Diabetes Health Analysis & Prediction

Introduction

The Behavioral Risk Factor Surveillance System (BRFSS) [1] is an annual health-related telephone survey conducted by the Centers for Disease Control and Prevention (CDC) since 1984. Each year, over 400,000 individuals participate in extensive surveys, answering questions regarding health-related risk behaviors, chronic health conditions, and the use of preventative services. Given the complexity and size of the dataset, extracting meaningful insights from it remains challenging.

The project aims to analyze the dataset with a specific focus on diabetes. We used various analytical techniques, including calculating data completeness, correlation rate as well as machine learning models to identify the most important factors related to diabetes. Additionally, we plotted our results to illustrate our key findings and facilitate a better understanding of the underlying patterns in the data. These insights can help inform the population about the risk factors associated with diabetes and highlight essential health strategies to prevent diabetes.

Methodology

To analyze the dataset, we utilized a Jupyter working environment along with PySpark.

First, we converted the format of the dataset to enable more efficient processing. Next, we performed data cleaning and preprocessing to ensure the dataset was ready for analysis. Afterwards, we conducted various analytical techniques to gain insights into the data. Finally, we implemented multiple machine learning models to predict the likelihood of diabetes in individuals and evaluated the performances of the models.

1) **Data form conversion**

   The original BRFSS dataset is available for download in either ASCII or SAS format. Given that working with SAS requires additional installations and that ASCII files are not optimized for PySpark, we opted to convert the ASCII file into Parquet format. Parquet is a columnar storage format that offers significant advantages for PySpark, including optimized storage through columnar compression and efficient querying that avoids the need to read entire rows when accessing specific column values.

   Given that the BRFSS dataset contains over 300 features, many of which are not relevant to diabetes, we carefully omitted these for more efficient processing. This reduction left us with over 100 features, which were further refined during the preprocessing stage.

2) **Preprocessing**

   In the following, the four main steps conducted in the preprocessing will be discussed.

   2.1 Data Completeness Percentage

   It was found that there are so many null values in some features while displaying the dataset. To identify the most completed features, i.e. the features with the non-null value,

the completeness percentage of each feature is calculated. After that, a new column, “response_rate_rank”, is added to indicate the rank of completeness of each feature.

2.2 Append Label Column & Calculate Correlation

To identify the most significant features associated with diabetes, we first added a column labeled "y" to classify individuals based on their diabetes status. In this classification, a value of 0 indicates no diabetes, while a value of 1 signifies the presence of diabetes, as derived from the "DIABETE4" feature in the BRFSS dataset. The question associated with "DIABETE4" was: “Have you ever been told you had diabetes?” The possible responses are outlined in the table below. We assigned a value of 1 to "y" when the response was 1 and a value of 0 when the responses were 2, 3, or 4. Additionally, values 7 and 9 were mapped to null to account for non-responses or irrelevant answers.

Value	**Value label**
1	Yes
2	During pregnancy
3	No
4	Pre-diabetic
7	don’t know
9	Refused to answer

Afterwards, to extract the most important features, we calculated the correlation of each feature with the target variable “y”. Besides, a new column, “Correlation_Rank”, is added to indicate the rank of correlation of each feature.

2.3 Feature Selection

With the completeness percentage data frame and correlation data frame, they are joined by using the “feature” column as the key and added a new column, “Sum_of_Rank”, is added to indicate the combined rank of each feature which considering both the completeness percentage and the correlation with the target variable. Moreover, the data frame is sorted by rank in descending order. The most useful features can be selected easily from the data frame. As a result, we selected the top 19 most important features that were also aligned in [2], as the paper conducted a literature survey on the most important factors related to diabetes.

2.4 Data Cleaning

Next, we performed data cleaning by removing rows that contained null values or had values of 9, 99, or 999 for most of the selected categories, as these were used to encode missing data. This process resulted in a dataset comprising 95574 rows and 20 columns (19 features + 1 target variable).

3) Machine learning models

The models were developed using the 19 features identified during the preprocessing phase. We first used one-hot encoding for the features "EMPLOY1", "ALCDAY4", "FALL12MN" as they were nominal data. The dataset was then split into training and test sets, with an 80-20 ratio. Then various hyper parameters were explored.

After training the model on the training data, we evaluated its performance by predicting whether diabetes was present in the test dataset. Finally, we concluded our analysis by examining the counts of true positives, true negatives, false positives, and false negatives. We also calculated the overall accuracy and weighted precision of the model to assess its effectiveness in predicting diabetes status.

The models that were used:

a. Linear Regression
   We tested various regularization parameters (0.001, 0.01, 0.1) and elastic net parameters (0, 0.001, 0.01, 0.1, 0.5). To identify the optimal parameters, we performed five-fold cross-validation.
b. Decision Tree
   The depth of the decision tree was explored by setting maximum depths of 5, 6, and 7, and adjusting the minimum number of instances per node to 1, 2, or 3. These hyperparameters were tuned to optimize the model's performance.
c. Logistic Regression
   We experimented with different regularization parameters (0.01, 0.1, 0.3), elastic net parameters (0, 0.5, 1.0), and maximum iterations (10, 20, 30). Cross-validation with five folds was used to select the optimal combination of these parameters.
d. Multilayer Perceptron
   We explored maximum iterations (15, 30) and block sizes (64, 128), using three-fold cross-validation to determine the best-performing configuration. The model's hidden layers and node count were set to 1 and 10, respectively.

Results

In this session, we present the results of the experiments we conducted for machine learning models. Since the results from the Linear Regression Model and Decision Tree Model are the most accurate, we dive deep into these models in the following.

We report the number of true positives and negatives as well as the false positives and negatives. In addition, we also present the (weighted) precision and accuracy of the models.

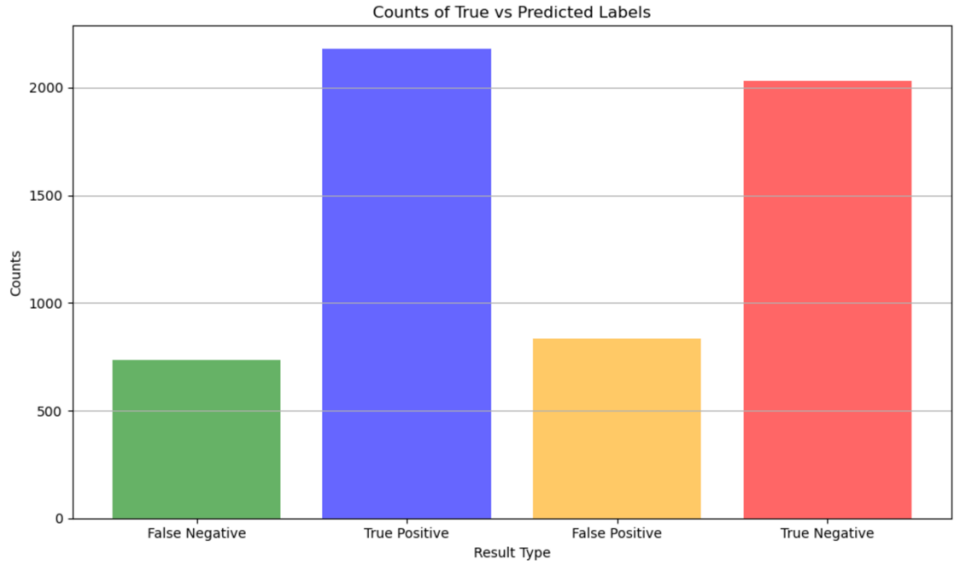
a. Linear Regression
   When conducting cross validation on our model, our root mean squared error on the test data was 0.4457 for our best model, which had 0.001 as the best regularization parameter

and 0.5 as the best elastic net parameter. We report the predicted values and the supposed values in the table below:

	True label 0	True label 1
Predicted label 0	16226 (true negative)	2783 (false negative)
Predicted label 1	102 (false positive)	189 (true positive)

The accuracy on the test data was 0.8505, while the weighted precision was 0.8222. We observed a significant number of false negatives compared to false positives. This discrepancy may be attributed to the dataset's imbalance, as there were considerably more individuals without diabetes than those with diabetes. To address this issue, we decided to conduct our analysis on a balanced subset of the dataset, consisting of 29359 rows with an equal distribution of individuals with and without diabetes.
After running the linear regression model on this balanced dataset, we obtained a more even distribution of predictions, as shown in the table below:

	True label 0	True label 1
Predicted label 0	2032 (true negative)	746 (false negative)
Predicted label 1	835 (false positive)	2178 (true positive)



This gave us an accuracy of 0.7282 and a weighted precision of 0.7284.

b. Decision Tree

To model the prediction of diabetes, we employed a decision tree classifier, a widely used algorithm for classification tasks due to its interpretability and ability to handle both numerical and categorical data.

Three-fold cross-validation was used to fine-tune hyperparameters, specifically the maximum tree depth and the minimum number of instances per node. This process ensures the model is not overfitting while optimizing for generalization.

The best model identified during cross-validation had the following characteristics:

- Maximum Depth: 5
- Number of Nodes: 11
- Number of Classes: 2
- Features Used: 4

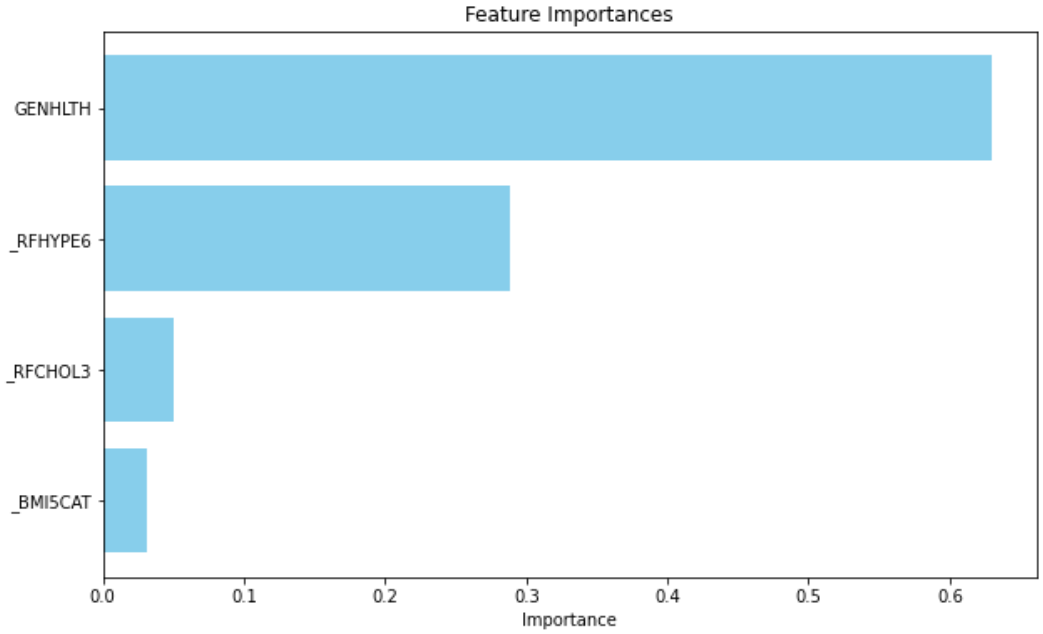
The model achieved an accuracy of 0.8501 on the test data, indicating good predictive performance, while the weighted precision was 0.8159. A detailed confusion matrix below highlights the classification results:

	True label 0	True label 1
Predicted label 0	15852 (true negative)	429 (false negative)
Predicted label 1	2448 (false positive)	507(true positive)

For this model, the three most significant predictors of diabetes were:

1. General Health (GENHLTH)
2. High Blood Pressure (_RFHYPE6)
3. High Cholesterol (_RFCHOL3)

In this diagram it is presented a general distribution of key features:



Conclusion

To sum up, we conducted an in-depth analysis of diabetes using the Behavioural Risk Factor Surveillance System (BRFSS) dataset, with the goal of uncovering key insights that could help us better understand and predict the prevalence of diabetes. We applied various analytical techniques, including calculating data completeness and correlation for each feature, and ranked the features to identify the most significant ones. Additionally, we built, trained, and evaluated several machine learning models. As a result, we developed two predictive models with an accuracy of 85% and identified critical features associated with diabetes.

References

[1] Data Source: Delaware Department of Health and Social Services, Division of Public Health, Delaware Behavioral Risk Factor Survey (BRFS), 2023: https://www.cdc.gov/brfss/annual_data/annual_2023.html

[2] Mohammad Mihrab Chowdhury, Ragib Shahariar Ayon, Md Sakhawat Hossain, An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset, Healthcare Analytics, Volume 5, 2024, 100297, ISSN 2772-4425: https://doi.org/10.1016/j.health.2023.100297