Logo, company name

Description automatically generated

Data Warehouse Data Mining

Semester Project

Instructor:Mam. Yella Mehroze

|  |  |
| --- | --- |
| Student Details | |
| Name | Bilal Ahmad |
| Course | Data mining |
| Reg No. | FA19-BCS-136 |
| Section | C |
| Date | 18/12/2021 |

**Student performance prediction**

**Dataset details**

**Description:**

Student performance prediction database will be use in this project to predict the engagement level of student with learning resource. Predicted class label will show high or low based on the interaction of student with study material. Dataset contains 13 attributes including class attribute and almost 500 records.

It includes:

* **Login:** How many times a student login to portal.
* **Content Read:** How many times a student read the study material content.
* **Forum Reads:** How many times a student read the problem/issue on community forum.
* **Forum post:** How many times a student posts the problem/issue on community forum.
* **Review:** How many times a student reviews the quiz before submission.
* **Lateness indicator:** Is student submit the late assignment. There are multiple lateness indicator attributes.
* **Average assignment submission time:** How many hours student takes to submit the assignment. There are multiple lateness indicator attributes.
* **Engagement level:** This is a class label that will show the level of student engagement to study materials based on previous attributes.

**Source:**

Source of the dataset is from github (open source platform) and you can get this dataset from [here](https://github.com/Western-OC2-Lab/Student-Performance-and-Engagement-Prediction-eLearning-datasets/blob/main/Student%20Engagement%20Level%20Prediction%20-%20Binary%20Case/Student%20Engagement%20Level-Binary.csv).

**Problem type:**

I want to predict class labels, so I will use classification because classification is used to predict the nominal value of class attribute.

**Techniques:**

I will use weka tool to classify to train model and predict the class labels. I will also use feature selection to select the relevant attributes for model training.

**Process**

**Preprocessing:**

Dataset is already preprocessed. There is no negative, null, or empty value. Class attribute label is nominal and contains H (high) or L (low) labels.

**Classification:**

I will use classification because classification is used to predict the nominal value of class attribute. I will train the model using different algorithm and test options.

Here are some steps to perform classification:

* Load training data into weka.
* In classify tab, choose tree and select J48 algorithm.
* Select cross validation (10 folds) in test option.
* Select class attribute
* Start the process.

A picture containing table

Description automatically generated

Result is showing that dataset is 99% correctly classified. There are other details like total instances, root mean, absolute error, confusion matrix, accuracy details etc.

Text

Description automatically generated with medium confidence

Right click on the result item and select “save the model” for test data label prediction.

Now, create the copy of training data, shuffle the order, and remove all the labels of class attribute to check either trained model can predict class labels or not.

Here are some steps to perform prediction:

* In classify tab, choose **supplied test data** in test options.
* Load the test data file (without labels).
* Choose class attribute.
* Click on more options, choose **Plain text** in output prediction.
* Reevaluate the model on current dataset.
* Result will show the predicted class labels.

Result is showing the predicted class labels as high or low based on trained model.

Table

Description automatically generated

**Association rules:**

Association rules are not applicable on this type of dataset and this

feature is out of scope. Association rules are only applicable if dataset contain itemset.

**Feature selection:**

The attribute selection task essentially consists in selecting a subset of originally available attributes to be subsequently used for model creation. For this purpose, I will use selectAttribute tab to select top 10 attribute in dataset that will be use in training model.

Here are some steps to perform feature selection:

* Load training data into weka.
* In classify tab, choose info gain in attribute evaluator.
* Set number option to 10 in Ranker.
* Select cross validation (10 folds) in test option.
* Select class attribute
* Start the process.

Table

Description automatically generated

Result is showing relevant attributes which are necessary for model training. Attributes are short on the base of average rank. Select top rank attributes for training model. Class attribute will be mandatory.