Data Science Homework 4

Andrés Ponce, **彭思安** P76107116

May 7, 2022

1 Introduction

Our data's properties will determine the pre-processing steps we take to ensure our model performs well with new data. For instance, since machine learning methods can for the most part deal only with numerical values, its is necessary to convert string-based values into numbers. Such methods range from just assigning a numerical value to a class index to using hash tables to store indices of the column to which it belongs.

Another common issue in the data processing stage is imbalanced data. This happens when some categories in our data happen less often than others, sometimes much less often. These categories are known as **minority class**, and more common classes or categories are known as **majority class**. To ensure our model is exposed to the category, we can either remove elements from the majority class with **undersampling methods** or create new samples from the minority class using **oversampling methods**.

In this assignment we investigate the relation between datasets and different encoding and sampling methods. First, we describe the datasets used, then conduct experiments on different combinations of encoding and sampling methods.

2 Datasets

In this assignment, we used 15 classification datasets, most of which have a class imbalance. The datasets were mostly obtained mostly from Kaggle, the UCI repository, and OpenML.

3 First Question

Dataset	Features	Categorical	Numerical	Size	Task
Heart Disease	14	11	3	319,795	Multi-class
BankMarketing	20	9	11	32561	Binary
Cover Type	55	0	55	581012	Multi-Class
Income Evalutaion	15	9	6	32,562	Binary
Telco Customers	21	19	2	7044	Binary
Wine Quality	12	12	0	1144	Multi-Class
Wisconsin Breast Cancer	20	0	20	570	Binary
Abalone	8	1	8	4178	Multi-Class
Arcene	100	0	100	100	Multi-Class

Table 1: Properties of the datasets used.