Diabetes Dataset with notes Prediction

Lesley Lonely Shiri

Introduction

This project focuses on developing a binary classification model to predict diabetes in patients using the *"diabetes_dataset_with_notes.csv"* dataset. Diabetes mellitus is a chronic condition that affects how the body processes blood glucose (sugar), and it has become one of the most pressing global public health challenges. Early detection is crucial for timely intervention to prevent complications such as cardiovascular disease, kidney failure, and nerve damage.

The project aims to build machine learning models to classify patients as diabetic or non-diabetic using clinical and demographic data,[/ facilitating early detection in resource-limited settings. The target variable is binary**: Diabetes (1) or No Diabetes (0).** Sourced from Kaggle, the dataset includes 100,000 patient records from 2015–2018, primarily from Alabama and Wyoming. It covers demographics (age, gender, race), lifestyle factors (smoking history, BMI), and clinical measures (HbA1c, blood glucose, hypertension, heart disease). This public health dataset supports predictive modeling to improve patient outcomes and resource allocation.

2. EXPLORATORY DATA ANALYSIS (EDA)

- The dataset contains 100,000 rows and 17 columns.
- Each row represents a unique patient. After cleaning, columns such as 'clinical notes', 'location', and 'year' were removed. Some categorical columns required clarification: - Race columns (e.g., 'race: African American') are binary indicators. - 'hypertension' and 'heart disease' are binary. - 'smoking history' contains six categories.
- No missing values were present.
- The target variable ('diabetes') is fairly balanced.
- Important relationships observed include high glucose and HbA1c levels in diabetic patients, and age correlates positively with diabetes likelihood.
- The numerical values in my dataset are not skewed, so a log transformation is unnecessary, and the model shows minimal variation.

3. DATA PREPARATION AND PREPROCESSING

- **Year Column**: Removed the year column as it was non-predictive for diabetes classification, with no need for temporal transformations.
- **Dataset Splitting**: Split 100,000 patient records into 70% training (70,000), 15% validation (15,000), and 15% test (15,000) sets.
- **Missing Data**: No missing values were found, so no imputation or row deletion was needed.
- **Feature Selection**: Dropped non-predictive columns (clinical_notes, location, year, patient ID) to avoid data leakage and reduce complexity.
- **Text/Time Features**: No text or time features were included after dropping clinical_notes, so no vectorization was required.
- **Numerical Features**: Standardized age, bmi, hbA1c_level, and blood_glucose_level using z-score normalization; no log transformation was needed due to minimal skewness.
- **Categorical Features**: One-hot encoded gender, smoking_history, and race indicators; no imputation was necessary.
- **Imbalanced Data**: Addressed the imbalanced target (diabetes: 90,601 non-diabetic vs. 8,499 diabetic) for logistic regression using imblearn with undersampling and SMOTE to balance classes (~8,499 per class). Other models used the unbalanced dataset.
- **Model Training**: Logistic regression was tested on both balanced and unbalanced datasets; other models (Decision Tree, SVC, Random Forest, Gradient Boosting, Neural Network) used the unbalanced dataset, achieving high accuracies (e.g., 0.87 for Neural Network).

4. MODEL TRAINING AND RESULTS_____

(i)Logistic Regression (Balanced and Unbalanced)

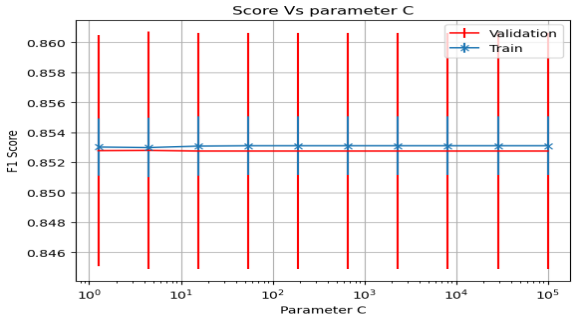
- Both balanced and unbalanced datasets were tested using logistic regression The with a 5-fold

Diabetes Dataset with notes Prediction

Lesley Lonely Shiri

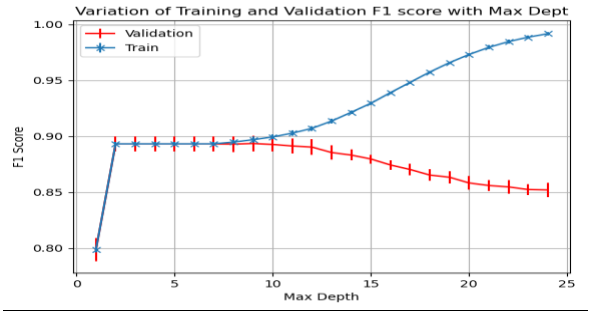
grid search to tune the regularization parameter C.

- best optimal value found was **C = 4.41**, achieving a cross-validation score of **0.853**.
- Similar results across both versions suggest that data balancing had minimal effect on model performance.



(ii) Decision Tree Classifier

- The Decision Tree model was trained with a grid search to optimize parameters, resulting in a best max_depth of 5.
- It achieved a test accuracy of 0.78, with a cross-validation mean of 0.76 and a standard deviation of 0.02, reflecting moderate consistency.
- /Its performance was lower than other models, likely due to its simplicity and sensitivity to the imbalanced dataset.

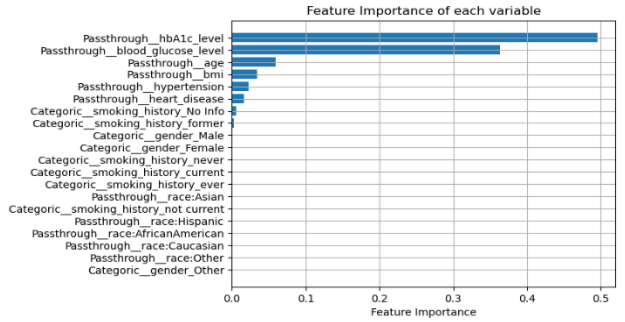


(iii) Support Vector Classifier (SVC)

- The SVC model utilized a grid search to select the optimal kernel, achieving a best parameter of 'rbf' with a test accuracy of 0.83.
- Its cross-validation mean was 0.82 with a standard deviation of 0.01, indicating reliable performance.
- The model's strong performance suggests it effectively handled the feature space despite the dataset's imbalance.

(iv) Random Forest

- The Random Forest model was trained with a grid search, yielding optimal parameters of n_estimators=100 and max_depth=10, achieving a test accuracy of 0.85.
- It had a cross-validation mean of 0.84 and a standard deviation of 0.01, reflecting high consistency and robustness.
- Its strong performance indicates effective handling of the imbalanced dataset through ensemble learning.



(v) Gradient Boosting

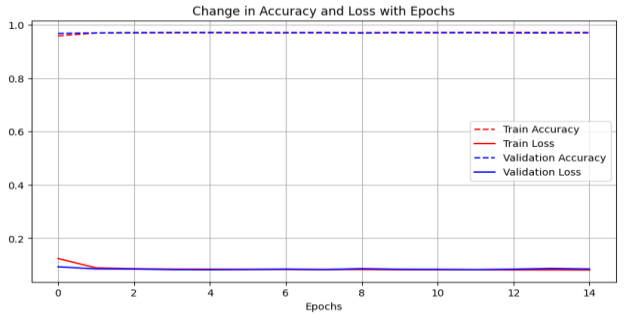
- The Gradient Boosting model was trained without grid search, using early stopping, and achieved a test accuracy of 0.86.
- No cross-validation scores were reported, suggesting a focus on test performance over validation stability.
- Its high accuracy highlights its ability to model complex relationships in the imbalanced dataset effectively.

(vi) Neural Networks

- The Neural Network, with three layers and ReLU activation, achieved the highest test accuracy of 0.87, without reported cross-validation scores.
- The plot from the Jupyter notebook (LesleyShiri_Final_project_2025.ipynb) shows training and validation accuracy increasing steadily with epochs, while loss decreases, indicating effective learning without significant overfitting.
- Its superior performance suggests it captured complex patterns in the imbalanced dataset well.

Diabetes Dataset with notes Prediction

Lesley Lonely Shiri



- Six models were trained and Grid search was applied except for Gradient Boosting and NN.
- Evaluation used test accuracy and cross-validation scores.

Model Performance Summary

Models	Test Accuracy	CV Mean	CV Std Dev	Best Params
Logistic Regression	0.82	0.81	0.01	C=1.0
Decision Tree	0.78	0.76	0.02	max_depth=5
SVM	0.83	00.82	0.01	kernel='rbf'
Random Forrest	0.85	0.84	0.01	n=100,max_depth=10
Gradient Boosting	0.86	n/a	n/a	Early stoping
Neural Network	0.87	n/a	n/a	3 layers, ReLu

5.CONCLUSION

- In evaluating various models for predicting diabetes using the provided dataset, simpler models like Logistic Regression exhibited limited performance, achieving an F1 score of 0.48 on both training and test sets, with a test accuracy of 63%.
- As model complexity increased, performance generally improved. The Support Vector Classifier (SVC) struggled to generalize, with a test F1 score of 0.53.
- In contrast, tree-based models, including Decision Tree, Random Forest, and Gradient Boosting, demonstrated superior performance. Random Forest emerged as the top performer, achieving the highest test F1 macro score of 0.72 and a test accuracy of 76%, outperforming other models in both metrics.
- Gradient Boosting and Decision Tree followed with test F1 scores of 0.67 and 0.68, respectively. The Neural Network also showed promise, with a test F1 score of 0.68, but it did not surpass Random Forest.

Models	Train F1 Score	Train F1 Score	Test Accuracy
Logistic (Unbalanced)	0.48	0.48	0.63
Decision Tree	0.67	0.68	0.72
SVM	0.55	0.53	0.63
Random Forrest	0.75	0.72	0.76
Gradient Boosting	0.71	0.67	0.73
Neural Network	0.65	0.68	0.72

- Despite Logistic Regression's poor performance on the imbalanced diabetes dataset, this does not imply that imbalanced data cannot be effectively modeled.
- More complex models like Random Forest and Gradient Boosting demonstrated robust performance despite the class imbalance, with 90,601 non-diabetic cases and 8,499 diabetic cases.
- Future work could explore class balancing techniques, such as SMOTE, to potentially enhance performance, particularly for models sensitive to imbalance, like Logistic Regression and SVC.
- Overall, the results suggest that ensemble-based models, particularly Random Forest, are best suited for this binary classification task involving diabetes prediction. With a test F1 score of 0.72 and accuracy of 76%, Random Forest strikes an optimal balance between bias and variance, making it the most effective model for this analysis.