1. Introduction

The issue of imbalanced data can impact the effectiveness of machine learning models, particularly for classification methods that assume a balanced data sample size across each class. Imbalance data occurs in dataset that have classes with more majority classes than minority classes. Data imbalance can introduce bias in the classification algorithm, leading to lower accuracy when it comes to classifying the minority class. The unequal distribution of data between classes can cause the algorithm to favor the majority class, as it has more samples for training. As a result, the minority class may receive less attention and may be misclassified more frequently [1]. This is a challenge for handling problems in the classification algorithm, especially data on companies that are experiencing financial distress. In the real world, a problem may not often occur, such as the data problem for a company experiencing financial distress, therefore causing an error in classification which results in a high cost to overcome it [2]. So, this imbalance needs to be addressed, to ensure fairness and accurate classification, potentially through techniques like oversampling or under sampling to balance the dataset and improve performance for the minority class especially in the case of financial distress classification.

Financial distress refers to a situation in which a company is unable to meet its obligations to creditors. It's also described as a state when a company faces financial challenges. This circumstance arises prior to a company being officially declared bankrupt [3]. Financial distress classification has been conducted by research [1]. In that study, classification was performed using tree-based models, specifically the Decision Tree. The dataset used consisted of data from a selected group of business units accounting in the double-entry bookkeeping system for 3 periods, namely in the years 2016, 2018, and 2019, located in the Republic of Slovakia. The dataset comprised 599 companies, with 27 companies classified as experiencing financial distress and 532 companies classified as healthy. One of the findings from the research scenario was the accuracy of 99.47% for the healthy company class and 29.41% for the financial distress company class.

In research [2], prediction on companies experiencing financial distress with imbalance data was performed using SMOTE. The dataset consisted of 2628 samples, with a ratio of 2190 normal companies and 438 companies experiencing financial distress. The sample data was collected from companies listed on the Shanghai Stock Exchange and Shenzhen Stock Exchange in China. SMOTE was applied to balance the data between companies experiencing financial distress and normal companies, addressing the data imbalance issue. The results of the research demonstrated that the data balancing process significantly improved the performance of the model for companies experiencing financial distress.

Several studies have been conducted on the classification of financial distress using SVM and Naive Bayes algorithms. The study [4] utilized the SVM algorithm without hyperparameter tuning for the classification of financial distress. The results of the research showed an accuracy of 81.06%, an error rate of 18.94%, a precision of 89.09%, and a recall of 59.04%. In study [5], the prediction of financial distress was performed using SVM with hyperparameter tuning. The research yielded an accuracy of 92%, a sensitivity of 93%, a Matthews Correlation Coefficient (MCC) of 85%, and a precision of 90%. In a similar field of research, study [6] conducted bankruptcy prediction using the Naïve Bayes algorithm. The best results obtained from the research showed an accuracy of 92.47%, an error rate of 7.52%, and a model building time of 0.13 seconds. Therefore, researchers utilize SVM and Naive Bayes algorithms as methods for classifying companies experiencing financial distress.

Based on the explanation above, a crucial extension of the data imbalance issue is the adverse impact that affects the performance of the classification model, particularly in cases of financial distress companies. When accuracy decreases in the minority group, the risk of the potential

financial crisis going undetected within the company also increases. The Kmeans-SMOTE approach has not been widely used in real-world problems, and it is necessary to find optimal hyperparameters. So, this research proposes to implement the Kmeans-SMOTE approach to the imbalance data of financial distress companies to see how this approach influences classification using SVM and Naïve Bayes. Although the extensive research conducted on SMOTE, Kmeans-SMOTE has not been widely adopted and observed. Therefore, Kmeans-SMOTE is used to solve imbalance problem. So that after the imbalance data has been solved, data is obtained in good condition for the classification process [7]. Hence, Kmeans-SMOTE will be used to balance the data classes, especially in the financial distress problem and emerging algorithms that will be used to complete the research objectives.

The next section will discuss the research methods employed in this study, encompassing data preprocessing, handling imbalanced data, model building, and model evaluation (Section 2). Subsequently, Section 3 will present a discussion of the obtained results. Finally, Section 4 will delve into the conclusions drawn from this research.