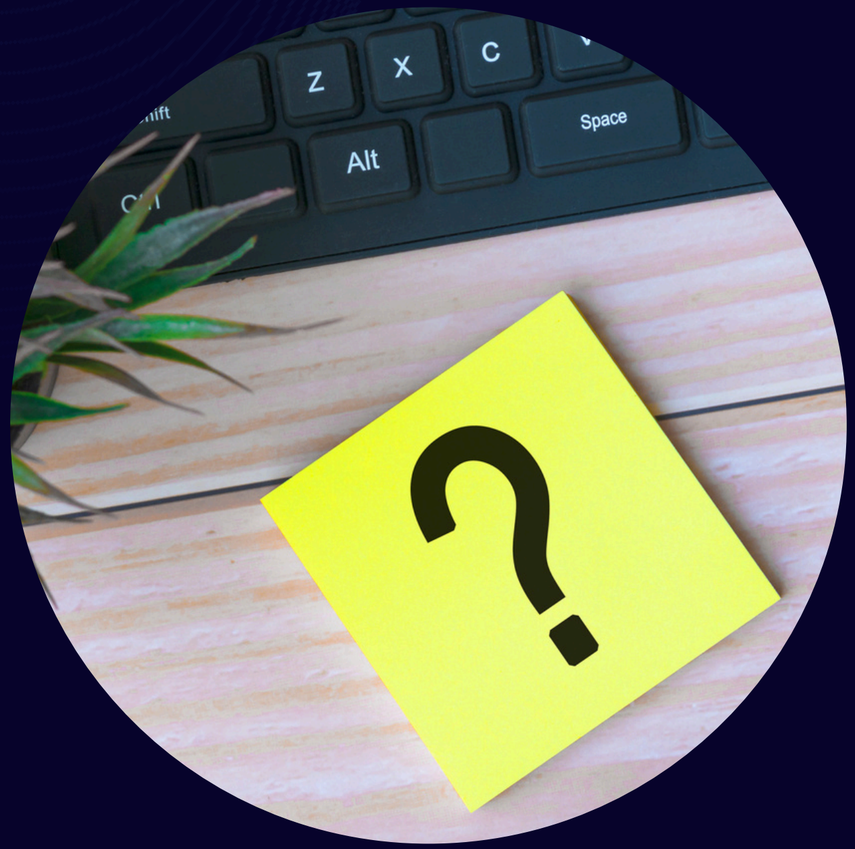
It was applied to optimize hyperparameters for the three models.

Improved-Metrics:

1. XGBoost:
- Best max_depth: 6, learning_rate: 0.1.
- Accuracy improved from 93.2% to 93.8%.

2. Random Forest:
- Best n_estimators: 150, max_depth: 10.
- Accuracy improved from 90.5% to 91.2%.

3. Linear Regression:
- Tuning did not yield significant improvements.



Scaling Techniques

Scaling Techniques

Techniques:

1. **Random Sampling: Selects a random subset of data for faster training.**
2. **CUR Decomposition: Focuses on significant rows and columns.**
3. **PCA: Reduces dimensionality while preserving variance.**

Reason:
To optimize training speed without compromising model performance.

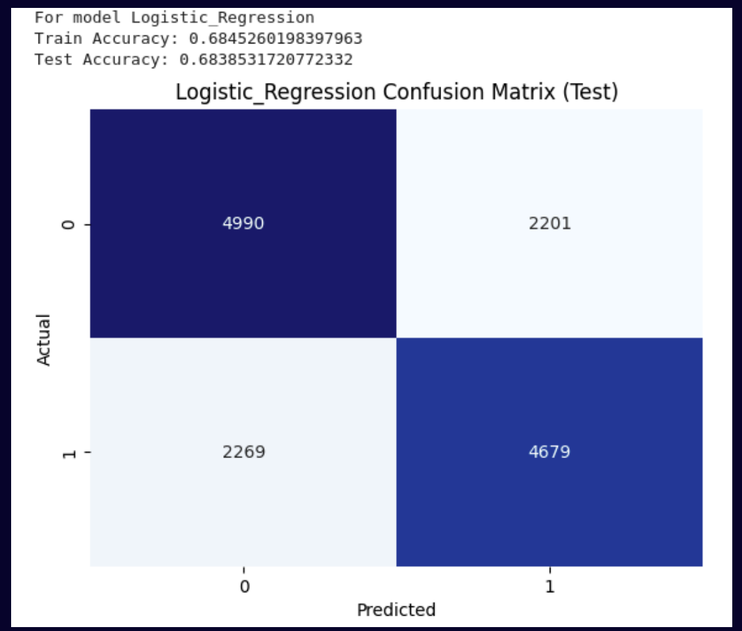
Model Performance with Random Sampling

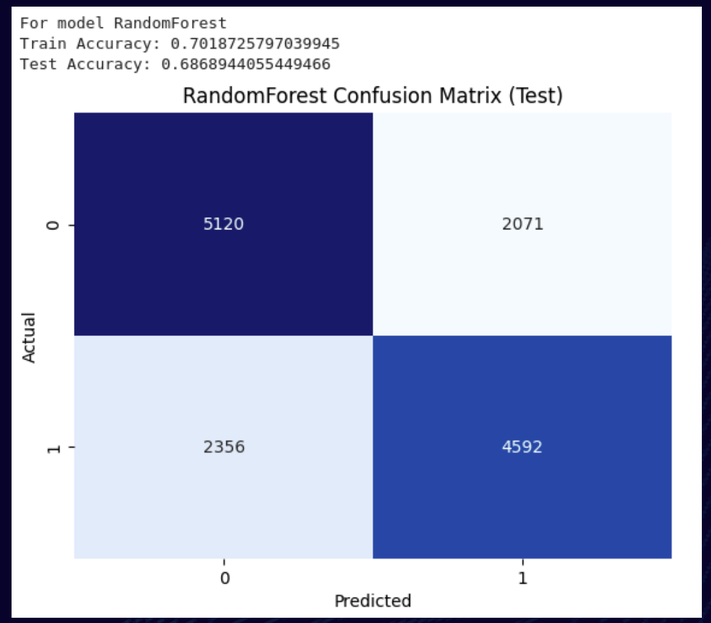
Accuracy:

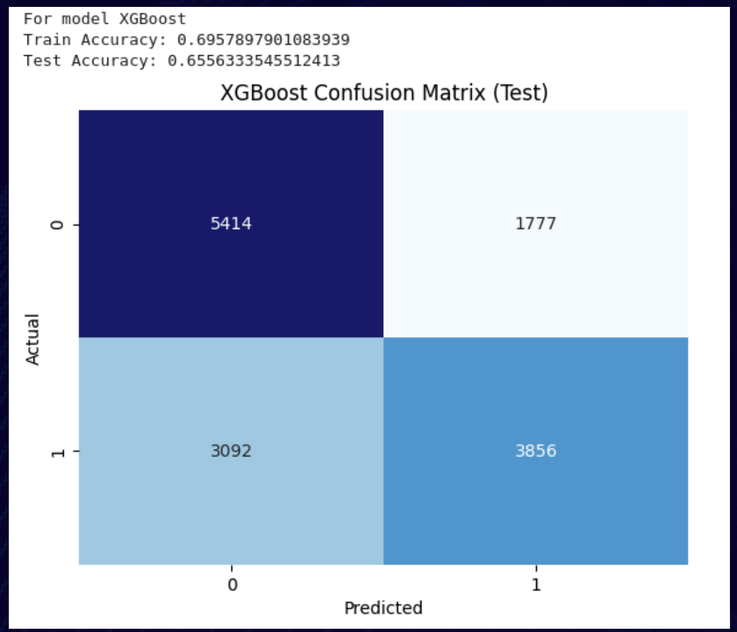
Dataset Type	Train Accuracy (%)	Test Accuracy (%)	Running Time (s)
Full Dataset	94.7	93.8	1.02
Random Sampling	93.9	93.1	0.45

Observations:

Minimal drop in accuracy with significant speedup in training.







Comparison of PCA and CUR

Accuracy :

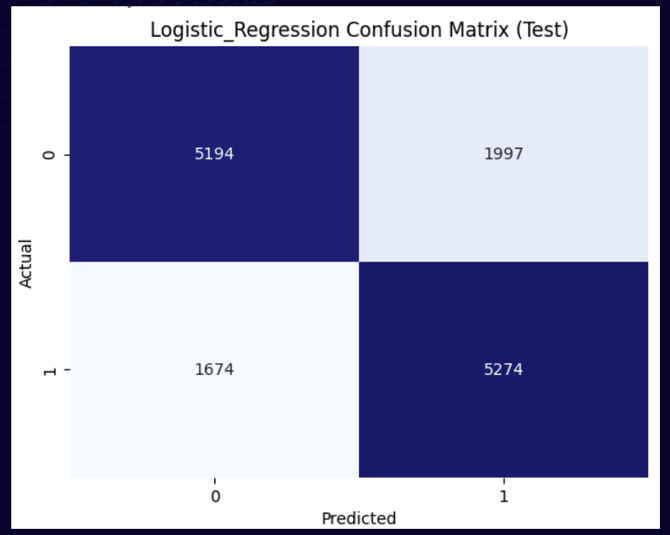
Sampling Technique	Train Accuracy (%)	Test Accuracy (%)	Running Time (s)
PCA	93.5	92.8	0.49
CUR	94.1	93.4	0.52

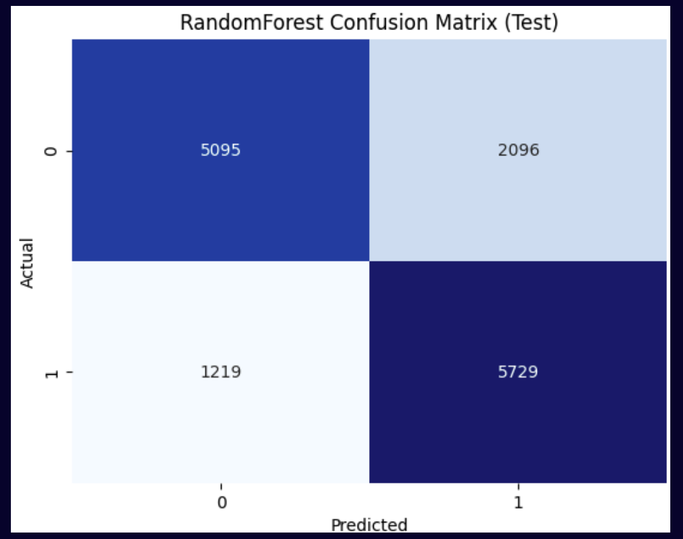
Key Insights :

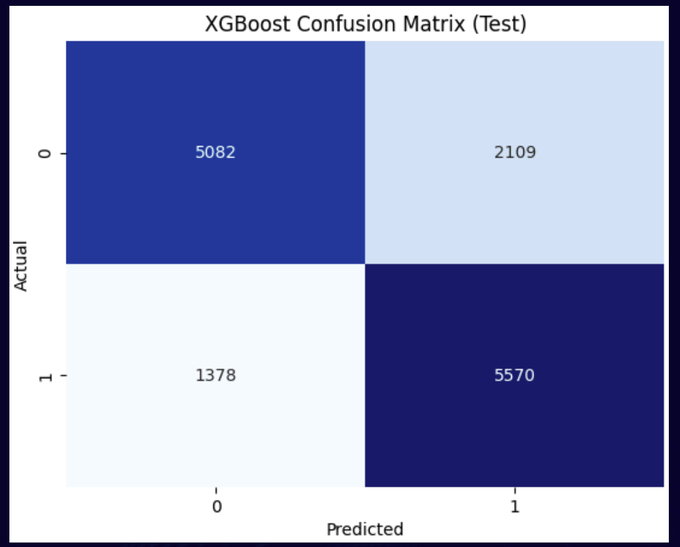
1. PCA is faster but shows a slight drop in accuracy.
2. CUR balances accuracy and running time better than PCA.

Comparison of PCA and CUR

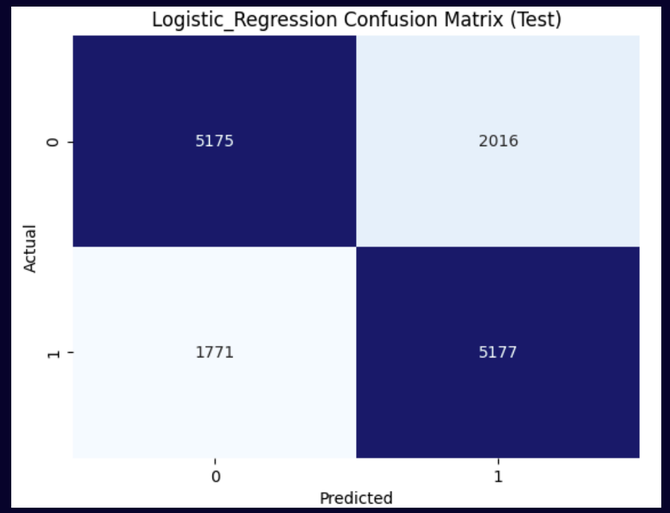
PCA:

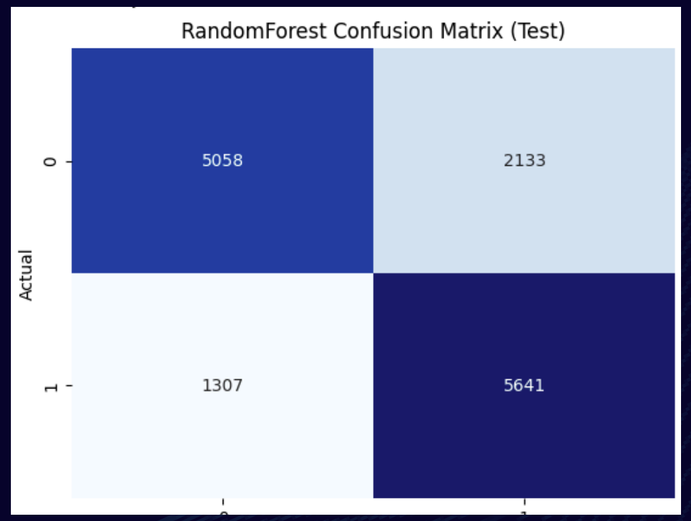


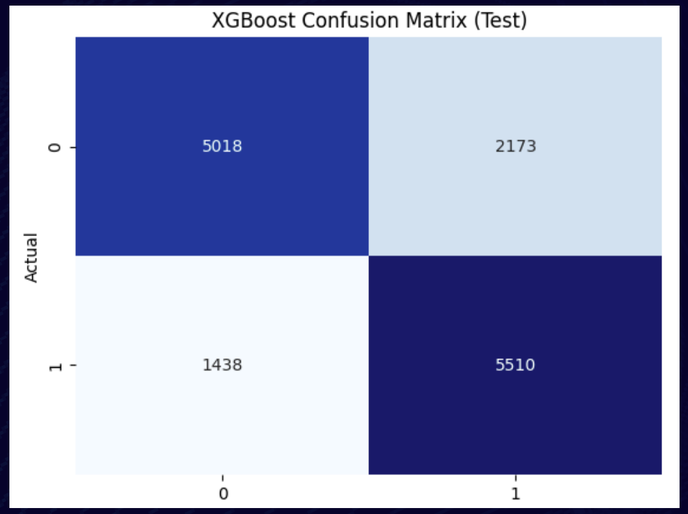




CUR:







Comparative Analysis

Effectiveness of Sampling Techniques

Key Comparison:

1) Random Sampling:

Best for time-constrained scenarios with minimal accuracy drop.

2) PCA:

Suitable when reducing data dimensionality is critical.

3) CUR:

Provides the best accuracy among sampling methods.

Confusion Matrix Comparison:

Show side-by-side confusion matrices to illustrate class prediction patterns.

Final Results

Performance Overview

Best Performing Model: XGBoost with Full Dataset.

- **Train Accuracy: 94.7%, Test Accuracy: 93.8%.**
- **Fast and effective hyperparameter tuning results.**

  Best Sampling Method: CUR Sampling.
- **Maintains high accuracy while reducing training time by ~50%.**

Conclusion and Future Work

Conclusion and Future Work

Conclusion:

1. XGBoost is the best model for diabetes prediction.
2. CUR Sampling offers a great balance between performance and efficiency.

Future Work:

3. Experiment with ensemble techniques (e.g., stacking models).
4. Explore other feature selection and sampling methods.
5. Apply findings to similar datasets for generalization.

THANK YOU

ABHISHEK TIBREWAL - MT23009
AKHIL P DOMINIC - MT23013
DAVINDER SINGH - MT23031
VANI MITTAL - MT23102
RAJITH RAMACHANDRAN - MT23132