NORTH AMERICAN UNIVERSITY
INSPIRATION INNOVATION GLOBAL COMPETENCE

COMP 5353
Data Mining
SPRING - 2021
Assignment #3

Full Name: ________Emine Balpetek______________________

TASK: **READ ANY SCIENTIFIC ARTICLE WITH TITLE OF "DATA MINING FEATURE SELECTION".**

TITLE:

Software Fault Prediction with Data Mining Techniques by Using Feature Selection Based Models

AUTHOR(S):

Amit Kumar Jakhar and Kumar Rajnish

AIM OF THE RESEARCH:

The purpose of the research paper is depth study and comparison of the Data Mining Techniques by Using Feature Selection Based Models.

Feature selection, as a data preprocessing strategy, has been proven to be effective and efficient in preparing data (especially high-dimensional data) for various data-mining and machine-learning problems. The objectives of feature selection include building simpler and more comprehensible models, improving data-mining performance, and preparing clean, understandable data. The recent proliferation of big data has presented some substantial challenges and opportunities to feature selection. Software engineering activities comprise of several activities to ensure that the quality product will be achieved at the end. Some of these activities are software testing, inspection, formal verification and software defect prediction. Many researchers have been developed several models for defect prediction. These models are based on machine learning techniques and statistical analysis techniques. The main objective of these models are to identify the defects before the delivery of the software to the end user. This prediction helps project managers to effectively utilize the resources for better quality assurance. Sometimes, a single defect can cause the entire system failure and most of the time they drop the quality of the software system drastically. Early identification of defects can also help to make a better process plan which can handle the defects effectively and increase the customer satisfaction level. But the accurate prediction of defects in software is not an easy task because this is an indirect measure. Therefore, it is important to find suitable and significant measures which are most relevant for finding the defects in the software system. This paper presents a feature selection based model to predict the defects in a given software module.

The most relevant features are extracted from all features with the help of seven feature selection techniques and eight classifiers are used to classify the modules.

A. Feature Selection Approaches:

All the attributes which are given in the data set are known as “Features” or “characteristics of the data set. Feature selection is an important technique for extracting the most essential feature of the data set which has hundreds or thousands of features. This problem has occurred in several machine learning tasks such as: prediction, regression, and classification etc. The main objective of feature selection approaches is to find the most appropriate feature or a subset of features from a given data set so that these selected features will improve the effectiveness of a model.

B. Classification Techniques

1. Decision Tree (DT):

Decision Tree is an effective technique in many data mining domains for classification of data. . Every node in the tree represents a feature and the branch represents the value. Classification starts at the root node and moves to the leaf node for prediction of the class that a particular instance belongs to.

2. Neural Networks (NN):

Neural networks consists of numerous interconnected neurons (processing elements). Basically the NN is made up of three things: input layer (which takes the input), hidden layer (perform complex functions), and the output layer (where the output is produced).

3. Support Vector Machine (SVM):

SVM was developed by Vapnik [2] in 1995 for addressing the problem of pattern recognition. Initially, it was developed only for binary classification but later it extended to solve the multiclass classification problems .

4. Naïve Bays Classifier :

This is the simplest classifier . This classifier is used for both the feature which is independent to every class and also for those features where independence is no further valid. This classification technique works in two stages: training stage and prediction stage.

5. K-star:

K-star [12] classifier is an instance-based classifier. In this classifier, the class of the test instances or cases is relying upon the class of training instances of those are similar to it, which is determined with the help of some similarity function.

TECHNIQUES, ALGORITHMS AND METHODOLOGIES USED IN THE ARTICLE:

In this work Subjective and Quantitative Analysis. Observations, Documents , Archives and records used as data collection methods.

The authors determined to use many analysis ways like Prescriptive Analysis and Descriptive Analysis. But main analysis is Text analysis used in the article.

CONCLUSION:

In this study, the authors proposed a feature selection model for software defect prediction. Public NASA data sets CM1, MC1, MC2, PC3, and PC4 from PROMISE data repository, are used. The different classification models are used for defect prediction, and two test groups are performed on each data set. In the first test group, all the features of each data sets are used with the classifier models for fault prediction, but in the second test group, several feature selection techniques are used for assigning weight to every feature of the data set, then only high weighted features are used for software defect prediction. For this work, only four features to twenty three features are used for the second test group experiment. The result of both the test group was calculated with the help of a confusion matrix which has been generated by each classifier and both the test groups were analyzed on the basis of the several performance parameters. The measured results of each table indicated the best value of the actual data set with all the classifiers which are concerned in this work. After analyzing the result of several performance parameters of both the test groups, it is found that, the proposed model performed better in several phases like: accuracy, PD, TNR, FNR, f-measure and AUC and it can be concluded that the proposed test group model have better ability for software fault prediction.