Introduction of project 3

Diabetes is a seriously pervasive chronic disease that disrupts the body’s ability to regulate blood glucose levels, leading to a diminished quality of life and reduced life expectancy. It stands as one of the most prevalent chronic illnesses in the United States, impacting millions of Americans annually and imposing a significant economic burden on the nation.

In this project, we will use the `diabetes binary health indicators` dataset obtained from [Kaggle]( <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/>) to conduct comprehensive exploratory data analysis (EDA) and develop predictive models. This dataset comprises 253,680 survey responses to the CDC’s BRFSS (Behavioral Risk Factor Surveillance System) from year 2015. The primary target variable, `Diabetes\_binary`, offers binary classification, distinguishing between 0 for no diabetes, and 1 for prediabetes or diabetes. This dataset encompasses 21 feature variables and is not balanced. Detailed information of variable can be found [here](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/?select=diabetes\_binary\_health\_indicators\_BRFSS2015.csv)

Our analysis will primarily focus on a subset of key variables, including High blood pressure (HighBP), High cholesterol (HighChol), cholesterol check (CholCheck), Body Mass Index (BMI), Smoker, Fruits, Veggies, and Age.

In our EDA phase, we will start by summarizing basic statistics visualizing variable frequencies. This will be followed by the exploration of correlations between variables and the creation of contingency tables to better understand the interplay of these factors.

Based on the results from EDA, we will split the dataset into training (70%) and test (30%) subsets for each educational level. Subsequently, we will employ the training data to fit six distinct models, including a logistic regression, a LASSO logistic regression, a classification tree model, a random forest model, a partial least squares model, as well as a **regularized logistic regression** model. The performance of these models will be rigorously evaluated using the test dataset, and we will determine the most effective model for predicting diabetes outcomes.

**What is log loss and why may we prefer it to things like accuracy?**

**Log loss**, also known as **logistic loss** or **cross-entropy loss**, is a loss function used in classification problems to evaluate the performance of a model, especially in scenarios with binary response variables. It quantifies the performance of a model by measuring the difference between the predicted probability distribution produced by the model and the actual probability distribution. Log loss is calculated as the negative logarithm of the predicted probability assigned to the true class label. It penalizes the model for incorrect predictions and a lower log loss value indicates better model predictive performance**.**

**Accuracy,** on the other hand, is a metric used to evaluate classification model performance by measuring the percentage of correctly classified instances. While accuracy is a useful metric, it can be misleading in some cases. For example, in imbalanced datasets where one class significantly outweighs the other, a model that always predicts the majority class will yield high accuracy but may not be practically useful. In such cases, log loss can offer a more reliable metric as it considers the model’s predictions uncertainty.

In summary, log loss is a preferred metric over accuracy when the data is imbalanced or when the cost of false positives and false negatives is different.

**Model summarizations:**

**What is a logistic regression and why we apply it to this kind of data?**

A logistic regression model is a statistical method used for analyzing the relationship between a binary outcome variable and one or more predictor variables. It is commonly used in classification problems, where the goal is to predict one of two possible outcomes, typically represented as 0 and 1 or “Yes” and “No”.

The relationship between response variable and the predictor variables is evaluated using the logistic function (also known as the sigmoid function), which ensures that the predicted probabilities fall between 0 and 1. The model parameters are estimated using a process called maximum likelihood estimation. The key output of a logistic regression model includes the coefficients (log-odds) associated with each predictor variable, which describe the strength and direction of their influence on the probability of the outcome.

Logistic regression is designed to handle **categorical dependent variables.** In our situation, Diabetes\_binary is a binary variable and is suitable for fitting logistic regression model.

**What is a LASSO logistic regression and why we might try to use it over basic logistic regression?**

LASSO (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the resulting statistical model. It uses glmnet package in R to fit the model.

The primary goal of LASSO regression is to find a balance between model simplicity and accuracy. It achieves this by adding a penalty term to the traditional linear regression model, which encourages sparse solutions where some coefficients are forced to be exactly zero. This feature makes LASSO particularly useful for feature selection, as it can automatically identify and discard irrelevant or redundant variables. That’s why we might try to use LASSO logistic regression over basic logistic regression.

**What is a classification tree model and why we might try to use it?**

A classification tree model is a type of supervised learning algorithm used in statistics, data mining, and machine learning. It is used to predict the value of a categorical response variable based on one or more predictor variables. The model is built by recursively partitioning the data into smaller and smaller subsets, with each partition being based on a different predictor variable. The result is a tree-like structure where each internal node represents a test on a predictor variable, each branch represents the outcome of the test, and each leaf node represents a predicted value for the response variable. The tree is constructed in such a way that the most important predictor variables are at the top of the tree, and the least important ones are at the bottom. The goal of the algorithm is to create a tree that is as small as possible while still accurately predicting the response variable.

Classification treesare a powerful and flexible machine learning algorithm that can be used for both classification and regression problems. They are particularly useful when the data has non-linear relationships, variable interactions, or outliers, and when interpretability is important. However, classification trees also have limitations, such as being prone to overfitting when they become too complex. To address, ensemble techniques like random forests and gradient boosting are often used to to enhance their performance.

**What is a random forest and why we might use it instead of a basic classification tree?**

A random forest regression model, also known as a random forest, is a versatile machine learning algorithm used for both regression and classification tasks. It is an ensemble technique that uses multiple decision trees to make predictions. In this model, each decision tree is trained on a different subset of the data, and the final output is the average of all the outputs of the individual decision trees. This technique helps to reduce overfitting and improve the accuracy of the model. Random forest can predict continuous values, such as stock prices, temperature, or sales figures, as well as performing classification variables.

Random forests have several advantages over the basic decision trees: reduced overfitting, improved accuracy, less sensitive to outliers, and identifying the most important predictors and improving the interpretability of the model. Therefore, we prefer random forest over basic classification trees.

**A summary of a partial least squares model**

A partial least squares (PLS) model is a supervised learning algorithm used to model the relationship between two sets of variables. The PLS model is used when there are multicollinearity, i.e., the explanatory variables are correlated, and high dimensionality in the data. PLS allows us to reduce the dimensionality of correlated variables and model the underlying, shared, information of those variables (in both dependent and independent variables). Moreover, PLS can model multiple outcome variables, which many statistics and machine learning models cannot directly deal with.

**A summary of regularized Logistic Regression**

A regularized logistic regression model is a variation of logistic regression that includes regularization techniques to prevent overfitting and improve the model's generalization ability. It's used for binary classification problems, where the goal is to predict one of two possible outcomes.

Regularized logistic regression is particularly useful when dealing with high-dimensional data with many predictor variables. It helps to improve model performance, maintain model simplicity, and select the most relevant features. It's commonly used in machine learning and data analysis when dealing with classification problems, especially when overfitting is a concern.

**References:**

<https://en.wikipedia.org/wiki/Lasso_(statistics)>

<https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/>

[Decision tree learning - Wikipedia](https://en.wikipedia.org/wiki/Decision_tree_learning)

<https://en.wikipedia.org/wiki/Partial_least_squares_regression>

<https://towardsdatascience.com/partial-least-squares-f4e6714452a>