Predicting Diabetic Outcomes

By Hana Wasif, Agustin Rodriguez, Floris Cash, Latifah Jones, and Jessica Velasquez

Project Overview

Objective:

- *"To predict diabetes outcomes using patient health data, identify key factors contributing to diabetes, and evaluate machine learning models to determine the best-performing approach."*

Dataset:

- *"We used a public dataset containing patient health information with features like BMI, GeneralHealth, and Diabetes outcomes."*

Approach:

- **Data Preprocessing:** Cleaned and balanced the dataset.
- **Exploratory Data Analysis:** Uncovered relationships between features and diabetes.
- **Modeling:** Compared Logistic Regression and Random Forest for classification.
- **Evaluation:** Used metrics like accuracy and ROC AUC to select the best model.

Key Question:

- *"Which factors (e.g., BMI, genetics, health habits) most influence diabetes, and which model performs better at predicting outcomes?"*

Data Cleaning and Preprocessing

Overview of the Dataset:

- Original dataset: 237,630 rows and 35 columns.
- Contains patient health records, with features like BMI, GeneralHealth, and `HadDiabetes` (target variable).

Key Cleaning Steps:

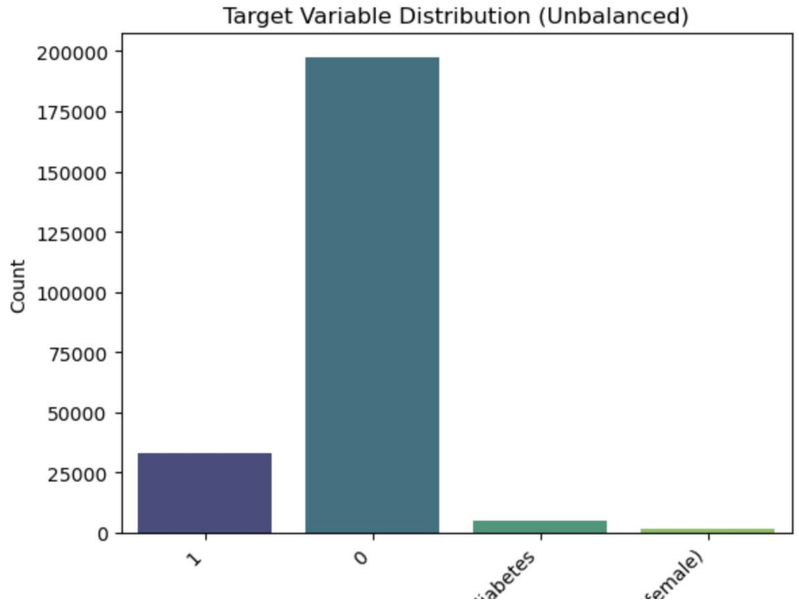
- Handled missing values (e.g., median imputation for `BMI`).
- Dropped irrelevant columns such as `PatientID`.
- Addressed outliers in numerical columns where necessary.

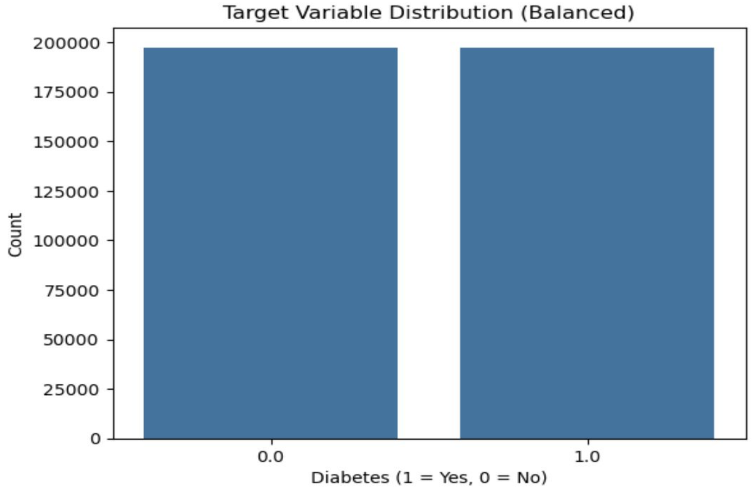
Preprocessing Steps:

- Encoded categorical variables (e.g., `GeneralHealth`) using OneHotEncoder.
- Scaled numerical features (`BMI`, `WeightInKilograms`) using StandardScaler.
- Balanced the target variable (`HadDiabetes`) using oversampling to ensure equal representation of both classes.

Result After Preprocessing:

- Final dataset: 237,630 rows and 34] columns.
- Balanced classes: Equal number of 1s (diabetes) and 0s (non-diabetes).





Exploratory Data Analysis

Exploratory Data Analysis helps us identify relationships, patterns, and trends in the dataset to prepare for modeling.

Key Insights from the Heatmap:

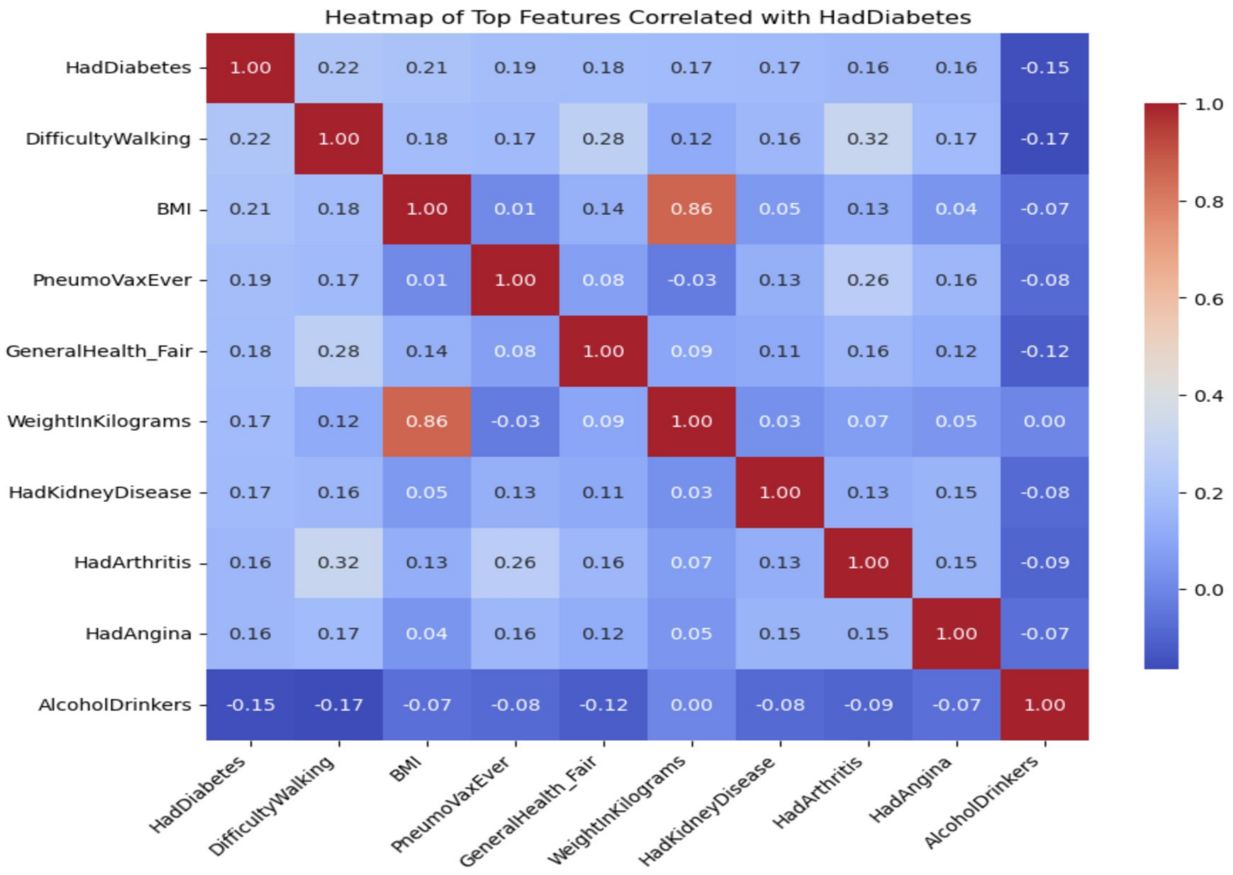
- The heatmap shows the top features correlated with `HadDiabetes` (target variable).

Top Correlations:

- `DifficultyWalking`: Strong positive correlation with diabetes.(0.22)
- `BMI` and `WeightInKilograms`: Show moderate positive correlations.(0.21)
- `HadKidneyDisease` and `HadArthritis`: Moderate predictors.

Negative Correlation:

- `AlcoholDrinkers` is weakly negatively correlated with diabetes (-0.15)



Heatmap of Top Features Correlated with HadDiabetes

	HadDiabetes	DifficultyWalking	BMI	PneumoVaxEver	GeneralHealth_Fair	WeightInKilograms	HadKidneyDisease	HadArthritis	HadAngina	AlcoholDrinkers
HadDiabetes	1.00	0.22	0.21	0.19	0.18	0.17	0.17	0.16	0.16	-0.15
DifficultyWalking	0.22	1.00	0.18	0.17	0.28	0.12	0.16	0.32	0.17	-0.17
BMI	0.21	0.18	1.00	0.01	0.14	0.86	0.05	0.13	0.04	-0.07
PneumoVaxEver	0.19	0.17	0.01	1.00	0.08	-0.03	0.13	0.26	0.16	-0.08
GeneralHealth_Fair	0.18	0.28	0.14	0.08	1.00	0.09	0.11	0.16	0.12	-0.12
WeightInKilograms	0.17	0.12	0.86	-0.03	0.09	1.00	0.03	0.07	0.05	0.00
HadKidneyDisease	0.17	0.16	0.05	0.13	0.11	0.03	1.00	0.13	0.15	-0.08
HadArthritis	0.16	0.32	0.13	0.26	0.16	0.07	0.13	1.00	0.15	-0.09
HadAngina	0.16	0.17	0.04	0.16	0.12	0.05	0.15	0.15	1.00	-0.07
AlcoholDrinkers	-0.15	-0.17	-0.07	-0.08	-0.12	0.00	-0.08	-0.09	-0.07	1.00

Importance of Findings:

- These correlations help refine feature selection for Logistic Regression and Random Forest models.
- Highlights the importance of `DifficultyWalking`, `BMI`, and other significant predictors.

Random Forest Classifier and Analysis

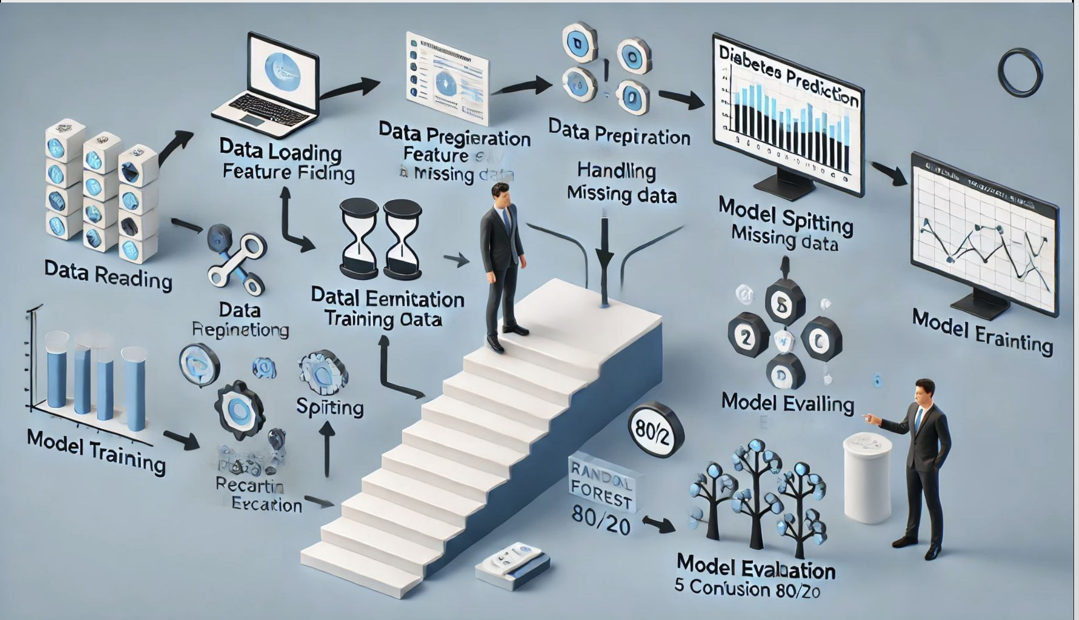
Data Preparation: Loaded data, handled missing values, and one-hot encoded categorical variables.

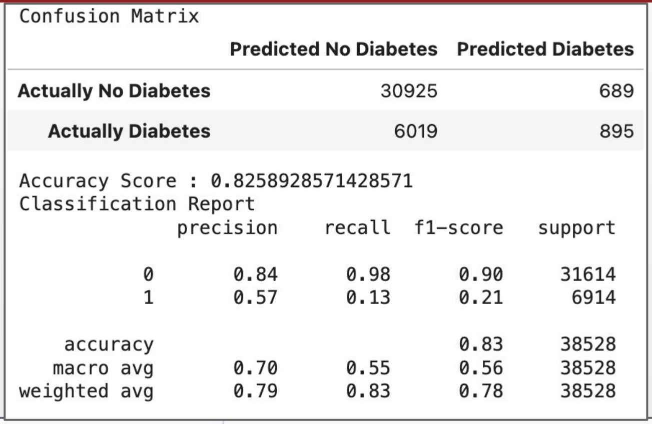
Data Splitting: Split dataset into 80% training and 20% testing sets.

Model: Used Random Forest Classifier with 100 estimators.

Performance: Achieved 0.83. Evaluated using classification report and confusion matrix.

Key Takeaway: A robust, modular pipeline adaptable for similar classification tasks.

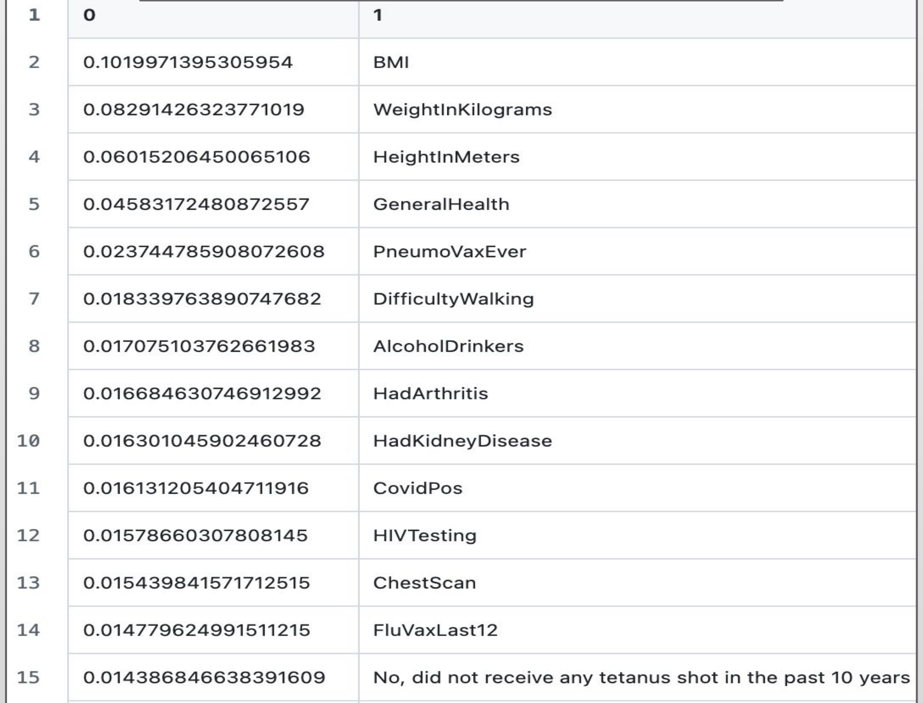




```
Confusion Matrix
```

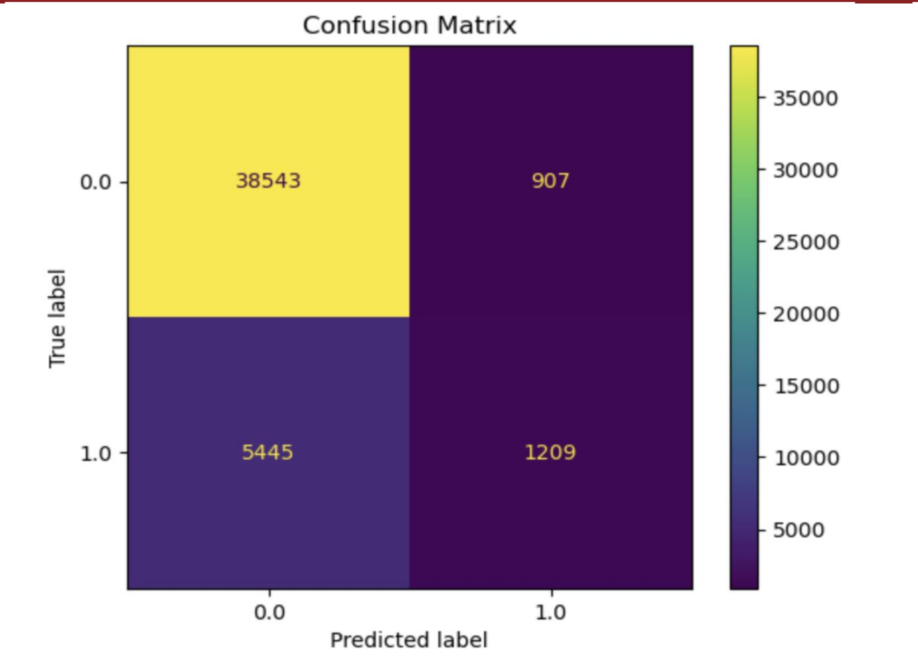
	Predicted No Diabetes	Predicted Diabetes
Actually No Diabetes	30925	689
Actually Diabetes	6019	895

```
Accuracy Score : 0.8258928571428571
Classification Report
              precision    recall  f1-score   support

           0       0.84      0.98      0.90     31614
           1       0.57      0.13      0.21      6914

    accuracy                           0.83     38528
   macro avg       0.70      0.55      0.56     38528
weighted avg       0.79      0.83      0.78     38528
```

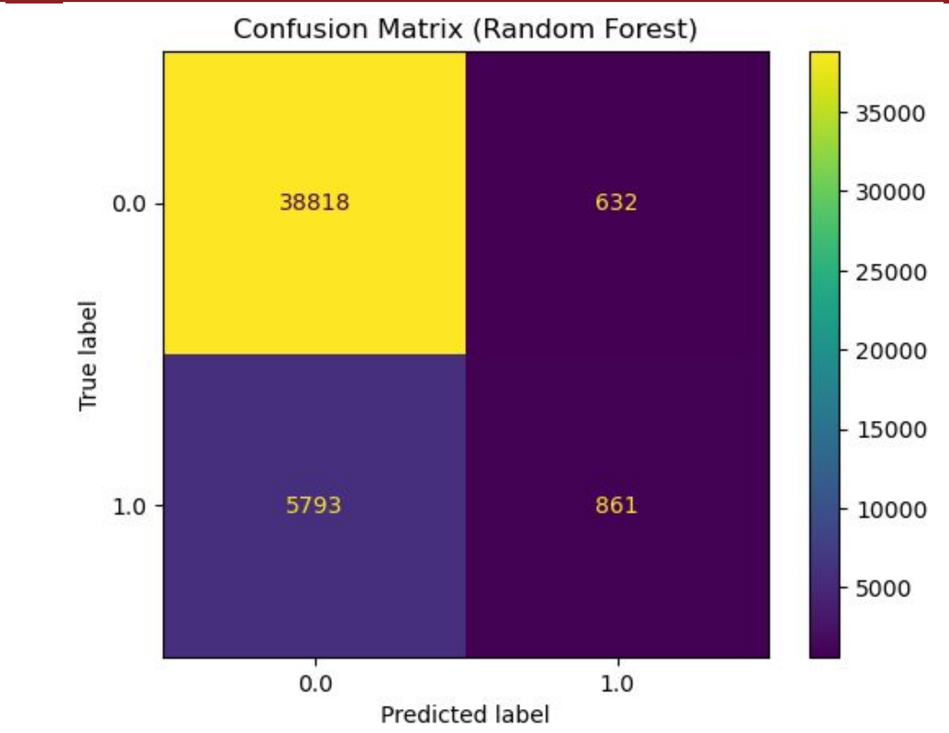


1	0	1
2	0.1019971395305954	BMI
3	0.08291426323771019	WeightInKilograms
4	0.06015206450065106	HeightInMeters
5	0.04583172480872557	GeneralHealth
6	0.023744785908072608	PneumoVaxEver
7	0.018339763890747682	DifficultyWalking
8	0.017075103762661983	AlcoholDrinkers
9	0.016684630746912992	HadArthritis
10	0.016301045902460728	HadKidneyDisease
11	0.016131205404711916	CovidPos
12	0.01578660307808145	HIVTesting
13	0.015439841571712515	ChestScan
14	0.014779624991511215	FluVaxLast12
15	0.014386846638391609	No, did not receive any tetanus shot in the past 10 years

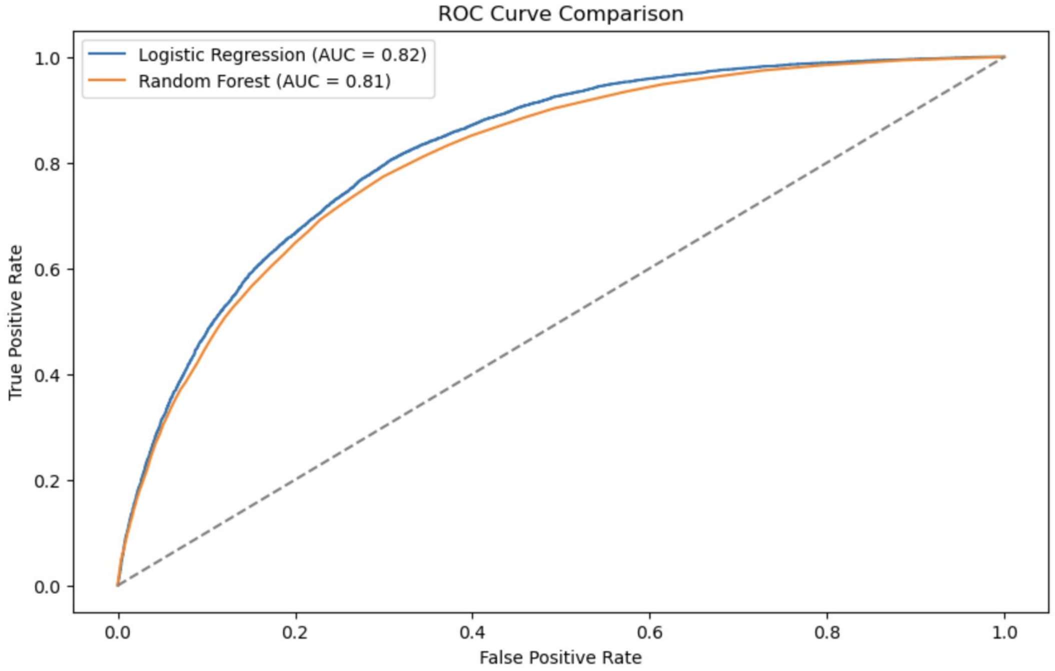
Modeling



Model Performance: Logistic Regression



Evaluation Of Data



ROC Curve Analysis

1. **Logistic Regression:**
   - The Area Under the Curve (AUC) for Logistic Regression is 0.82.
   - This indicates that the model performs well in distinguishing between positive (diabetes) and negative (non-diabetes) cases.
2. **Random Forest:**
   - The AUC for Random Forest is 0.81, which is slightly lower than Logistic Regression.
   - Despite its ability to capture complex relationships, Random Forest does not outperform Logistic Regression in this case.
3. **Insights:**
   - Both models perform comparably, with Logistic Regression slightly edging out Random Forest based on the AUC score.
   - Logistic Regression may be favored for its simplicity and interpretability in this scenario.

Conclusion

- Our project highlights the power of data-driven insights in healthcare by analyzing and predicting diabetes outcomes using Logistic Regression and Random Forest models. Through thorough Exploratory Data Analysis, we identified key predictors such as BMI, DifficultyWalking, and WeightInKilograms. By addressing data imbalance with oversampling, we improved model accuracy and ensured fair and reliable predictions, contributing to early detection and prevention of diabetes.