Available online at https://www.irjst.com/
International Research Journal of Science and Technology
ISSN:2707-3955
DOI: https://doi.org/10.46378/irjst.2020.010202

A Model to Determine Factors Affecting Students Academic Performance: The Case of Amhara Region Agency of Competency, Ethiopia

Aklilu Mandefro Messele[1]*, Melkamu Addisu[2]
[1]Department of Computer Science, Bahirdar University, Bahirdar, Ethiopia
[2]Senior Lecturer, Department of Computer Science, Bahirdar University, Bahirdar, Ethiopia

Paper Status

Received : January 2020
Accepted : March 2020
Published : March 2020

Key Words
Algorithm
Classification
Data Mining,
J48
Naive Bayes
Students Performance

Abstract

At this time the amount of data stored in educational institutions is increasing rapidly. These data contain hidden information for improvement of students' performance, guidance, teaching, planning, and so on. Identifying factors that influence students' academic performance help educational stakeholders to take remedial measurements to improve performance of their students. In this paper total of 7,561 students' data covering the period from 2008-2011 with 28 attributes is used to determine the most influential factors. The classification algorithms J48 algorithm and Naive Bayes algorithm is used to develop the model. Design science research methodology is used as a frame work while the hybrid six-step Cios model is followed to develop the model. Many experiments were done with J48 algorithm and Naive Bayes classifier by changing the default values and reducing the number of attributes. However, 8 experiments are presented for analysis which shown better accuracy than the rest. The results of this study have shown that the data mining techniques are valuable for students' performance model building and J48 algorithm resulting in highest accuracy (70.3468% & 83.3552%) for practical and theory exams respectively. It also reveal that Education mode of training experience, Level, Purpose of Assessment, Candidate's category, Age, Sector, Sex, and Employment type found to be the most influential factors for students' academic achievement. Hence, future research directions are pointed out to come up with an applicable system in the area.

Citation: Aklilu Mandefro Messele and Melkamu Addisu.. "Model to Determine Factors Affecting Students Academic Performance: The Case of Amhara Region Agency of Competency, Ethiopia", International Research Journal of Science and Technology, 1(2), 75-87, 2020.

1. Introduction

Education is a type of learning in which the habits, beliefs, values, skills and knowledge of people are shared among generations through narration, discussion, teaching, training, or research. Education gives people critical skills and tools to help them better for themselves and others. Education helps people work better and can create opportunities for sustainable and viable economic growth now and in the future. Education takes place in a school environment, with classrooms of multiple students learning together with a trained teacher. It appears in the form of pre-school, primary, secondary, tertiary, vocational and special.

Technical and Vocational Education and Training (TVET) is focused on direct and practical training that prepares people for specific trades, crafts and careers at various levels from a trade, a craft, technician, or a high professional practitioner position in careers such

* Corresponding author: Aklilu Mandefro Messele
Department of Computer Science, Bahirdar University, Ethiopia
Email: bepositive2712@yahoo.com,aklilu.mandefro@yahoo.com

as engineering, accountancy, nursing, medicine, architecture, law etc [1]. Craft vocations are usually based on manual or practical activities and are traditionally non-academic but related to a specific trade, occupation. It is sometimes referred to as technical education as the trainee directly develops expertise in a particular group of techniques. It can interact with the apprenticeship or internship system.

As described in [2], TVET helps learners to acquire skills, knowledge and attitudes needed to enter the world of work. A quality TVET programs plays an essential role in promoting a country's economic growth and contributing to poverty reduction as well as ensuring the social and economic inclusion of marginalized communities. United Nations educational, scientific and cultural organization /UNESCO/ strongly supports the development of competency-based and employment-led TVET programs that are adapted to a country's socio-economic context and to worldwide technological development [2].

In Ethiopia, TVET is established to create a competent, motivated, adaptable and innovative workforce and to transfer accumulated and demanded technologies in the country, thus contributing to poverty reduction and social and economic development through facilitating demand-driven, high quality technical and vocational education and training, relevant to all sectors of the economy, at all levels and to all people [3].

Students are main assets of educational institutions. Their academic performance plays an important role in producing the best quality graduates who will become great leader and manpower for the country thus responsible for the country's economic and social development. The performance of students is a concern not only to the administrators and educators, but also to corporations in the labor market. Academic achievement is one of the main factors considered by the employer in recruiting workers.

Students in TVET learn level based programs categorized as level one – level four respectively. Up on obtaining satisfactory result in grade ten, a student is admitted to TVET to join level one and continue to the next levels when performing a good result in the certificate of competency (CoC) examination for his/her respective profession. Thus, students have to place the greatest effort in their study to obtain a good grade in order to fulfill their future career.

Analyzing the past performance of enrolled students would provide a better perspective of the possible academic performance of students in the future. This can be achieved using the concepts of data mining. As stated in [4], students' achievement could be subject to diverse socio-demographic variables like personal, social, psychological, natural and other environmental variables. Identifying the factors that affect students' academic achievement using scientific technique could enable to discover the hidden reasons and relationships behind student performance which help in further decision making process. Possibly data mining provides powerful techniques for determining students' academic achievement [5]. The identified factors could enable management team of the institution to allocate appropriate resources and staff, improve their policy, strategies, and enhance curriculums and in turn improve the quality of their education system.

It is proposed that the first-step towards intervention is to investigate that there are many indicators of academic performance that affect different clusters of students. This calls for a thorough analysis of factors that affect academic performance among students.

2. Experimental Procedures, Materials, and Methods

The Design Science (DS) problem-solving paradigm has its roots in engineering and the sciences of the artificial. It is fundamentally a problem solving paradigm. It seeks to create innovations that define the ideas, practices, technical capabilities, and products through which the analysis, design, implementation, management, and use of information systems can be effectively and efficiently accomplished. DS attempts to create things that serve human purposes. It creates and evaluates IT artifacts intended to solve identified organizational problems [37]. It involves a rigorous process to design artifacts to solve observed problems, to make research contributions, to evaluate the designs, and to communicate the results to appropriate audiences. Design is both a process and a product or an artifact. It describes the world as acted upon (processes) and the world as sensed (artifacts).

The figure at 2.1 below represents the conceptual research framework adopted for understanding, executing and evaluating this research. The environment defines the problem space in which reside the phenomena of interest. It is consists of people, organizations, and their existing or planned technologies. In it are the goals, tasks, problems, and opportunities that define business needs as they are perceived by experts within the

organization. Business needs are assessed and evaluated within the context of organizational strategies, structure, culture, and existing business processes. They are positioned relative to existing technology infrastructure, applications, communication, architectures, and development capabilities. Such perceptions are shaped by the roles, capabilities, and characteristics of people within the organization. Given such an articulated business need, a design science research is conducted through the building and evaluation of artifacts designed to meet the identified business need.

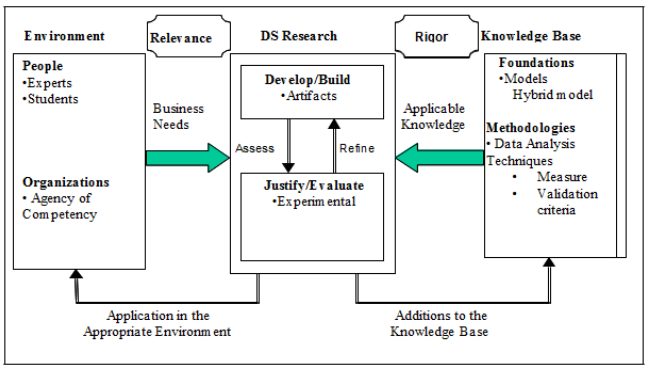

Figure 1. Research Framework [37]

2.1.Literature Review

The researcher has conducted a serious of literature reviews to assess the major issues and concepts in the field of data mining. Various books, journals, articles and papers from the Internet has been read to assess the importance and applications of data mining technology in general and its application on educational data in particular.

2.2. Method Selection

There are various kinds of data mining models such as knowledge discovery database/KDD/, Cross Industry Standard Process for Data Mining /CRISP - DM/, Sample, Explore, Modify, Model, and Access /SEMMA/, Hybrid, and the Two Crows process model. Hence, since this research was conducted for academic purpose and results of the paper may bring solutions to the agency of competency office /AoC/ office, the researcher has used a hybrid six-step Cios KDP model [12] which combines aspects of both industrial and academic models. The model has been chosen since it incorporates all the advantages of CRISP-DM model and includes some extensions which provide a more general, research oriented descriptions of the steps, offers more detailed feedback mechanisms, and also allows modification of the last step throughout the research work. Figure 2. shows descriptions of the six steps of the KDD process model [12]. Based on the hybrid model of the Cios six-step KDP methodology, the following procedures were identified in order to conduct the research work.

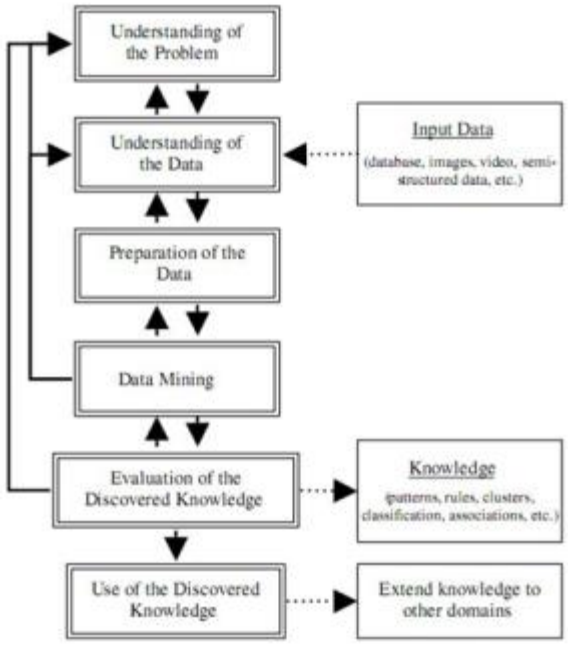

Figure 2. The six-step KDP model as taken from [12]

2.2.1.Understanding of the Problem Domain

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives [12]. Hence, the researcher has worked closely with the Agency of Competency's /AoC/ domain experts to define the problem as depicted in Section 1 and an effort is made to determine the project goals, and oversee current solutions to the problem. Then, the project goals were converted into data mining goals. Serious of interviews with the organization's domain experts has enabled the researcher to define the data mining problem. This has led the researcher to conduct a research by applying data mining to determine the most influential factors that affect the students' performance. For this reason classification technique were applied to develop a model.

2.2.2.Understanding of the Data

The data understanding phase starts with initial data collection and proceeds with activities that enable us to become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information [12]. After understanding the problem to be addressed clearly in this study, the next

step is analyzing and understanding the data available. During this phase, collection of original data is made. Various activities are performed in order to get familiar with the data. Efforts to identify data quality problems are made, which helped the researcher to discover first insights into the data and to detect interesting subsets to form hypotheses for hidden information. Hence, the researcher had collected a total of 7,561 candidate students' real data from the AoC of Amhara Region. The data consists of 28 attributes about students' background information and exam results.

The data used for this research is described thoroughly. The description includes listing out attributes, their possible values, data types, and evaluation of their importance to the problem domain. Careful analysis of the data and its format is made with domain experts by evaluating their relevance to the problem and the particular data mining tasks to be performed.

2.2.3. Preparation of the Data

According to Han and Kamber suggested at [15], attention should not be neglected to clean data for knowledge mining because the real world data is highly susceptible to noisy, inconsistency and incompleteness. The researchers state that the need for data preparation is, today's real-world data are highly vulnerable to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources. Low quality data will lead to low quality mining results.

According to [21] one of the most important tasks in data mining is preparing the data in a way that is suitable for the specific data mining tool or software package to be used. Data preparation involves data selection, data cleaning, data construction, data integration and data formatting.

There are a number of data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse. Data transformations, such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance. These techniques are not mutually exclusive; they may work together. For example, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a date field to a common format. Data processing techniques, when applied before mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining.
The candidate's real data, which is employed for the purpose of this study, were subject to a number of limitations. These include missing values, outliers and inconsistency in some of the attribute values. However, all these were considered critically during the study.

2.2.3.1. Data Selection

Originally there were around 7,561 records. Since this dataset contains irrelevant and unnecessary data, all are not used for training and testing. So, after eliminating irrelevant and unnecessary datasets, only a total of 5,335 datasets are used for the purpose of conducting this study.

2.2.3.2. Data Cleaning

As described by Han and Kamber at [15], real world databases contain incomplete, noisy and inconsistent data due to the size of databases. Such unclean data may cause confusion for the data mining process. Data cleaning attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Consequently, data cleaning has become a must in order to improve the quality of data so as to improve the performance of the accuracy and efficiency of the data mining techniques. The researcher used MS-Excel application for cleaning the data. Accordingly, different data cleaning tasks were carried out.

2.2.3.2.1. Handling of Missing Values

Missing attribute values in the data is most likely associated with unavailability of interesting information, lack of knowhow on the importance of data at the time of entry, misunderstanding of the data during encoding, the respondents him/herself may refuse to answer certain questions or they may not know the answer exactly or may answer in an unexpected manner. Susceptibility of data for noisy, in consistency and incompleteness is becoming dominant in real world; this is mainly related with its huge size and heterogeneous source of data [15]. Thus, it is important for the researcher to manage missing values efficiently.

As it is suggested in [15], Missing values for an attribute could be filled using various techniques. Among these; ignoring the tuple with missing value, fill in the missing value manually, use a global constant to fill in the missing value, use the attribute mean for all samples belonging to the same class as the given tuple, or use the most probable value to fill in the missing value.

The researcher observed missing values in three of the attributes selected for this study. Hence, different ways of missing value handling were employed for each attribute with missing values.

The researcher observed around 1873 missing values in the "Candidate's category" field. Candidate's category can be derived from "Education mode of training experience" and "Employment type" field. For example if candidate's education mode of training experience is Diploma and is employed, then most probably her/his category is either 1C-level teacher or Industry worker. Hence, the researcher has removed a total of 1852 records whose both attribute values were empty. The remaining 21 missing values were filled manually as it was derived from the candidate's education mode of training experience and employment type.

There were 811 missing values in the "Sex" field. Then, the researcher derived these values from the names of the candidates. Accordingly, 539 values were filled manually and the remaining 272 records were removed whose names were ambiguous to identify the gender.There were also 833 missing values in the "age" field. Then, the researcher used the mean value of the available age values from the record. Average Mean = Sum (all values of age) / total number of records, this was calculated in Microsoft Excel aggregate functions and result found is 25. All the age missing values were then replaced with this value using the filtering mechanism of Microsoft Excel.

2.2.3.2.2. Handling of Outlier Value

According to [15], outlier is a data value that does not (or is not thought to have) come from the typical population of data. Outliers are values that fall outside the boundaries that enclose most other values in the data. This can apply to values of an attribute, or of entire cases. Outlier treatment is the approach to replacing outliers in numerical data attributes. There are several techniques including specifying explicit boundaries, percentages in the tails of the distribution, and number of standard deviations, such that values outside the valid range are replaced either by null values or edge values.

In this study, the researcher observed 102 records whose age values were less than 12 (which are very rare to be a candidate for CoC examination) and such age value may affect the modeling result and the researcher considered it as an outlier). Therefore, the researcher discussed with the domain experts and decided to remove the records to avoid bias.

2.2.3.2.3. Handling of Noisy Value

According to Han and Kamber [15], Noise is a random error or variance in a measured variable. Hence the researcher found inconsistent values in several attributes and had corrected them manually based on the value of the corresponding data.

For this reason, the values for "Candidate's category" field originally lie on 1A-level teacher, 1B-level teacher, 1C-level teacher, Industry worker, TVET graduate, and/or others. However, there were 47 records whose values were found to be entered in numerical representation. On the other hand, the value for the field "Purpose of Assessment" was either 1 for employment or 2 for higher education. However, there were 91 records whose values were written as for higher education and/or for employment. Hence the researcher corrected these inconsistencies manually by cross-checking the value for each number.

2.2.3.3. Data Transformation

According to [15], in data transformation; data are transformed or consolidated into forms appropriate for mining process. It involves smoothing, aggregation, generalization, normalization, and attributes construction.

Generalization of the data were done in the effort to prepare the data ready for the data mining techniques to be undertaken in this research. During generalization, low-level data are replaced by higher-level concepts through the use of concept hierarchies. For example, categorical attributes, like street, can be generalized to city or country. In the same way, values for numerical attributes, like age, could be mapped to youth, middle-aged, and old.

Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute in to intervals [15]. Interval labels can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels there by reduces and simplifies the original data. This leads to a concise, easy to use, knowledge-level representation of mining results.

Accordingly, the researcher performed discretization on the attribute "Age" using a binning method. The attribute "Age" is binned into three equal frequency bins and the smoothed bins were taken as young, middle, and old as it is presented in table 4.1 below. It was discretized to reduce the distinct values of the attributes so that it suit the mining tool and obtained meaningful patterns. The reason for choosing binning method is that; the generalized data may be more meaningful and easier to interpret results, and

generalized dataset has special pattern identification capability. Table 1 shows discretization of the age values.

Table 1. Discretization of the age values into three classes

S.No	Age Group	Replaced Value
1	[17,30]	Young
2	[31,44]	Middle
3	[45,58]	Old

2.2.3.4. Data Reduction

As the data collected for the purpose of this study was huge, this could have slowed down the mining process. Hence, the size of the dataset had to be reduced before the data mining results.

According to [15], Data reduction obtains a reduced representation of the dataset that is much smaller in volume, yet produces the same (or almost the same) analytical results. There are number of strategies for data reduction. These include data aggregation (e.g., building a data cube), attribute subset selection (e.g., removing irrelevant attributes through correlation analysis), dimensionality reduction (e.g., using encoding schemes such as minimum length encoding), and Numerosity reduction (e.g., "replacing" the data by alternative, smaller representations such as clusters or parametric models). Data can also be "reduced" by generalization with the use of concept hierarchies, where low-level concepts, such as Woreda for candidate's address, are replaced with higher-level concepts, such as zone. A concept hierarchy organizes the concepts into varying levels of abstraction. Among those, Numerosity reduction was applied on the data collected for this study.

2.2.3.4.1. Attribute Subset Selection

The whole target dataset were not taken for the data mining task. Since the main objective of this research was to apply data mining technology to determine factors affecting students' performance; we have excluded many features that are not necessary for this study. Among these; the attributes like name, address, name of school graduated, type of ownership of training institution, employer, knowledge test signature, practical test signature, date, mobile number, registration number, score of the knowledge test, name of the assessment center, assessment center address, assessor's name, mobile number of the assessor, supervisor's name, and occupation with all their information were eliminated before starting the actual data mining function. According to the suggestion of the domain experts, these features are believed to have no value in the data mining process and are excluded with their information. Only data's relevant and believed by domain experts to the analysis task are encoded. Hence, the attributes that were finally selected for the mining process are described below in table 2.

Table 2. Final list of selected attributes with their descriptions

Attribute	Data Type	Description
Sex	Nominal	Gender of the candidate
Age		Age of the candidate
Candidate's Category	Nominal	Candidate's certification category
Purpose of Assessment	Nominal	Reason for taking the exam
Education mode of Training	Nominal	Educational level of the candidate
Experience		
Employment type	Nominal	Employment status of the candidate
Level	Nominal	TVET's level to which the candidate belongs
Sector	Nominal	Sector of the candidate's field of study
Knowledge Test	Nominal	Knowledge test result of the candidate
Practical Test	Nominal	Practical test result of the candidate

2.2.3.5. Dataset Format

WEKA is chosen as an implementation tool for this study. Thus, WEKA needs data to be prepared in some formats and file types. WEKA accepts records whose attribute values are separated by commas and saved in an ARFF (Attribute-Relation File Format) file format. The datasets provided to this software are prepared in a format that is acceptable for WEKA software.

In order to prepare the data in such format the records encoded into the Microsoft excel database are saved as a Comma Delimited (CSV) file. Once all processing is completed and the file is converted to .csv format, WEKA either process the .csv format itself or a file in the form of Attribute Relation File Format (.arff). For this study the data is given to the software in .arff format. A screen shot of the data is depicted.

2.2.4.Data mining

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary [12].

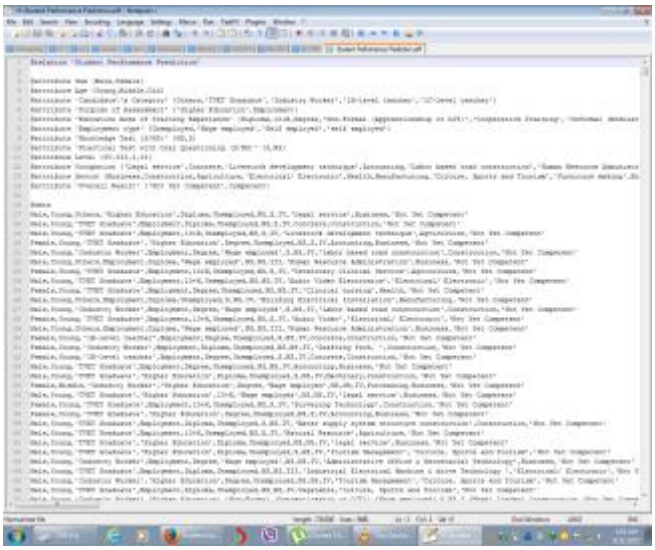

Figure 3. A screen shot of the data .arff file format

In this step of the data mining methodology, different algorithms supported by the selected data mining tool are examined by taking into consideration their application to the problem domain. Models are built and evaluations are carried out automatically to select the best model. Finally the appropriate model is recommended for use by the decision makers.

The main objective of this research work is to build model for determining factors affecting students' academic achievement. Having this purpose in mind, the supervised classification technique is employed. Classification is selected because it is the appropriate and more suitable technique for student performance [16]. For doing so, Decision tree (J48) and Naïve Bayes algorithm were applied to develop the model. The reason for selecting Decision Tree and Naïve Bayes was that, decision trees are easy to understand, implement and use, handles all types of attributes and missing values it is also computationally cheap. Naïve Bayes is also easy to implement, robust to isolated noise points, handles missing values by ignoring the instance during probability estimate conclusions, robust to irrelevant attributes and also various literatures and articles show that J48 and Naïve Bayes algorithms perform at high accuracy when applied in building model for student performance [36].

2.2.4.1. Implementation Tool Selection

For conducting this research work, the WEKA (Waikato Environment for Knowledge Analysis) version 3.7.4 toolkit is selected. WEKA is a widely used toolkit [13] for machine learning and data mining projects that was originally developed at the University of Waikato in New Zealand. WEKA contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. It was selected over the others [14] like Tanagra, KNIME (Konstanz Information Miner), and Orange in terms of computational perspective, wider range of algorithms, better data preparation tools and its support for very large datasets. It is also freely available and is platform independent since it is written in Java.

2.2.5.Evaluation of the Discovered Knowledge

Performance evaluation of the methods chosen to develop the model was evaluated to interpret the knowledge patterns properly. The methods are true positive rate, false positive rate, accuracy, precision and 10 folds cross validation and 75% split for training and 25% for testing, test option was used to measure the performance of the model. 25% of the dataset were selected on random set selection from the initial raw data to evaluate the accuracy of the developed model.

3.Results and Discussion

The study focuses on identifying determinant factors affecting students' performance and building a classification model. Various classification models have been built by using decision tree, and Naive Bayes algorithms. The models have been tested on different number of the selected attributes and the significance of the outputs of the most important model is presented for analysis to the domain experts. Finally, the model with the best performance is selected.

3.1.Classification Modeling Using J48 Decision Tree

Analysis of the decision tree models are made in terms of detailed accuracy of the classifier on the training dataset as tested on the tested data based on a confusion matrix of each model result. The confusion matrix is a useful tool for analyzing how well the classifier can recognize tuples of different classes. Confusion matrix shows four important numerical quantities namely: True-Positive, False-Positive, False-Negative, and True-Negative as shown below.

Table 3. The Confusion Matrix

	Predicted Class	
Actual Class	Yes	No
Yes	True-Positive (TP)	False-Negative (FN)
No	False-Positive (FP)	True-Negative (TN)

The True-Positive and True-Negative are correct classifications. A False-Positive occurs when the outcome is incorrectly predicted as yes when it is actually no. A False-Negative occurs when the outcome is incorrectly predicted as no when it is actually yes.

For starting the classification experimentation, J48 decision tree and Naive Bayes were selected. There were a total of twelve experiments to be done for practical and theory examinations having six experiments for each. The experimentations were experimented, analyzed and compared to each other in terms of different performance matrix values, accuracies, number of leaves, the size of trees generated and execution time. The models were also compared with the discovered knowledge and the judgment of the experts.

3.1.1.Classification Modeling For Practical Exam

3.1.1.1. Experimentation I: J48 with 10-Fold Cross-Validation

The first experimentation was conducted for practical exam using the eight independent attributes that are selected during the data preparation phase. These are Sex, Age, and Candidate's category, Purpose of assessment, Education mode, Employment type, Level, and Sector. The experiment is conducted using 10-fold cross validation technique using the classification model J48 decision tree. Table 4 shows the performance of the J48 decision tree model on the given dataset.

Test mode : 10-fold cross-validation
Number of Leaves : 37
Size of the tree : 51
Time taken to build mode l : 0.06 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances: 3753 and 70.3468 %
Incorrectly Classified Instances: 1582 and 29.6532 %

The base for calculating correctly classified instances and incorrectly classified instances is the confusion matrix. The confusion matrix of the class which is a base for calculating accuracy measures and performance is presented below.

Table 4. Confusion Matrix of J48 with 10-fold cross-validation

	Predicted Class	
Actual Class	Satisfactory	Not Satisfactory
Satisfactory	2332	538
Not Satisfactory	1044	1421

The number of True-positives in the above confusion matrix is 2332 records. Those records which were predicted as 'Satisfactory' class by the classifier and also happened in 'Satisfactory' actually are True-Positives. The number of the records (1421) which were classified to the 'Not Satisfactory' class by the classifier and they are actually in 'Not Satisfactory' are True-negative. The sum of the True-positive and True-negative give us correctly classified instances. The total number of records which were correctly classified to 'Satisfactory' and 'Not satisfactory' classes was 3753 (70.3468%) while 1582 (29.6532%) records are incorrectly classified.

3.1.1.2. Experimentation II: J48 with Percentage Split

The second experiment of the J48 decision tree for practical exam was conducted with percentage split 75% for training and 25% for testing. The first line reports the split point for training and testing dataset.

Test mode is split 75.0% train, remainder test and the
Number of Leaves : 37
Size of the tree : 51
Time taken to build model : 0.05 seconds
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances are 929 and 69.6924 %
Incorrectly Classified Instances are 404 and 30.3076 %

Table 5. Confusion Matrix of J48 with Percentage Split

	Predicted Class	
Actual Class	Satisfactory	Not Satisfactory
Satisfactory	570	132
Not Satisfactory	272	359

As it can be seen in table 5, the percentage split with 75% training set and remaining 25% for testing purpose was applied. In the above confusion matrix, the number of correctly classified instances is 929 (69.6924%) and 404 (30.3076%) incorrectly classified out of 1333 records. This shows that this experiment hasn't improved the accuracy of the model as shown in the confusion matrix. When compared with previous experiment, the accuracy measures and other matters of this experiment are lower with the accuracy of 69.69%.

3.1.1.3. Experimentation III: Naïve Bayes with 10-Fold Cross-Validation

The third experiment for practical exam was conducted using 10-fold cross validation technique using the classification model Naive Bayes. The performance of the Naive Bayes model on the given dataset.

Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Time taken to build model: 0.02 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances are 3334 and 62.493%
Incorrectly Classified Instances are 2001 and 37.507%

Table 6. Confusion Matrix of Naive Bayes with 10-fold Cross-Validation

	Predicted Class	
Actual Class	Satisfactory	Not Satisfactory
Satisfactory	2014	856
Not Satisfactory	1145	1320

The above table 6 shows the model confusion matrix and accuracy of Naive Bayes classification model with 10-fold cross-validation. As shown in the above table out of 5335 records of the dataset, 3334 (2014+1320) records are classified correctly, while 2001 (1145+856) are incorrectly classified with an accuracy of 62.493%. This still doesn't improve the accuracy of the previous model.

3.1.1.4. Experimentation IV: Naive Bayes with Percentage Split

The fourth experiment for practical exam was conducted with percentage split 75% for training and 25% for testing using the classification model Naive Bayes. The performance of the Naive Bayes model is on the given dataset.

Test mode: split 75.0% train, remainder test
=== Classifier model (full training set) ===
Time taken to build model: 0.01 seconds
=== Evaluation on test split ===
Time taken to test model on training split: 0.01 seconds
=== Summary ===
Correctly Classified Instances 811 60.8402%
Incorrectly Classified Instances are 522 and 39.1598%

Table 7. Confusion Matrix of Naive Bayes with percentage split

	Predicted Class	
Actual Class	Satisfactory	Not Satisfactory
Satisfactory	486	216
Not Satisfactory	306	325

The above table 7 shows the model confusion matrix and accuracy of Naive Bayes classification model with percentage split. The number of correctly classified instances is 811 (60.8402%) and 522 (39.1598%) incorrectly classified out of 1333 records. This shows that this model has deteriorated the accuracy of the previous models resulting in 60.84%.

3.1.2.Classification modeling for Theory exam

3.1.2.1. Experimentation I: J48 with 10-Fold Cross-Validation

The first experimentation for theory exam was conducted using the eight independent attributes that are selected during the data preparation phase. These are Sex, Age, and Candidate's category, Purpose of assessment, Education mode, Employment type, Level, and Sector. The experiment is conducted using 10-fold cross validation technique using the classification model J48 decision tree. Table 8 shows the performance of the J48 decision tree model on the given dataset.

Test mode: 10-fold cross-validation
Number of Leaves: 24
Size of the tree: 33
Time taken to build model: 0.04 seconds
=== Summary ===
Correctly Classified Instances are 4447 and 83.3552 %
Incorrectly Classified Instances are 888 and 16.6445 %

The confusion matrix of the class which is a base for calculating accuracy measures and performance is presented below.

Table 8. Confusion Matrix of J48 with 10-fold cross-validation

	Predicted Class	
Actual Class	Satisfactory	Not Satisfactory
Satisfactory	384	791
Not Satisfactory	97	4063

As it can be seen from the confusion matrix that resulted from the model developed by the J48 decision tree algorithm with the 10-fold cross-validation, the model scored an accuracy of 83.3552%. This shows that from the total 5335 test data, 4447 (85.3552%) of the records are correctly classified, while 888 (16.6448%) of them are misclassified. Hence, this model has improved the accuracy of the classification models gained so far.

3.1.2.2. Experimentation II: J48 with Percentage Split

The second experiment of the J48 decision tree for theory exam was conducted with percentage split 75% for training and 25% for testing. The following depicts the result gained from the decision tree.

Test mode: split 75.0% train, remainder test
Number of Leaves : 24
Size of the tree : 33
Time taken to build model: 0.04 seconds
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances are 1096 and 82.2206%
Incorrectly Classified Instances are 237 and 17.7794%

Table 9. Confusion Matrix of J48 with Percentage Split

Actual Class	Predicted Class: Satisfactory	Predicted Class: Not Satisfactory
Satisfactory	92	214
Not Satisfactory	23	1004

As shown in table 9, the percentage split with 75% for training and 25% for testing was applied. This resulted 1096 (82.2206%) records were classified whereas 237 (17.7794%) records were misclassified. This model has lowered the accuracy of the previous model which had shown an accuracy of 83.3552% however it is still better than the other models built so far.

3.1.2.3. Experimentation III: Naïve Bayes with 10-Fold Cross-Validation

The third experiment for theory exam was conducted using 10-fold cross validation technique using the classification model Naive Bayes. The table below at 2.10 shows the performance of the Naive Bayes model on the given dataset.

Table 11 Summary of the J48 decision tree and Naïve Bayes models

Experiment No		Type of classifier	Number of Leaves	Size of tree	Correctly classified instances	Incorrectly classified instances
I	Practical	J48 decision tree with 10-fold cross validation	37	51	70.3468%	29.6532%
II		J48 decision tree with percentage split	37	51	69.6924%	30.3076%
III		Naive Bayes with 10-fold cross-validation			62.493%	37.507%
IV		Naive Bayes with percentage split			60.8402%	39.1598%
V	Theory	J48 decision tree with 10-fold cross-validation	24	33	83.3552%	16.6448%
VI		J48 decision tree with percentage split	24	33	82.2206%	17.7794%
VII		Naive Bayes with 10-fold cross-validation			75.7638%	24.2362%
VIII		Naive Bayes with percentage split			75.994%	24.006%

3.1.2.4. Experimentation IV: Naïve Bayes with Percentage Split

The fourth experiment for theory exam was conducted with percentage split 75% for training and 25% for testing using the classification model Naive Bayes. Table 12 shows the performance of the Naive Bayes model on the given dataset.

Test mode: split 75.0% train, remainder test
Time taken to build model: 0.01 seconds
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances are 1013 and 75.994%
Incorrectly Classified Instances of 320 24.006%

Table 12.Confusion Matrix of Naive Bayes with percentage split

Actual Class	Predicted Class: Satisfactory	Predicted Class: Not Satisfactory
Satisfactory	82	224
Not Satisfactory	96	931

The above table 12 shows the model confusion matrix and accuracy of Naive Bayes classification model with percentage split. The number of correctly classified

Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Time taken to build model: 0.01 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 4042 75.7638%
Incorrectly Classified Instances 1293 24.2362%

Table 10.: Confusion Matrix of Naive Bayes with 10-fold Cross-Validation

Actual Class	Predicted Class: Satisfactory	Predicted Class: Not Satisfactory
Satisfactory	290	885
Not Satisfactory	408	3752

As shown in the above table out of 5335 records of the dataset, 4042 (75.7638%) records are correctly classified while 1293 (24.2362%) are incorrectly classified.

instances is 1013 (75.994%) and 320 (24.006%) incorrectly classified out of 1333 records.

3.1.3.Comparison of Naïve Bayes and J48 Decision Tree Models

A summary to the models built by the J48 decision tree and the Naive Bayes is presented in table 11.

As it can be seen from the above table, the first and the second four experiments are developed for practical and theory respectively. J48 decision tree has shown better accuracy than Naive Bayes at both the practical and theory model. It is also shown that the number of leaves and size of the tree is the same both on the first and the second experiments of the practical model. However, the first experiment has resulted better accuracy of 70.3468% than the second experiment which has an accuracy of 69.6934%. During the experimentation for the theory, the decision tree has brought the same number of leaves and size of the tree. However, J48 decision tree with 10-fold cross-validation has resulted an accuracy of 83.3552% while J48 with percentage split resulted an accuracy of 82.2206%. This shows J48 decision tree with 10-fold cross-validation has better accuracy of 70.3468% and 83.3552% both at practical and theory respectively.

3.1.4.Evaluation of the Discovered Knowledge

This stage of the data mining task includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared. The fact that, the classification capability of any model depends on the learning ability of the datasets provided, the model with the highest true positives and less false negatives means there is a likelihood of classifying the datasets correctly making the precision and the recall values to be maximum. Hence, the model with a maximum precision and recall value was chosen as an evaluation parameter. So in this research work, J48 algorithm was chosen having better prediction and accuracy. The developed model was evaluated using the test set data (1334) prepared for evaluation purpose.

4.Conclusion

This research attempted to study the application data mining on educational data to find interesting patterns and identify factors affecting students' performance. The study was conducted based on the data mining steps or process discussed in section 2: defining the problem domain, data understanding, data preparation, data mining, evaluation of the discovered knowledge, and finally use of the discovered knowledge.

Data collection, selection and cleaning were major tasks which took most of the experimental time of the research. This is due to higher volume or size of the data residing in the agency's record office. Originally, the total number of datasets was 7561 having 28 attributes. Hence, the total number of datasets becomes 5335 after it was fully prepared for building the model. The given datasets were also divided into two as practical and theory and only 8 attributes were found to be relevant for building the model after a deep discussion with the domain experts. This study reveal that Education mode of training experience, Level, Purpose of Assessment, Candidate's category, Age, Sector, Sex, and Employment type found to be the most influential factors for students' performance during the practical exam while Level, Candidate's category, Sex, Age, Employment type, Sector, Purpose of Assessment, and Education mode of training experience are the most influential factors for students' performance during the theory exam.

The outputs of the models were presented for analysis to domain experts for feedback. To achieve this goal the hybrid data mining methodology has been adopted and the WEKA 3.7.4 data mining tool has been used to implement the J48, and Naive Bayes classification models.

Various experiments are made iteratively by making change of the values to come up with a meaningful output of classification model. Usually, both practical and theory datasets with training set used to build the model and test set used to validate by 10-fold cross validation and percentage split. The comparison of the models using WEKA's experimenter showed better result using J48 algorithm with 10-fold cross validation with an accuracy of 70.3468% and 83.3552% respectively for both practical and theory datasets.

In general, the results from this study were encouraging. It was possible to identify the most influential factors affecting students' academic performance using data mining techniques that made good meanings to domain experts. It is the researcher's belief that a more thorough study using data mining techniques can help to understand more about students' performance in the region.

5.Acknowledgement

Above all I am thankful to God for letting me reach here and paving all the way in front of me.

It is my pleasure to forward my deep gratitude to my advisor Dr.Melkamu Addisu (PhD), for his encouragement, guidance, and unconditional support starting throughout my research work. Next to this my heartfelt thanks goes to my co-advisor Mr. Daniel Abebe whose comments, guidance, and consultations has realized this research to be completed. Zeme, Elias thank you for feeding me with the necessary data I want for this research work. Big respect goes to my friends and classmates for sharing all the fun and jokes during my stay in Bahirdar, Ethiopia.

Last but not least, I would like to extend thank to my father, mother, brothers, and sisters for all the time we spent since childhood. Dad! You have shown me all the way to be the man I am now. And thank you very much for your trust on me.

6. References

[1]. K. King, "Eight Proposals for a Strengthened Focus on Technical and Vocational Education and Training in the Education for All Agenda," Paper commissioned for the EFA Global Monitoring Report 2012, 2012.

[2]. B. Aboubakr, "TVET and Entrepreneurship Skills," UNESCO-UNEVOC Revisiting global trends in TVET

[3]. FDRE, "Education Sector Development Program IV," Federal Ministry of Education, 2010.

[4]. A. AL-Malaise, A. Malibari, and M. Alkhozae, "Student's Performance Prediction System Using Multi Agent Data Mining Technique," International Journal of Data Mining & Knowledge Management Process, vol. 4, no. 5, September 2011.

[5]. Smita and P. Sharma, "Use of Data Mining in Various Field: A Survey Paper," IOSR-Journal of Computer Engineering, vol. 16, no. 3, pp. 18-21, 2010.

[6]. M. Durairaj and C.Vijitha, "Educational Data Mining for Prediction of Student Performance Using Clustering Algorithms," International Journal of Computer Science and Information Technologies, vol. 5, pp. 5987-5991, 2008.

[7]. Edukans Foundation, "Technical Vocational Education and Training in Ethiopia Mapping," Edukans Foundation, January 2009.

[8]. B. Rahel and M. Wolfgang, "A Bayesian Approach to Predict Performance of a Student (BAPPS): A Case with Ethiopian Students," Proc. IASTED International Conference on Artificial Intelligence and Applications, 2005.

[9]. M. Ramaswami and R. Rathinasabapathy, "Student Performance Prediction Modeling: A Bayesian Network Approach," International Journal of Computational and Informatics, vol. 1, no. 4, January-March 2012.

[10]. V.O. Oladokun, A.T. Adebanjo, and O.E. Charles-Owaba, "Predicting Students' Academic Performance using Artificial Neural Network: A Case Study of an Engineering Course," The Pacific Journal of Science and Technology, vol. 9, no. 1, May-June 2008.

[11]. J.F. Superby, J.P. Vandamme, and N. Meskens, "Determination of factors influencing the achievement of the first-year University students using data mining methods," In Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems , pp. 37-44, 2006.

[12]. K. Cios, P. Witold, S. Roman, and A. Kurgan, 2007, "Data Mining: A Knowledge Discovery Approach," Springer, New York: USA.

[13]. B. J. Sudhir and B. G. Kodge, "Census Data Mining and Data Analysis Using WEKA," International Conference in Emerging Trends in Science, Technology and Management, 2009.

[14]. H. Abdullah, A. Qasem, N. Mohammed and M. Emad, "A Comparison Study between Data Mining Tools over some Classification Methods," International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence, pp. 18-26.

[15]. J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Second Edition, Morgan Kauffman Publishers, San Francisco.

[16]. Baker, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier.

[17]. G. Bontempi, "Data mining for prediction," University of Minnesota, 1999.

[18]. R. Khalid, "Application of Data Mining in Bioinformatics," Indian Journal of Computer Science and Engineering, vol. 1, no. 2, pp. 114-118.

[19]. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," AI Magazine, pp. 37-54, 1996.

[20]. G. Lijia, "Applying Data Mining Techniques in Property/ Casualty Insurance," University of Central Florida.

[21]. Two Crows Corporation, "Introduction to Data Mining and Knowledge Discovery," 2005.

[22]. O. Marban, G. Mariscal, and J. Segovia, "A Data Mining and Knowledge Discovery

Process Model," I-Tech Education and Publishing, 2009.
[23]. S. Umair and Q. Haseeb, "A Comparative Study of Data Mining Process Models (KDD, CRISP-DM, and SEMMA)," International Journal of Innovation and Scientific Research, vol. 12, no. 1, pp. 217-222, Nov. 2011.
[24]. S. Fadzilah and A. Mansour, "Knowledge-Oriented Applications in Data Mining," InTech, January 2011.
[25]. M. Bharati, "Data Mining Techniques and Applications," Indian Journal of Computer Science and Engineering, vol. 1, no. 4, pp. 301-305.
[26]. P. Neelamadhab, M. Pragnyaban and P. Rasmita, "The Survey of Data Mining Applications and Feature Scope," International Journal of Computer Science Engineering and Information Technology, vol. 2, no. 3, June 2012.
[27]. S.D. Gheware, A.S. Kejkar and S.M. Tondare, "Data Mining: Task, Tools, Techniques and Applications," International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, issue 10, pp. 8095-8098, October 2009.
[28]. A. Azwa, H. Nor and A. Fadhilah, "First Semester Computer Science Students' Academic Performances Analysis by Using Data Mining Classification Algorithms," Proceeding of the International Conference on Artificial Intelligence and Computer Science, 15-16 September 2011, Bandung, Indonesia.
[29]. S. Mamta and S. Jyoti, "Machine Learning Techniques for Prediction of Subject Scores: A Comparative Study," International Journal of Computer Science and Network, vol. 2, issue. 4, pp. 77-80, August 2009.
[30]. D. Lalit and R. Jayant, "A Decision Support System for Predicting Student Performance," International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, issue. 12, pp. 7232-7237, December 2010.
[31]. P. Jiban, "Usefulness and Applications of Data Mining in Extracting Information from Different Perspectives," Annals of Library and Information Studies, vol. 58, pp. 7-16, March 2011.
[32]. A. F. ElGamal, "An Educational Data Mining Model for Predicting Student Performance in Programming Course," International Journal of Computer Applications, vol. 70, no. 17, pp. 22-28, May 2010.
[33]. M. Abdous, He. W. & Yen. Cherng-Jyh, "Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade," Educational Technology & Society, vol. 15, no. 3, pp. 77-88, 2010.
[34]. A. Muluken, "Application of Data Mining Techniques for Student Success and Failure Prediction: The Case of Debre Markos University," International Journal of Scientific & Technology Research, vol. 4, issue. 04, pp. 91-94, April 2010.
[35]. P. Saurabh, "Mining Educational Data to Reduce Dropout Rates of Engineering Students," International Journal of Information Engineering and Electronic Business, vol. 2, pp. 1-7, 2012.
[36]. T. Mahendra, B. Manu and Y. Omprakash, "Performance Analysis of Data Mining Algorithms in Weka," IOSR Journal of Computer Engineering, vol. 6, issue.3, pp. 32-41, 2012.
[37]. R. H. Alan, T. M. Salvatore and P. Jinsoo, "Design Science in Information Systems Research," MIS Quarterly, vol. 28, no. 1, pp. 75-105, March 2004.