1. Introduction

Organelle biogenesis in eukaryotic cells requires the selective assembly of specific proteins, lipids, and other constituents from a shared pool of biosynthetic intermediates. An extensively present eukaryotic microbe is yeast. Yeast vacuole biogenesis was chosen as a model system. It has garnered much interest due to its rapid development and remarkable metabolic efficiency. Therefore, it is possible to generate extensive collections of mutants with defects in unbalanced vacuole assembly. With this in mind, we must find the structural balance of data in yeast[1].

Imbalanced data is when there is an unbalanced distribution of data classes and the number of data classes is either more or lower than the number of other data classes[2]. In this research, unbalanced data is a situation with significantly more observations in one class than in the other. This problem is predominant in cases where anomaly detection is critical, for example, fraud detection in banks, healthcare, insurance, etc.

This paper uses SVM classification to compare the results of different sampling methods, namely Destiny-Based Majority Sampling and Near Sampling Techniques. We applied undersampling techniques to correct class imbalances in yeast data sets. Previous studies rarely compare the performance of undersampling techniques, particularly in using undersampling DBMUTE and NearMiss on yeast data sets. We use the two performances to share the results of imbalanced data on the data sets. Furthermore, to find out the extent to which the use of f1 score performance and balanced accrued good value against the imbalanced data sets.

In undersampling, several algorithms can be used. In this research, we use DBMUTE and NearMiss undersampling. The DBMUTE undersampling algorithm eliminates majority noise cases that overlap with minority cases. And the NearMiss algorithm reduces information loss during undersampling in the majority class[3]. In this study, we also used a classification method, namely Support Vector Machine. This classification method was chosen because it deals with imbalanced data well. The categorization operates by creating an N- dimensional hyperplane that divides the data into two groups in the best way possible. This research uses references from previous studies relevant to imbalanced data, classification methods, undersampling methods, preprocessing, and evaluation stages. Undersampling is often used to balance data by reducing the size of abundant classes.