Problem statement

The advancement of technology and the Internet has significantly affected learning and education. Within that context, e-learning was developed and can be defined as ‘‘the use of computer network technology, primarily over an intranet or through the Internet, to deliver information and instruction to individuals’’ [1,2]. However, there are various challenges regarding e-learning, such as the assorted styles of learning, and challenges arising from cultural differences [3]. Other challenges include pedagogical e-learning, technological and technical training, and the management of time [4]. That is why the need for more personalized learning has emerged. Personalized learning can be considered as one of the biggest challenges of this century [5], where the personalization of elearning includes adaptation of courses to different individuals. One of the biggest learning differences includes the level of knowledge an individual has, and it is being assessed through the learner profile. Learner profile is the most crucial step of the personalization process [6,5]. To make learning more personalized, adaptive techniques can also be implemented [7,8]. Data can be automatically collected from the e-learning environment [8] and then the learner’s profile can be analysed. Accordingly, this report uses the comparative analysis gained from various classification algorithms to predict student’s performance at earlier stages of the course delivery.

Related Work

DM methods have great potential when it comes to analysing educational data. There is a big interest for understanding the needs of students and their actual level of knowledge. Many researchers have been interested in this problem during the last few years. In 2000, researchers tried to determine lowperforming students by using association rules [12], so that they could involve them in additional courses. Luan [13,24] tried investigating which students are most likely to fail the course by using clustering, neural networks and decision tree methods. In 2003 [25], Minaeli-Bidgoli et al. used classification for modelling online student grades, while in [26] authors were investigating how students’ performance can be influenced by demographic characteristics and performance.

Pardos et al. [27] used LR to predict the test score in math based on students’ individual characteristics, while Superby et al. [28] used decision tree techniques, RF method, Neural networks, and Linear discriminant analysis for predicting students who will most likely drop-out. Vandamme et al. [29] also used Decision tree methods, neural networks and linear discriminant analysis for their prediction of students who will fail the course by classifying them into three groups: low, intermediate, and high-risk students. In 2008, Cortez and Silva [30] compared DM algorithms from four different approaches, namely Decision Tree, RF, Neural Network and SVM for prediction of students’ failure. Kovacic [31] developed a profile of students who would most likely fail or succeed by using classification techniques. He used socio-demographic and learning characteristics as variables for predicting students’ success. Ramaswami et al. [32] tried developing a predictive model that will be used for identifying students who are slow at learning by using Chi-square Automatic Interaction Detector (CHAID) decision tree algorithm.

Pandey [33] used NB classification to accurately distinguish the bright students from the slow ones. Their model was able to predict students’ grades based on their previous grades. In 2012, authors conducted comparative research to make a best guess of the student’s performance [34]. The study used decision tree algorithms and it was aimed at finding the best decision tree algorithm that can accurately predict students’ grades. The authors found that CART algorithm that was designed as a decision tree algorithm was the most efficient as it produced the most desired results and concluded that it is desirable to try different classifiers first and then decide which one to use based on the precision and accuracy it gives. Kabakchieva in [15] used four DM algorithms — OneR Rule Learner, Decision Tree, Neural Network and k-NN. Results indicated that the highest accuracy was achieved using the Neural Network algorithm, where the most influencing factors on the classification process were students’ score upon admission and the frequency of failures in the first-year examinations.

Model Architecture

K-nearest neighbours model was used to train the model for solving the problem mentioned above. Before implementing the model, several pre-processing steps was done on the data as follows:

* Data visualisations to take a glimpse of the features’ distributions and to compare between the percentage of the data labelled as ‘G’ and the data labelled as ‘W’.
* Principal component analysis to take a glimpse of the important features in the data set and to reduce the data to two columns for visualisations.

After that, a grid search was conducted using 3-fold cross validations and K-NN model for finding the best parameter k for training the model.

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

**How does K-NN work?**

The K-NN working can be explained on the basis of the below algorithm:

* Step-1: Select the number K of the neighbours
* Step-2: Calculate the Euclidean distance of K number of neighbours
* Step-3: Take the K nearest neighbours as per the calculated Euclidean distance.
* Step-4: Among these k neighbours, count the number of the data points in each category.
* Step-5: Assign the new data points to that category for which the number of the neighbour is maximum.
* Step-6: Our model is ready.

Evaluation strategies

Because the data is not balanced, the following evaluation strategies were used:

* Recall

Text

Description automatically generated

* Precision

Text

Description automatically generated

* F1 score:

Text

Description automatically generated with low confidence

* ROC-AUC.

Evaluation results

At the 20% stage:

After applying grid search to the model, the model was trained with the best k = 4. The training precision was 99 percent, while the testing precision was 93.75%.

At 50% stage:

After applying grid search to the model, the model was trained with the best k = 4. The training precision was 99 percent, while the testing precision was 93.75%.