MET CS699 – Project Assignment

Final Report

Brandon Bartol – Chuong Nguyen

Introduction

In the realm of public health, identifying risk factors and predicting health outcomes is crucial for effective intervention and support. The Behavioral Risk Factor Surveillance System (BRFSS) Survey Data, a cornerstone of public health research in the United States, provides a wealth of information on various health-related behaviors, chronic conditions, and preventive health practices among adults.

Our project aims to leverage the BRFSS Survey Data, specifically the 2020 dataset, to develop and evaluate classification models for predicting the occurrence of depressive disorders among survey respondents. With depression being a prevalent and debilitating mental health condition, early detection and intervention are essential for improving outcomes and reducing the burden on individuals and healthcare systems.

Dataset Overview

The dataset provided for this project, named "project\_dataset-5K.csv," comprises 5000 tuples and 276 attributes. Each tuple represents an individual surveyed in the BRFSS, with attributes covering a wide range of demographic, behavioral, and health-related factors. The key attribute of interest is the "Class," which indicates whether a person has experienced a depressive disorder ("Y") or not ("N").

Preprocessing

Before diving into model building, preprocessing of the dataset is essential to ensure data quality and model performance. Initial preprocessing steps have already been undertaken, including handling missing data, removing columns with low variance, and addressing multicollinearity. Additionally, feature selection techniques such as Boruta, information gain, and principal component analysis (PCA) have been applied to identify relevant attributes for modeling.

Project Goal

The primary objective of our project is to develop and compare multiple classification models to predict depressive disorder occurrence based on the provided dataset. Through rigorous evaluation and comparison of these models, we aim to identify the most effective approach for predicting depressive disorders. Ultimately, our goal is to contribute to the development of reliable tools for early detection and intervention in mental health disorders.

Team Collaboration

This project represents a collaborative effort between two students, Brandon Bartol and Chuong Nguyen. By pooling our expertise and leveraging each other's strengths, we aim to deliver a comprehensive analysis and evaluation of classification models for depressive disorder prediction. Our partnership enables us to tackle the complexities of data preprocessing, model building, and evaluation more effectively, ensuring a robust and insightful outcome.

**I. Introduction**

* Brief overview of the project objectives
* Description of the dataset and its source
* Explanation of the preprocessing steps performed on the dataset
* Clear statement of the project goal: to build and compare multiple classification models for predicting depressive disorders

**II. Data Preprocessing**

* Initial data inspection and summary statistics
* Handling missing data
* Removing columns with high missing values
* Imputing missing values using mode or median, stratified by the class column
* Removing columns with near-zero variance
* Removing highly correlated columns
* Scaling the data for model compatibility

**III. Feature Selection and Imbalanced Data Treatment**

* Splitting the data into training and testing sets
* Feature selection methods:
  + Boruta
  + Information Gain
  + Principal Component Analysis (PCA)
* Imbalanced data treatment methods:
  + Over-sampling
  + Under-sampling
  + Combined sampling
  + Bootstrap resampling

**IV. Model Building**

* Splitting the training and testing sets
* Building and evaluating multiple classification models:
  + Naive Bayes
  + Logistic Regression
  + Decision Tree
  + Random Forest
  + K-Nearest Neighbors (KNN)
  + Neural Network (NN)
* Summary of parameter tuning and performance evaluation for each model

**V. Model Evaluation & Interpretation**

* Evaluation metrics computation and interpretation
* Comparison of model performance using various balancing methods
* Discussion on the effectiveness of each classification model
* Interpretation of the results and implications for predicting depressive disorders

**VI. Conclusion**

* Summary of key findings and insights from the project
* Reflection on the effectiveness of different classification models
* Suggestions for future research or improvements in methodology
* Closing remarks

**VII. References**

* List of all sources referenced in the paper, including datasets, packages, and methodologies used