Coronary Heart Disease and Myocardial Infarction Risk Assessment Based on Multi-Dimensional Health Factors Using Tabular-based and Text-based Machine Learning Methods

George Matlis, Konstantinos Tsintzas, Efstratios Skaperdas

gmatl@csd.auth.gr, tsintzask@csd.auth.gr, skaperdas@csd.auth.gr, *Department of Computer Science, Aristotle University of Thessaloniki*

1. Overview

This project aims to compare different tabular and text-based classification methods in the task of assessing certain heart disease risks, based on survey data from the CDC which details behavioral risk factors. The dataset used is from the 2021 survey.

This QR code links to a GitHub repository of all code and data used for this project.

2. The Dataset

Behavioral Risk Factor Surveillance System

The Centers for Disease Control and Prevention, a health agency based in the USA, conducts yearly anonymous surveys assessing various behavioral risk factors and health outcomes for people in different areas across the United States. The data collected in this manner is made publicly available. Examples of some features included in the dataset are whether a person smokes, how much they exercise, and whether they have experienced certain health issues recently.

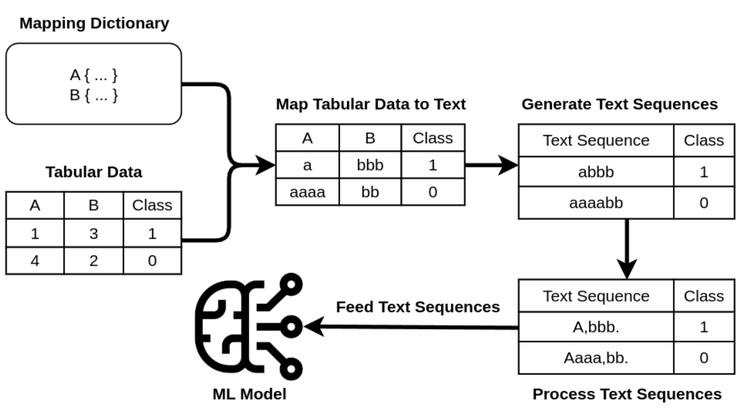
Two of the health issues asked about are incidence of coronary heart disease and/or myocardial infarction. Coronary heart disease is a condition that occurs when the blood vessels supplying the heart muscle become narrowed or blocked due to the buildup of plaque. A myocardial infarction, commonly known as a heart attack, occurs when there is a sudden blockage of blood flow to a part of the heart muscle. Heart attacks require immediate medical intervention to avoid permanent damage or death. For this project, we utilized survey data collected in 2021 for Selected Metropolitan/Micropolitan Area Risk Trends (SMART). The surveys also include geographical and demographic information.

The target variable we used is called _MICHD in the original dataset, and expresses whether respondents have ever reported having coronary heart disease or myocardial infarction.

3. Data Preprocessing

The dataset was initially in a proprietary format, which we converted to Excel tabular data.

For the next step, we needed to additionally construct a text dataset corresponding to the tabular data.



To accomplish this, we utilized the tabular data in combination with a manually created dictionary that contained mappings of values to short text descriptions for each feature. As this was survey data, most features had a limited amount of possible values.

Next, we applied the mappings to the dataset, and merged each row into comma-separated strings. The resulting documents were used to train a variety of text-based machine learning models on the classification task.

4. Methodology

The classification process consists of evaluating the following methods and models:

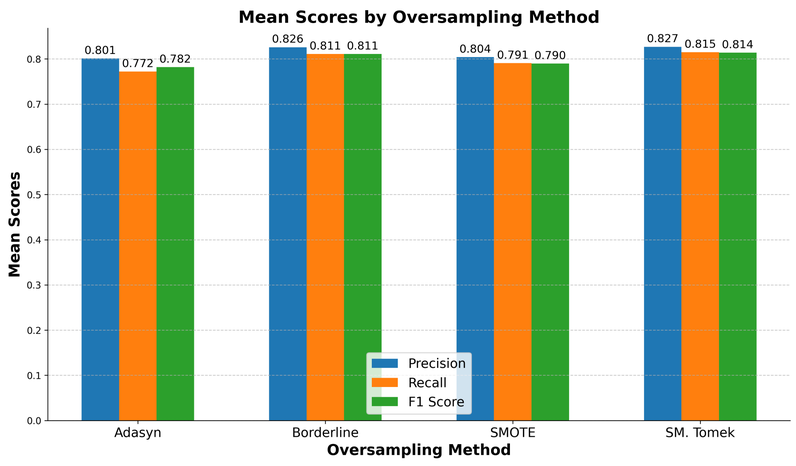
- XGBoost and Random Forests (RF)
- Decision Tree
- Gaussian NB, Multinomial NB, Bernoulli NB, and Binary NB
- KNN (K-Nearest Neighbors)
- Logistic Regression
- BERT and GPT-2 for sequence classification

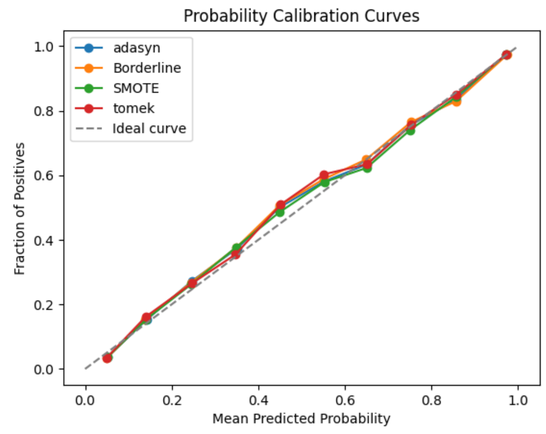
The classes in the original dataset are imbalanced. To address this issue, we used oversampling by generating synthetic data and undersampling. The following oversampling methods were considered:

- Adasyn
- Borderline
- SMOTE
- SMOTE Tomek

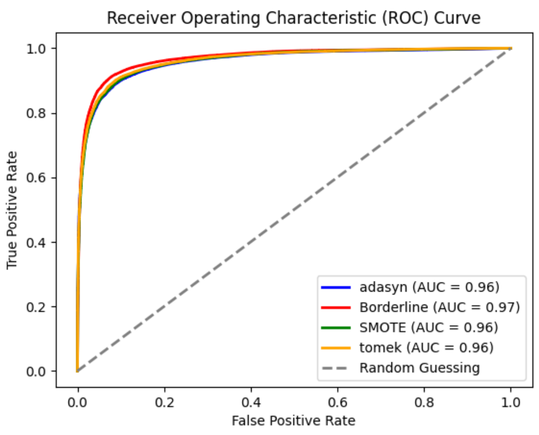
For undersampling, we randomly removed elements from the majority class to ensure both classes would have an approximately equal amount of examples.

5. Classification with Oversampling

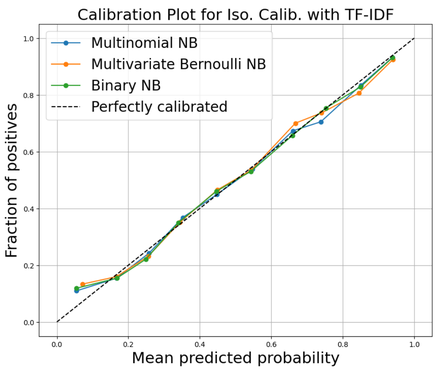


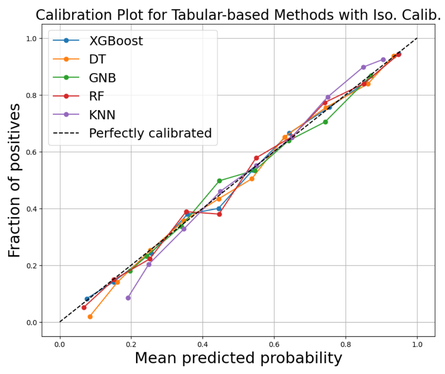


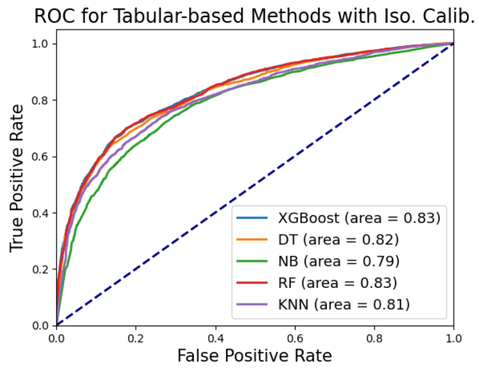
- Random Forest executed the task exceptionally well (BDL P. 0.915 R. 0.912 F1 0.912 C. 2927)
- We obtain an exceptionally high-quality calibration plot for the SMOTE method.
- We obtain a nearly perfect ROC curve for all oversampling methods.
- Gaussian NB performed quite poorly (SMOTE P. 0.741 R. 0.722 F1 0.718 C. 7998).
- The tabular-based methods outperformed the text-based methods (P. 0.7577 R. 0.7305 F1 0.7305 C. 7362.25).
- The variables _AGE_G, _INCOMG1, _SEX, _EDUCAG, _RFHLTH had the greatest impact on the models.



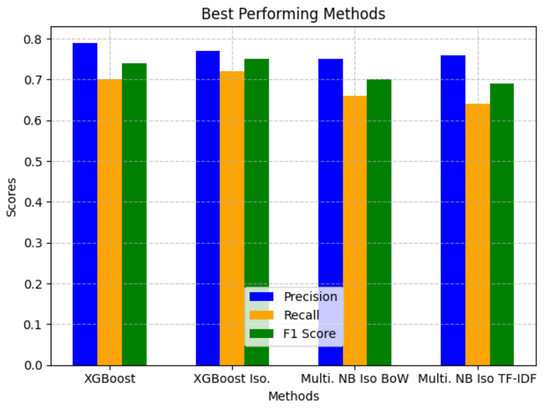
6. Classification with Undersampling



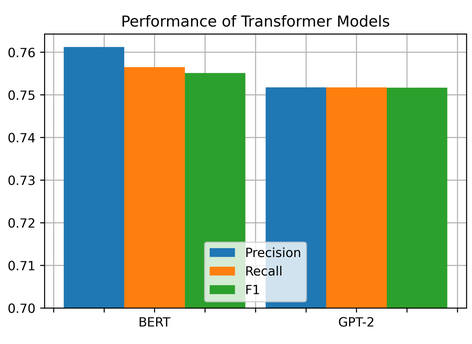


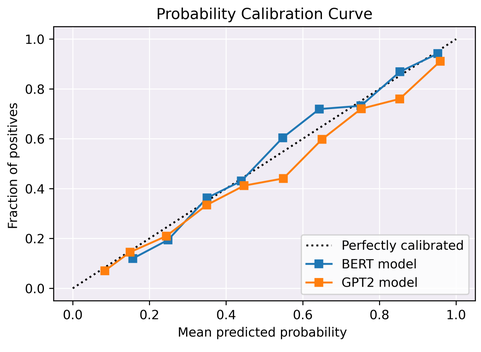


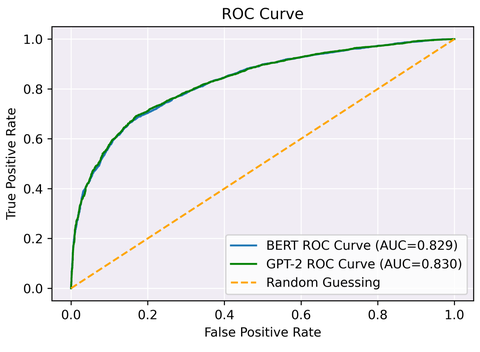
- The text-based methods exhibit a nearly perfect calibration curve, whereas the tabular methods fall shorter in comparison.
- Tabular-based methods often outperform text-based methods due to the meticulous fine-tuning process involved in selecting optimal parameter values for the former. Unlike tabular-based methods, most text-based methods, except for models like BERT and GPT, typically do not undergo fine-tuning.
- XGBoost with isotonic calibration outperforms all other tested methods.
- The tabular-based methods with isotonic calibration achieve the highest AUC scores.



7. Text-Based Transformer Models







- Pre-trained models were used, which were fine-tuned to the undersampled CDC survey dataset for 4 epochs.
- BERT achieved the highest recall and F1 score among the tested text-based models.
- Both models exhibited very good probability calibration, and near-identical ROC curves with high AUC scores.

8. Conclusions

- For the given classification task, Random Forests were the best overall method of those tested, achieving the highest precision, recall, and F1 scores.
- Text-based methods have better probability calibration naturally, while tabular methods need to be adjusted for it.
- When provided with tabular data, tabular methods are expectedly better-performing.
- From the tested methods, BERT exhibits the highest potential for better performance due to its capacity for fine-tuning on additional survey data.