Diabetes Prediction Over Telephonic Health Survey

Name:	**Smita Sarkar**
Roll No.:	2311005
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	August 17, 2023
Date of Submission:	November 19, 2023

1 Introduction

Diabetes is one of the most prevalent chronic diseases across various countries affecting millions of individuals globally, imposing a significant burden on both healthcare systems and economies. It has a negative impact on the life expectancy and quality of life of persons who are diagnosed with the illness. For successful diabetes prevention and control, accurate and timely diabetes identification is essential.

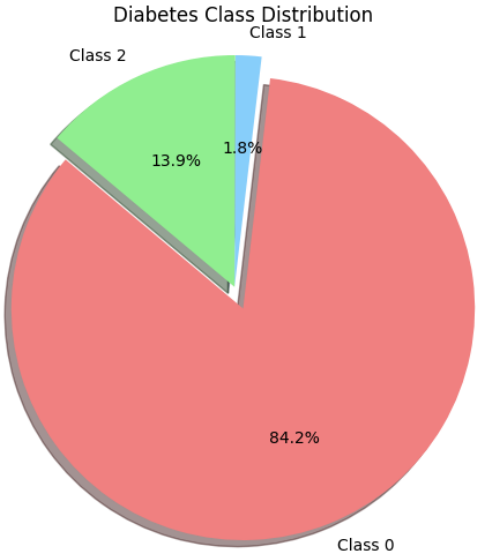
The **Behavioral Risk Factor Surveillance System (BRFSS)** is an extensive health-related telephone survey conducted annually by the Centers for Disease Control in the USA. This survey collects valuable data from over 400,000 US citizens on various health-related risk behaviors, chronic health conditions, and the utilization of preventative services. One pressing question that arises is whether the data collected through the BRFSS can be harnessed effectively for the prediction of diabetes using machine learning techniques.

This project aims to investigate the feasibility of developing diabetes risk prediction models using telephone health surveys, therefore contributing to global efforts in diabetes prevention and management. To accomplish this, we embark on a multifaceted journey involving data analysis, feature selection, and machine learning.

Data Description: The dataset used in this analysis contains 21 features that have been collected through the BRFSS, namely Diabetes_binary, HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, and Income.

The target variable, Diabetes Status, is the focal point of this analysis. It classifies individuals into one of three distinct classes, each indicative of a different diabetes status:

Class	**No. of instances**
0 (No diabetes or only during pregnancy)	192333
2 (Prediabetes)	31811
1 (Diabetes)	4168



2 Methods

2.1 Proposed Methods

There were 20456 duplicate values in the training dataset which were removed to get 207856 unique data instances. Then, several traditional machine learning classifiers such as Logistic Regression, Decision tree and Random Forest were explored. Aditionally, ensemble methods such as Adaptive Boosting (AdaBoost) with different base classifiers like Decision Trees and Random Forests were also employed. [2] Furthermore, various feature selection methods were employed to enhance model performance. The dataset was oversampled using Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance and improve model generalization. The goal was to assess the performance of these models in predicting diabetes based on features obtained from telephonic health surveys. [1] Please find my code here.

2.2 Parameter Tuning

For optimal model performance, hyperparameter tuning was crucial. Grid search with cross-validation was applied to find the best hyperparameters for the classifiers. This iterative process helps enhance the model's predictive capabilities. The parameter grids were defined for each classifier to explore different combinations efficiently. [3]

3 Experimental Setup

3.1 Evaluation Criteria

Various evaluation criteria were used for assessing the model performance such as precision, recall, accuracy and F1-score. These metrics provide a comprehensive understanding of the model's ability to correctly identify positive and negative instances, considering both false positives and false negatives.

3.2 Experimental Settings

The experiments were conducted using several libraries like numpy, pandas, matplotlib, scikit-learn, imbalanced-learn, etc in Python. The dataset was split into training (80%) and testing (20%) sets, and the models were evaluated using stratified k-fold (5 folds) cross-validation to ensure robust results. Four feature selection methods, namely SelectFromModel, Mutual Information, Recursive Feature Elimination (RFE) and Recursive Feature Elimination with Cross-Validation (RFECV), were used to select relevant features (10).

4 Results and Discussions

Results before applying SMOTE:

Table 1: Performance of Different Classifiers using SelectFromModel

Classifier	Precision	Recall	Accuracy	F1-score
Adaptive Boosting	0.7846	0.8329	**0.8329**	**0.7872**
Decision Tree	0.7807	0.8301	**0.8301**	**0.7895**
Logistic Regression	0.7680	0.8267	**0.8267**	**0.7733**
Random Forest	0.7845	0.8329	**0.8329**	**0.7832**

Table 2: Performance of Different Classifiers using RFE

Classifier	Precision	Recall	Accuracy	F1-score
Adaptive Boosting	0.7848	0.8330	**0.8330**	**0.7878**
Decision Tree	0.7803	0.8299	**0.8299**	**0.7890**
Logistic Regression	0.7672	0.8261	**0.8261**	**0.7742**
Random Forest	0.7844	0.8328	**0.8328**	**0.7843**

Table 3: Performance of Different Classifiers using Mutual Information

Classifier	Precision	Recall	Accuracy	F1-score
Adaptive Boosting	0.7819	0.8318	**0.8318**	**0.7854**
Decision Tree	0.780	0.830	**0.830**	**0.788**
Logistic Regression	0.775	0.828	**0.828**	**0.784**
Random Forest	0.782	0.832	**0.832**	**0.781**

Table 4: Performance of Different Classifiers using RFECV

Classifier	Precision	Recall	Accuracy	F1-score
Adaptive Boosting	0.7770	0.8301	**0.8301**	**0.7757**
Decision Tree	0.7804	0.8307	**0.8307**	**0.7868**
Logistic Regression	0.7773	0.8295	**0.8295**	**0.7841**
Random Forest	0.7831	0.8323	**0.8323**	**0.7837**

Results after applying SMOTE:

Table 5: Performance of Different Classifiers using SelectFromModel

Classifier	Precision	Recall	Accuracy	F1-score
Adaptive Boosting	0.5791	0.5777	**0.5777**	**0.5782**
Decision Tree	0.8224	0.8240	**0.8240**	**0.8207**
Logistic Regression	0.4936	0.4944	**0.4944**	**0.4934**
Random Forest	0.8608	0.8600	**0.8600**	**0.8581**

Table 6: Performance of Different Classifiers using RFE

Classifier	Precision	Recall	Accuracy	F1-score
Adaptive Boosting	0.5786	0.5768	**0.5768**	**0.5773**
Decision Tree	0.8119	0.8138	**0.8138**	**0.8104**
Logistic Regression	0.4914	0.4896	**0.4896**	**0.4817**
Random Forest	0.8438	0.8431	**0.8431**	**0.8407**

Table 7: Performance of Different Classifiers using Mutual Information

Classifier	Precision	Recall	Accuracy	F1-score
Adaptive Boosting	0.5489	0.5483	**0.5483**	**0.5483**
Decision Tree	0.652	0.651	**0.651**	**0.649**
Logistic Regression	0.504	0.507	**0.507**	**0.500**
Random Forest	0.660	0.658	**0.658**	**0.656**

Table 8: Performance of Different Classifiers using RFECV

Classifier	Precision	Recall	Accuracy	F1-score
Adaptive Boosting	0.6060	0.6042	**0.6042**	**0.6048**
Decision Tree	0.8430	0.8436	**0.8436**	**0.8407**
Logistic Regression	0.5359	0.5381	**0.5381**	**0.5364**
Random Forest	0.8912	0.8895	**0.8895**	**0.8887**

Observations:

- Decision tree is giving the best results before applying SMOTE, using all the four feature selection methods which were used, in terms of f1 score.
- Applying SMOTE significantly impacts precision, recall, accuracy, and F1-score.
- Adaptive Boosting shows a decrease in performance, while Decision Tree and Random Forest exhibit improvement.
- Logistic Regression's performance is adversely affected.
- Feature selection using mutual information after applying SMOTE is also giving poor results.
- Random Forest classifier, when coupled with RFECV after employing SMOTE, demonstrated the best F1 score of **0.8887**. After hyperparameter tuning, the best hyperparameters were found to be: {'max_depth': None, 'n_estimators': 200}

Analysis:

- The Random Forest model, with 200 estimators and unlimited tree depth, performed exceptionally well.
- The large number of estimators can lead to robust ensemble learning, capturing diverse patterns in the data.
- Not restricting the maximum depth allows the trees to grow deeper, potentially improving the model's ability to generalize to complex data.

5 Conclusion

The comprehensive evaluation and comparison of classifiers provide valuable insights for healthcare practitioners and public health officials aiming to implement effective diabetes prevention and management strategies.
Our study contributes to the ongoing global efforts in leveraging machine learning for healthcare applications. Moving forward, further research could explore additional features, consider more advanced models, and delve deeper into the interpretability of the selected features.

References

[1] Lakshmi H.N., A. Srinivasa Reddy, and Kritika Naidu. Analysis of diabetic prediction using machine learning algorithms on brfss dataset. pages 1024–1028, 2023.

[2] Tasin I, Nabil TU, Islam S, and Khan R. Diabetes prediction using machine learning and explainable ai techniques. *Healthc Technol Lett*, 10(1-2):1–10, Dec 14 2022.

[3] Butt UM, Letchmunan S, Ali M, Hassan FH, Baqir A, and Sherazi HHR. Machine learning based diabetes classification and prediction for healthcare applications. *J Healthc Eng*, 2021:9930985, 2021.