Intelligent Admission:The Future Of University Decision Making With Machine Learning

Abstract:

Taking appropriate decisions in the academic processes at a university has a great impact on improving the quality of education and can have an important benefit for students, faculty members, and the entire academic community. In this paper, we propose a decision support solution providing accurate analysis, better decision support, and reporting and planning capability to assist decision-makers in order to enhance the quality of educational processes. To achieve this goal, a set of machine learning is used. Experiments are conducted on real data describing the College of Computer Science and Engineering (CCSE) at Taibah University in Saudi Arabia. Results show that we can predict graduation rates in a real case study to support decision-making. In addition, a comparison between four techniques of machine learning namely Support Vector Machine, Naïve Bayes, Decision Tree, and Random Forest is held using accuracy, recall, precision, and F-measure.

Keywords—component, academic process, decision-making, machine learning, decision support, Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest.

Introduction

Improving educational quality is among the top priority for decision-makers in many universities. There is a huge amount of data collected each year at the university level; however, these data are not properly exploited to extract knowledge and insights useful to improve the quality of education within the university, therefore an automated process is highly- demanded to achieve this goal is in high-demand. In literature, several systems are proposed to support the academic process. Yet, most of them focus on the descriptive side and disregard the predictive one, which aims to analyze current and historical data to make future decisions or otherwise unknown events. Among methods, Machine Learning (ML) has shown outstanding performances in many fields by providing useful descriptive and predictive information. The main advantage of ML methods is their ability to create models that may be integrated into the decision-making process as in [1] [2].

In this paper, we start by briefly presenting our previous work by developing a business intelligence (BI) solution to support academic affairs at Taibah University as in [3]. Later, we illustrate the proposed data warehouse in order to store the data collected from the academic process. Then, we describe four ML methods named Support Vector Machine, Naïve Bayes, Decision Tree, and Random Forest which included in the BI solution. The proposed approach will involve several stockholders such as students, faculty members, heads of departments, deans of colleges, and university presidents. Our approach will be evaluated in real data from five programs in the College of Computer Science and Engineering (CCSE) for a period of 8 years' worth of information.

BACKGROUND

At the beginning of each semester, the newbie students are assigned to academic advisors who are responsible for providing a full copy of the Academic Advising (AA) guide to the students to discuss and ensure understanding all the rights and duties. Additionally, the academic advisor gives pieces of advice to their

advisees in selecting or registering courses at the beginning of each semester. According to the policies followed in most of the Saudi universities, students can register their chosen courses one-week earlier.

Because of inexperience or even randomly distributing students over the sections, students might want to change the assigned sections for any reason. The advising process, at Taibah University (TU), e.g., stars by raising a request from the student to change a section or to add some courses which are non-regular sequence. The academic advisors explain to the students how such a change affects their future progress, as follow:

- The academic advisor asks advisees to bring updated documents of their curriculum and registered courses during the early registration period.

- The academic advisor reviews the students' progress and recommends courses based upon their study plan.

Problems and Issues in Supervised learning:

Before we get started, we must know about how to pick a good machine learning algorithm for the given dataset. To intelligently pick an algorithm to use for a supervised learning task, we must consider the following factors [4]:

Heterogeneity of Data:

Many algorithms like neural networks and support vector machines like their feature vectors to be homogeneous numeric normalized. The algorithms that employ distance metrics are very sensitive to this, and hence if the data is heterogeneous, these methods should be the afterthought. Decision Trees can handle heterogeneous data very easily.

Redundancy of Data:

If the data contains redundant information, i.e. contain highly correlated values, then it's useless to use distance based methods because of numerical instability. In this case, some sort of Regularization can be employed to the data to prevent this situation.

Dependent Features:

If there is some dependence between the feature vectors, then algorithms that monitor complex interactions like Neural Networks and Decision Trees fare better than other algorithms.

Bias-Variance Tradeoff:

A learning algorithm is biased for a particular input x if, when trained on each of these data sets, it is systematically incorrect when predicting the correct output for x, whereas a learning algorithm has high variance for a particular input x if it predicts different output values when trained on different training sets. The prediction error of a learned classifier can be related to the sum of bias and variance of the learning algorithm, and neither can be high as they will make the prediction error to be high. A key feature of machine learning algorithms is that they are able to tune the balance between bias and variance automatically, or by manual tuning using bias parameters, and using such algorithms will resolve this situation.

Curse of Dimensionality:

If the problem has an input space that has a large number of dimensions, and the problem only depends on a subspace of the input space with small dimensions, the machine learning algorithm can be confused by the huge number of dimensions and hence the variance of the algorithm can be high. In practice, if the data scientist can manually remove irrelevant features from the input data, this is likely to improve the accuracy of the learned function. In addition, there are many algorithms for feature selection that seek to identify the relevant features and

discard the irrelevant ones, for instance **Principle Component Analysis** for unsupervised learning. This reduces the dimensionality.

Overfitting:

The programmer should know that there is a possibility that the output values may constitute of an inherent noise which is the result of human or sensor errors. In thi case, the algorithm must not attempt to infer the function that exactly matches all the data. Being too careful in fitting the data can cause overfitting, after which the model will answer perfectly for all training examples but will have a very high error for unseen samples. A practical way of preventing this is stopping the learning process prematurely, as well as applying filters to the data in the pre-learning phase to remove noises.

Only after considering all these factors can we pick a supervised learning algorithm that works for the dataset we are working on. For example, if we were working with a dataset consisting of heterogeneous data, then decision trees would fare better than other algorithms. If the input space of the dataset we were working on had 1000 dimensions, then it's better to first perform PCA on the data before using a supervised learning algorithm on it.

PROPOSED APPROACH

The purpose of this paper is to provide decision support for AA users in order to assist them with useful knowledge that helps improve academic operations. The main idea is to build a tool able to provide accurate analysis, reporting and planning, better decision support, and improved data quality. This will result in enhancing the quality of the educational process. In previous work, we proposed a business intelligence-based solution to support academic affairs at Taibah University [3]. The proposed process is based on three steps: collecting data from different sources, developing a multi-dimensional solution describing the academic process, and visualizing information through a dashboard. Experiments are made using SQL Server Data Tools provided by Microsoft to provide statistical and predictive decisions required in the academic process. Fig. 1 depicts the proposed data warehouse.

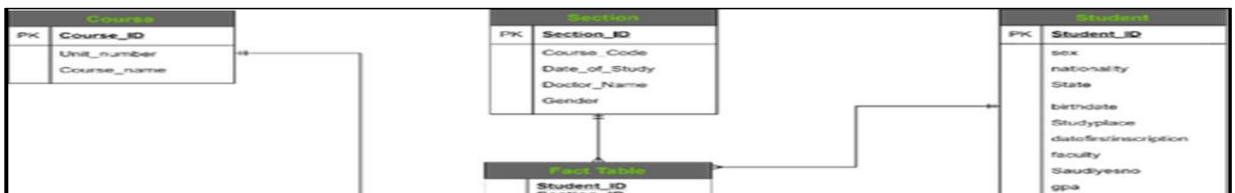

Fig. 1. The proposed data warehouse.

The proposed data warehouse is composed of five dimension tables named student, mark, course, dropped course, and section; and a fact table grouping the following measures count, average, sum, min, and max. In this paper, we propose to reinforce the proposed BI tool by making output decisions more profitable. This will make decision-making in the AA process more automated with minimal risk using descriptive/predictive analytics. We start by providing a brief background of five ML algorithms used to provide descriptive/predictive analytics in our proposed tool namely Support Vector Machine, Naïve Bayes, Decision Tree, and Random Forest.

A. Support vector machine

Support vector machine (SVM) is a supervised machine- learning algorithm that was first proposed by Cortes and Vapnik [12]. The main idea of this algorithm is to achieve a high-prediction accuracy by plotting data as points in n- dimensional space [13]. Then, it creates a model by finding the optimal hyperplane from the training set. In general, SVM has significant good results when it is applied to data outside the training set compared to other existing methods [13].

- *Naïve Bayes*

Naïve Bayes (NB) is defined by Vijayarani and Muthulakshmi [14] as a "probabilistic classifier that evaluates a set of probabilities by calculating the frequency and arrangements of a value in a given dataset". The NB is based on two main steps: (1) training to estimate parameters for the probability distribution and (2) prediction to compute the posterior probability of that sample belonging to each class. NB has been used in many classification tasks due to its ability to provide competitive classification accuracy and computational efficiency and many other desirable features [15].

B. Decision tree

Decision Tree (DT) algorithm is a popular ML technique used in classification, regression, and prediction [16]

[17]. A DT is constructed through an algorithm that splits the dataset based on several input variables. DT is a non-parametric supervised learning method that aims to predict the value of a variable by learning decision rules inferred from the data features. The tree is composed of non-leaf nodes, arcs, and leaf-nodes. A DT is composed of non-leaf nodes, leaf-nodes, and arcs. Non-leaf nodes constitute internal nodes labeled with input features. Each leaf represents a value of the target variable, which is labeled with a class or a probability over the classes. Arcs link nodes (non-leaf or leaf) and label possible values of features. Decision rules correspond to the path from the root to the leaf and they are written as if-then-else statements [4].

C. *Random forest*

Random forest (RF) consists of an ensemble of individual decision trees, which constitutes an ensemble learning method for classification, regression, and prediction. The same process in DT is carried out for each tree in the RF and a voting process is performed to choose the last decision among all trees. The main advantage of using RF is bagging, which refers to reducing the deviation in results by combining multiple decision outputs obtained by different DTs [18].

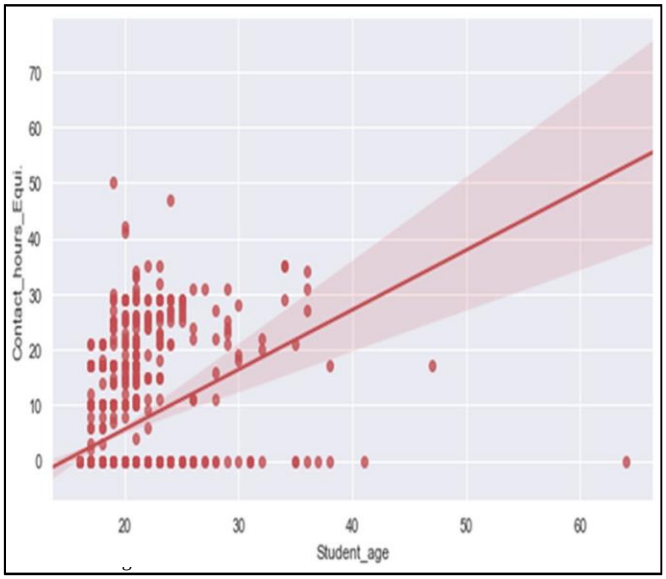

Fig. 3. Correlation between student age and amount of equivalent hours.

Fig. 3 shows the correlation between "student age" and number of "equivalent hours". Even if this graph does not tell us more information. However, zooming on the age period between [23, 30], we noted that most of the equivalent hours fall in this interval. This can be explained by two reasons: (1) The first reason is the educational regulations followed at Saudi universities that allow students to transfer from other branches or universities. So, when the students find themselves near to fail or hear that the study in other universities is easier than the study in their current university, they -especially fresh and junior students- might want to transfer. (2) the second finding is related to the Deanship of Admission and Registration that tries to count such hours as soon as they can. However, in some cases are delayed and the students start their study even if the equivalence process does not finish yet.

In addition to the earlier findings, the head of departments, as well as the deans of colleges, might want to give a general overview of the performance of students. Hence, our tool provides them with such a requirement. Fig. 4 presents how is the distribution of students' grades over the years. The figure shows that a notable number of students gained a "very good" GPA.

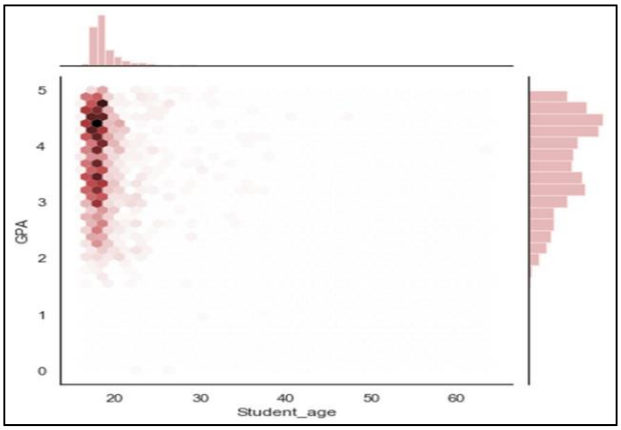

Fig. 4. Distribution of Students' GPA over the years.

Fig. 5 despites the distribution of students' grades with respect to their gender shows that the female section performs well compared to the male section and also the number of females who obtain high marks is greater than the number of males.

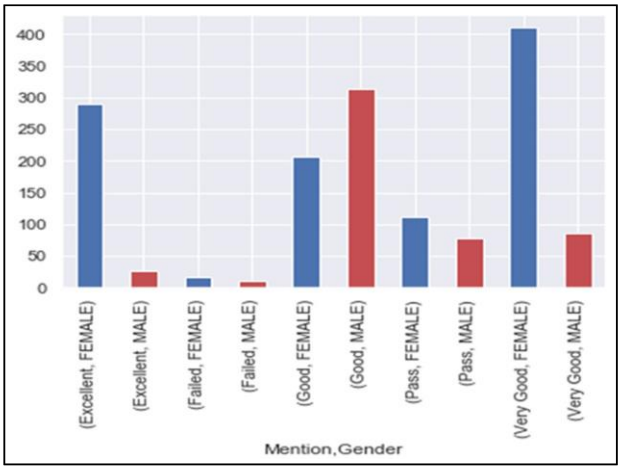

Fig. 5. Performance of Girls verse of Boys.

The decision-makers might want to see the performance of foreign students at the university. Fig. 6 shows that even if the low number of non-Saudi students inscribed in the university, the majority of students is obtained very good marks and they perform well during their studies.

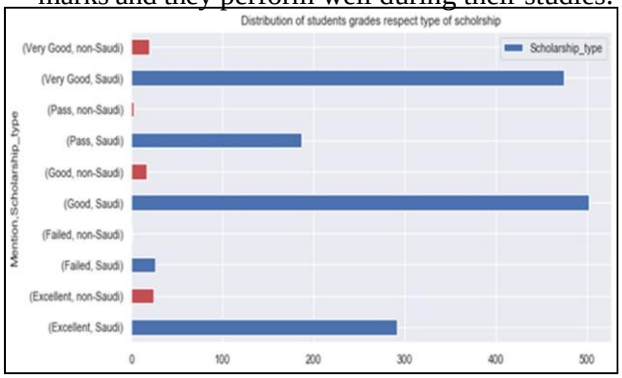

Fig. 6. Distribution of students grades respect scholarship type.

- Predictive analysis

Besides the descriptive analysis that our system provides to decision-makers, it can predict the students' GPA based on their performance during different semesters. As stated in section IV, we employed four well-known classifiers named, SVM, NB, DT, and RF. Table 1 shows the proposed classifiers perform well for predicting the GPA. However, the RF overcomes other classifiers in terms of accuracy, recall, precision, and f-score.

TABLE I. PERFORMANCE OF CLASSIFIERS

	Accuracy	*Recall*	*Precision*	*F-measure*
SVM	89.59%	0.828	0.918	0.7056
NB	90.29%	0.9036	0.9036	0.8964
DT	93.08%	0.9306	0.9	0.891
RF	96.92%	0.9684	0.963	0.963

CONCLUSION

This paper briefly presented our data warehouse system implemented at Taibah University (TU) which is used to support the academic advisory units as well as the decision- makers at the departmental and presidential level.

The proposed approach reinforced the proposed BI tool by making output decisions more profitable. The outputs are categorized into two levels: the descriptive and the predictive level. From the descriptive level, our BI tool showed how the students' grades are distributed over the years as well as the performance of students regarding their gander and type of scholarship. In terms of predictive analysis, four well-known ML algorithms were used to compare with named as support vector machine, naïve Bayes, decision tree, and random forest. The results show while comparing that the random forest algorithm outperformed other algorithms.

An interesting perspective of the proposed work would be integrating uncertainty modeling in the reasoning steps to improve decision-making in the AA process [19] [20]. Moreover, considering the case of big data related to the academic process is a high research topic that should be addressed in future works.