Mining CMS Data to Understand Students' Learning Issues

Prakhar Gautam
M.Tech Student CSE,
Gyan Ganga Institute of Technology and Sciences,
India.

Santosh K. Vishwakarma
Associate Professor,
Gyan Ganga Institute of Technology and Sciences,
India.

ABSTRACT
Students face a lot of problems in their college/engineering life. CMS (Content Management System) is a platform for students to post their problems and let the authorities know what exactly their issues are. The data collected from students is huge. It's important to extract some useful 'knowledge' from this data. Data Mining, which is a process of extracting useful information from a huge dataset, is applied to the CMS data to understand students' learning issues. This way, they can have a better future and a good academic career. Traditionally, educational researchers have been using methods such as surveys, interviews, to collect data, which is very time consuming and inefficient. Also, these methods have not given much insight into students' problems. Researchers have also used social media data, but the social media data is unreliable, unauthentic and mostly anonymous. In this dissertation work, the focus is on mining CMS data, which is authentic and real, as it doesn't allow users to go anonymous. CMS data is much more reliable as compared to other platforms.

In this dissertation work, data mining technique known as Classification (where the Engineering students' problems are classified into certain classes) is used to implement a model where students' problems can be analysed which they face in their day to day college life, and also suggest the solutions for the same.

The knowledge extracted after applying Data Mining algorithms will be very useful for policy makers and educators in making informed decisions. The data generated by engineering students in future can also be mined and solutions can be provided instantly.

Keywords
Students' problems, Engineering Students', Data Mining, RapidMiner, Text Mining, CMS, Classification, Naive Bayes Classifier.

1. INTRODUCTION
At 315 million, India has most students in the world. Particularly, "Engineering graduates" constitute a significant part of the nation's future workforce. Thus; Engineering Students' have a direct impact on the nation's economic growth. Aspiring Minds, (a private company) did a survey and found that over 80% of engineering students in India are unemployable[1]. This shows how serious this issue is and how important it is to get a long term solution for better future of Engineering Students. It's important that Students' are free of problems and distractions; so that they can concentrate on their studies and build a better career/future. In this work, the analysis is done on Engineering Students' problems which they face in their college life. 3 years of Data comprising of Students' problems is taken; and those problems are analysed so that the solutions can be provided for the same. The data collected is huge and manually extracting such useful information from this huge dataset is not feasible. One such approach which can be used for this task is Data Mining. Data Mining is defined as a process of extracting useful and hidden information from large databases. Data will be of no use if it is not converted it into something useful. Another term used for Data Mining is that it is a "knowledge discovery" process. In other words, Data mining is the procedure of mining "knowledge" from data. This knowledge can be used to increase revenue, cuts costs, or both. The retailers of grocery products increase their sales through Data Mining (through analysing customer purchase history). From Loan Payment prediction (weather a loan should be given to a customer based on his/her past data) to detecting financial fraud detection; Data Mining is used everywhere. Data Mining techniques are not limited to text. Mining can be performed on image/video data termed as Multimedia Mining; or web data called as web mining. In this dissertation work, the focus is on Text mining. Text Mining is defined as deriving knowledge from Text; which can be either stored in a Notepad text file, word file or an excel sheet.

Dataset is taken from CMS (Content Management System) which is a software application used to create and manage digital content. CMS is widely used by institutes to store students' records in an excel sheet (text data) which allows Students' to post their problems and the same is stored in the system so that the solutions can be provided for these issues/problems. There are various data mining techniques such as Classification, Clustering, Rule-Association, and Regression which are used to discover knowledge from huge Dataset.

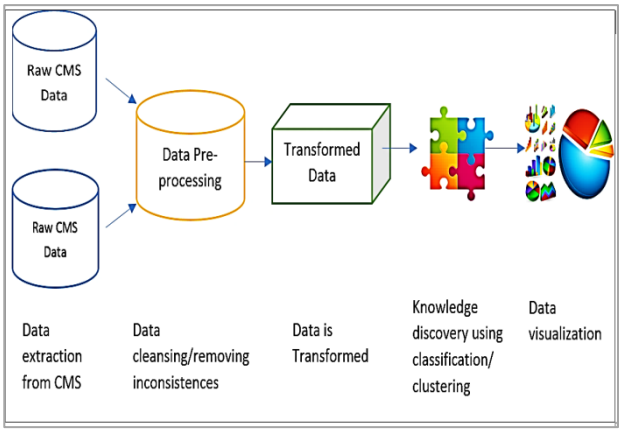

Fig1: Data Mining Process

In this work, a very popular Data Mining technique called as Classification is used to classify Students' problems. In classification, each item in a set of data is classified into predefined set of classes or categories. The goal of classification is to accurately predict the target class for each record in the dataset. For example, a classification model could be used to identify loan applicants as low, medium, or

high credit risks; and on that basis, Loan or Credit card could be approved or rejected to the customers. To classify students' records, Naive Bayes Classifier has been used.

Naive Bayes classifier is based on Bayes' probability theorem. It learns the probability of an object with certain features belonging to a particular group/class. So based on these certain set of features, it predicts the target class for the items.

To perform our Data mining task, software tool called as RapidMiner has been used. RapidMiner is an open source Data science platform which is free to use and is build using Java. RapidMiner provides excellent set of tools for predictive analytics models and makes it easy to get the results without getting into the hassle of writing codes as it eliminates the need to write a code and also provides fast results.

RapidMiner is used to perform predictive analysis on past 3 years of Students' records. This Dataset is taken from CMS software application comprising of Students' problems. The CMS data is taken from Gyan Ganga Institute, Jabalpur.

2. LITERATURE REVIEW

After doing extensive research, researchers have realized that it's not feasible to understand Students' problems/issues through surveys, interviews, etc. Manually extracting knowledge from such huge dataset is also a very difficult task.

Xin Chen et al. proposed a research work on "Mining Social Media Data for Understanding Students' Learning Experiences" [2]. In this work, the dataset is taken from social networking website: "Twitter" to understand students' problems. The methods used are Classification, Inductive Content Analysis; to mine social media data which is available for free using Twitter API. The researchers have classified problems after manually going through the dataset. The Training Dataset is taken from Twitter and the Testing Dataset is taken from Purdue University. The Naive Bayes Classifier is used to classify the dataset. 6 categories are made where if a record doesn't belong to a classified category, it goes to the "other" category. There are certain limitations-1) Twitter has very less Global/Indian users as compared to Facebook and WhatsApp. So to a certain extent, this data is not very relevant.2) Very few users access Twitter; students can't express much, as there is a 140 char limit in a tweet.3) Mining social media data is of great importance as students are free to write what they want to; but at the same time, it is a platform to create anonymous identities. These unreal profiles create problems as no one can be sure whether the data is authentic or not. 4) No solution is given to students' problems.5) The "other" category is huge which could have been further divided as there were a lot more problems in the dataset which are not covered.

Cui Yuan proposed a research work on "Data Mining Techniques with its Application to the Dataset of Mental Health of College Students" [3]. In this work, the author has done a study on the mental health problems of students by the means of Data Mining technology. ESX, which is a Data mining tool, is used for performing analysis along with iData analyzer. The sample dataset is of 500 students' taken from Chengdu Medical College Science, China. The analysis is done on the basis of number of suicides, how students behave with their juniors, how good they feel in the campus; what are their areas of interest, etc. IDA Supervised learning is used to train the system for making predictions. The training is done on various parameters such as thinking of committing suicide, fear of speaking in public; and pressure of study. The result section shows how students' mental health is affected in the college campus. The analysis shows that majority of students feel their college life is boring and ordinary; which can be the reason for suicides as it affects their mental health. Only 14% of students believe their college life is interesting. The fresher is more likely to suffer mental health issues as compared to student of higher grade. The limitations are: 1) No solution is provided for the problems discussed. 2) There are different techniques in Data Mining such as Clustering/Classification; which are not used for proper comparative study.

Banumathi et al. proposed a research work on "A Novel Approach for Upgrading Indian Education by Using Data Mining Techniques" [4]. In this paper, the authors have analysed the problems students' face in their school life. Clustering (which is a data mining technique) is applied to the dataset of 10 students to classify them in different clusters as low, average, high, on the basis of their marks. This is a sample dataset created by the authors themselves. As k-means clustering requires initial fixed value of k; the authors have used UCAM clustering which removes this drawback by fixing the threshold value. The result section shows five clusters formed on the basis of students' marks. The analysis shows that a student needs extra care in one subject if he/she is scoring low marks in that particular subject. There are certain limitations: 1) Data of only 10 students is taken which is very less for performing mining. 2) Dataset is created by the authors and not taken from any real institute/school.

Bo Guo et al. proposed a research work on "Predicting Students Performance in Educational Data Mining" [5]. In this paper, the researchers have applied Data Mining techniques and deep learning approach to predict students' performance. A system called, "Students Performance Prediction Network" (SPPN)" is proposed to predict students' performance which is demonstrated to be highly accurate with large datasets. The system is trained using Supervised and Unsupervised Learning both. Graphical processing unit (GPU) is used for fast execution and training. Six layers Neural Network is used to implement deep learning algorithm. Dataset is taken from 100 junior high schools of Hubei province, China. The result section shows a comparative study of SPPN with other classifiers such as SVM (Support Vector Machine) and Naive Bayes classifier; where SPPN is shown to have better accuracy. The limitations are: 1) The system is trained but is not applied on any testing dataset to show how accurate the predictions are. 2) The system is very complex which makes it costly; also may not work with small dataset. 3) The system requires GPU's for processing which are very costly as compared to CPU's. 4) No proper solutions are recommended or given by the system.

Nyalleng Moorosi et al. proposed a research work on "Privacy in mining crime data from social media: A South African perspective" [6]. This paper discusses the privacy issues of users when the data is taken without their permission from Twitter and Facebook API's. The research is focused on the crime reports in South Africa where users use social media to inform authorities about the crimes happening around them. The concern is raised as to how much data must be taken from social media for mining, as the privacy of the users is breached or compromised. The result section shows the data analysis of user accounts and also mentions the laws in South Africa to protect users' personal information. Stricter laws are recommended against using user's personal data from social media. The limitation in this work is that it only shows the concerns of users' privacy and doesn't apply any methodology/workflow or provide any solution to protect the same.

3. METHODOLOGY

Classification [7] is a very popular data mining technique based on machine learning. It is a supervised based learning approach where the system is trained based on past decisions and predictions are made for the future; as to which class a particular object will belong to. In this work, Classification technique is used to classify Students' problems. Students' problems are classified into certain classes and a system is build which can predict the target class for Students' problems and solutions can be suggested for the same. These classes are made after performing Inductive content analysis (i.e., manually going through the data) on CMS records of past 3 years. Classification makes decision based on certain features and uses a training sample set to predict class for the new object. Training sample set is the known data where decisions are already made for certain objects; and based on this past data; the system predicts the class for the new object. In this work, Naive Bayes classifier (which is based on Bayes theory) is used to classify students' problems.

Naive Bayes classifier [8] makes use of apriori probabilities which is already known before. A coin has 50% chance of coming heads; which makes apriori probability as 1/2 for heads. For example, suppose there are two classes a and b. Then p(a)&p(b) are the apriori probability for class a & b respectively.

To determine class for an object, a new feature say "x" is used to determine the deciding feature of an object for the class. This makes the decisions more valid &logical. This is also called as observation; which gives a certain feature to decide the target class for the object & it is represented by p(x/a).

The probability density function of x, taken from the known objects of class a is shown through p(x/b). This is probability density function of x, taken from the known objects of class b.

These are called class conditional probability density function.For an unknown object, feature "x" can be measured and then it can be decided which class the object belongs to.

Next, the observation value of "x" is known, calculate the following

p(a/x) = probability that an object belongs to category p(a) given feature "x"

p(b/x) = probability that an object belongs to category p(b) given feature "x"

These two features are combined to make decisions more logical. This will be called joint probability. Two classes are taken as a_i where i=1 and 2.

$p(a_i,x) = p(a_i/x).\ p(x)$ or
$p(a_i,x) = p(x/a_i).\ p(a_i)$
$p(a_i/x).\ p(x) = p(x/a_i).\ p(a_i)$
$p(a_i/x) = p(x/a_i).\ p(a_i) / p(x)$

where $p(a_i/x)$ is the probability that an object belongs to category $p(a_i)$ given feature "x".

p(x) is the apriori probability of feature x.

To determine class for an object, following equation is used:

$$p(a_i/x) = \frac{p(x/a_i).\ p(a_i)}{p(x)}$$

where $p(x) = \sum_{i=1}^{2} p(x/a_i).\ p(a_i)$

$p(x/a_i)$ is the class conditional probability and $p(a_i)$ is the apriori probability.

Using these two, $p(a_i/x)$ can be calculated which is called as posterior probability.

p(a/x) > p(b/x) => a

p(a/x) < p(b/x) => b

This is the decision rule for determining/predicting classes for the objects.Naive Bayes classifier is used in RapidMiner using "Naive Bayes" operator. The Naive Bayes classifier takes the dataset for training and the system is trained to correctly predict the class of students' problems.

In this work, the software platform used is RapidMiner. RapidMiner[9] is one of the most popular software tool used in Data Mining. It is an open source platform which empowers all organizations to put data science behind every decision. RapidMiner is used for both research work and real-world data mining tasks. It has become immensely popular among the researchers because of its ease of use; as it eliminates the need to write a code. Data Mining has grown very rapidly in past few years and as Data Mining is very useful for Businesses; RapidMiner provides excellent set of tools for predictive analysis. There are various operators which are used to perform tasks in RapidMiner. These operators can be brought to the main window by drag and drop operations; which make RapidMiner easy to use and also provides fast results. To perform training in the system (so that students' problems are categorized correctly), the data is stored in an excel sheet; which is brought inside RapidMiner using a "read excel" operator. This operator reads the excel file which is further used to train the system for performing predictive analysis on the dataset.

Next operator used is "Process Documents from Data". This operator is used to perform Pre-processing which is a necessary step to remove errors. It removes inconsistency and noise from the data. It first "tokenizes" the data and the text is broken into small tokens using "Tokenize" operator. These tokens are transformed into one common format to remove inconsistencies; and the data results in anappropriate form for mining. This is done using "Case Transform" operator. The stopwords which are common English language words such as "a", "an", "the" are removed as they are irrelevant for the analysis. This is done using "Filter Stopwords" operator. "Stemming" is used after this to reduce words to their root/base. E.g.there are three words as "car", "cars", "car's" will be reduced to 'car'. As car is the root word and they all carry similar meaning. Once the pre-processing is done, the data is classified intocertain classes which include all the problems Students face in their day to day college life. Validation operator is used to check the accuracy of the system.Store operator is used to store the wordlist and training data so that it can be retrieved later for testing dataset.

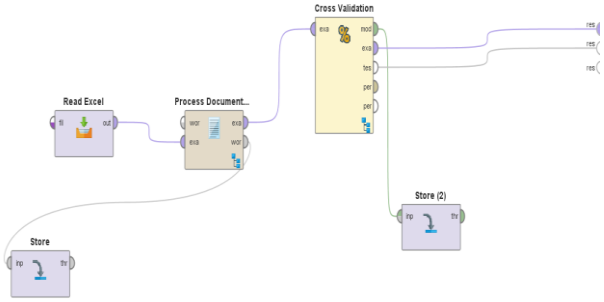

Fig2:RapidMiner main process

After performing Inductive content analysis on the huge dataset (manually going through the dataset) certain classes are formed such as BUS problem, LIBRARY problem, WASHROOM problem; which are the most common issues among Engineering Students' in the Gyan Ganga Institute. Several students responded as they don't have any problem so a new category was made as "NO_PROBLEM".

3.1 Training Dataset

The system is trained once the classes are made. 1000 records of students' problems collected after months of research is taken in an excel sheet to train the system. The system is trained as if the problem is related to "bus", it goes to the BUS category. Majority of students faced issues with BUS. Same goes for other categories such as LIBRARY and WIFI. These are some common problems which students face in their day to day college life.

Table 1. Students' Problems Training Dataset

S.No.	Problems	Classification
1	No problem till now.	NO_PROBLEM
2	There are usually no seats available in the bus.	BUS
3	Problem of the bad conditions of the college buses.	BUS
4	My bus has got defective seats	BUS
5	Not getting to issue the books from the library.	LIBRARY
6	No problem. All ok.	NO_PROBLEM
7	Washrooms are very dirty	WASHROOM
8	Books not available in the library sir. Out of stock.	LIBRARY
9	Break time must be increased.	BREAK_TIME
10	WASHROOMS aren't clean. They stink even in the morning.	WASHROOM
11	Library is of no use. Books are never available.	LIBRARY
12	I Have No Problem With Ggits.	NO_PROBLEM
13	very over crowded bus	BUS
14	wifi should be provided	WIFI
15	wifi was promised in college	WIFI

After training the system, testing Dataset is taken.

3.2 Testing Dataset

Testing Dataset contains records of 51 students. There are 51 text files for 51 students; one text file for each. The problems which students face are stored in these text files. This dataset is used for analysing the problems of students which they face in their college life; and also suggest the solutions for the same. The testing dataset helpsus to check the accuracy of the system; i.e. how accurate the system is making predictions

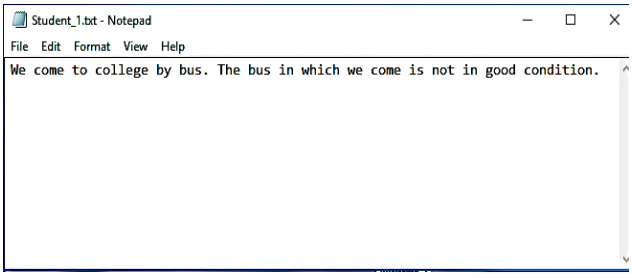

Fig3:Testing dataset.

3.3 Content Management System

In this work, Students' problems are taken from CMS. A content management system [10] (CMS) is a software application that is used to create and manage digital content. CMS is used by institutes to store students' information in a structured and tabular form. CMS allows Students to post their problems which is stored in an excel sheet.

Table 2. Example of CMS Dataset

TG Name-2	SatyendraSonare	
1	0206CS 161038	ANUSHREE AGRAWAL	Books are never available in library also the books are of very old edition in our library sir.....
2	0206CS 161034	ANSHU SONI	Bus Is always late and no place to sit in bus
3	0206CS 161041	APOORVA DIDHATE	Washrooms are not good.
4	0206CS 161033	ANSHIKA PATEL	There are no proper cooling fans or coolers in the workshop....so we will face a problem when we work in the workshop..
5	0206CS 161032	ANKIT PARIHAR	I have no problem with this institution and faculty.

The above table shows how data is stored in CMS. Students post their problems and they are stored in the system. It comes with a user login and password to prevent unauthorized access. Data of 3 years is taken from CMS comprising of Students' problems. The reason for choosing CMS data is that the data is authentic and real. Every record includes the details of students such as name and roll no. This makes the data authentic; unlike social media, where mostly the data is fake and anonymous. Social media allows users to create a fake identity and the real identities of users are hidden. This makes the social media analysis unreliable. CMS data comprises of problems taken directly from students' which makes it real and reliable too.

4. RESULT AND ANALYSIS

The records of 51 students are used to check the accuracy of the system. RapidMiner is used in this work, where we create a new process and run it to get the desired output. All the operators brought in the main window are connected to each other and with the input/output ports. The figure below shows the result once the process is executed.

Row No.	label	prediction(...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	metadata_file	metadata_d...	metadata_p...
1	Student_problem	BUS	1	0	0	0	0	0	1.txt	Apr 20, 2017 ...	C:\Users\Pra...
2	Student_problem	BUS	1	0	0	0	0	0	10.txt	Apr 20, 2017 ...	C:\Users\Pra...
3	Student_problem	NO_PROBL...	0	1	0	0	0	0	11.txt	Apr 20, 2017 ...	C:\Users\Pra...
4	Student_problem	NO_PROBL...	0	1	0	0	0	0	12.txt	Apr 20, 2017 ...	C:\Users\Pra...
5	Student_problem	BUS	1	0	0	0	0	0	13.txt	Apr 20, 2017 ...	C:\Users\Pra...
6	Student_problem	BUS	1	0	0	0	0	0	14.txt	Apr 20, 2017 ...	C:\Users\Pra...
7	Student_problem	BUS	1	0	0	0	0	0	15.txt	Apr 20, 2017 ...	C:\Users\Pra...
8	Student_problem	BUS	1	0	0	0	0	0	16.txt	Apr 20, 2017 ...	C:\Users\Pra...
9	Student_problem	LIBRARY	0	0	1	0	0	0	17.txt	Apr 20, 2017 ...	C:\Users\Pra...
10	Student_problem	NO_PROBL...	0	1	0	0	0	0	18.txt	Apr 20, 2017 ...	C:\Users\Pra...
11	Student_problem	LIBRARY	0	0	1	0	0	0	19.txt	Apr 20, 2017 ...	C:\Users\Pra...
12	Student_problem	BUS	1	0	0	0	0	0	2.txt	Apr 20, 2017 ...	C:\Users\Pra...
13	Student_problem	LIBRARY	0	0	1	0	0	0	20.txt	Apr 20, 2017 ...	C:\Users\Pra...
14	Student_problem	NO_PROBL...	0	1	0	0	0	0	21.txt	Apr 20, 2017 ...	C:\Users\Pra...
15	Student_problem	BUS	1	0	0	0	0	0	22.txt	Apr 20, 2017 ...	C:\Users\Pra...
16	Student_problem	NO_PROBL...	0	1	0	0	0	0	23.txt	Apr 20, 2017 ...	C:\Users\Pra...
17	Student_problem	NO_PROBL...	0	1	0	0	0	0	24.txt	Apr 20, 2017 ...	C:\Users\Pra...
18	Student_problem	BUS	1	0	0	0	0	0	25.txt	Apr 20, 2017 ...	C:\Users\Pra...
19	Student_problem	BUS	1	0	0	0	0	0	26.txt	Apr 20, 2017 ...	C:\Users\Pra...
20	Student_problem	LIBRARY	0	0	1	0	0	0	27.txt	Apr 20, 2017 ...	C:\Users\Pra...

Fig 4:Result of Naive Bayes classifier,predicting class for Students' problems.

Here we can see, the system has correctly predicted class for Student's problems. This helps us to analyse what problems students' face and how it can be resolved. We can see that majority of students' have issues with BUS and LIBRARY. The students' who are not facing any such problems; a new categoryhas been made for them as NO_PROBLEM. Out of 51 students, 14 students have problem with BUS, 11 with Library, 3 with BREAK_TIME being not enough, 6 with WASHROOMS not clean, 14 have no problems and 3 have issues with WIFI. This is automatically predicted by the system and no manual effort is required.

The analysis shows that there are some common problems which students' face and they must be resolved for their better future. There are several problems such as BUS problem, WIFI, Library, Break Time, and Washroom problem.

From the results of 51 students' records, the analysis shows that 28% of student's have no issues with the Institute; 28% students' face issues with BUS regularly, 22% have issues with Library, 6% with Break Time and WIFI; and 12% have issues with Washrooms.

In this paper, Naive Bayes classifier has been used; but there are several other classification algorithms which can also be applied to the dataset and a comparative study can be made on the basis of their performance and accuracy in making predictions. Decision Tree, K-NN, Random Forest, Generalized linear model are some of the most popular classifiers.

The table below shows the comparative study for various classification techniques along with their performance parameters such as accuracy, classification error rate, Kappa, precision and recall.

Table 3. Performance Table

Classifier	Naive Bayes	K-NN	Decision Tree	Random Forest(modernize)	Generalized Linear Model
Accuracy	99.10%	99.70%	**99.90%**	84.45%	97.61%
Classification error	0.90%	0.30%	**0.10%**	1.39%	2.39%
Kappa	0.989	0.996	**0.999**	0.854	0.970
Weighted mean recall	99.27%	99.62%	**99.79%**	87.10%	96.41%
Weighted mean precision	98.76%	99.73%	**99.85%**	94.2%	98.50%

Decision tree has given best performance in terms of high accuracy and low classification error rate. K-NN and Naive Bayes has also achieved high accuracy. Random forest which uses combination of decision trees with random function is giving little high error rate asit requires more features for classifying data. This may not be appropriate to classify students' problems from the CMS dataset as the emphasis is on problems rather than different features. Generalized linear model has also achieved high accuracy. Classifiers which can effectively handle polynomial data have given high accuracy results.

The graph below shows the analysis of performance parameters for different classifier-

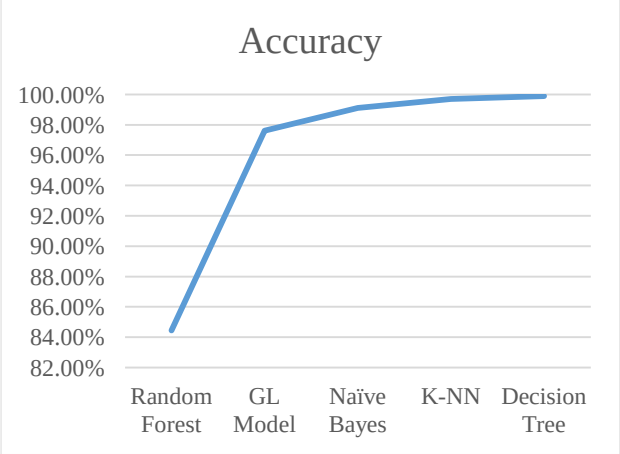

Fig 5: Accuracy Graph

Accuracy is how correct the predictions are. The above graph shows that decision tree has given highest accuracy with 99.90%. Other classifiers have also achieved high accuracy.

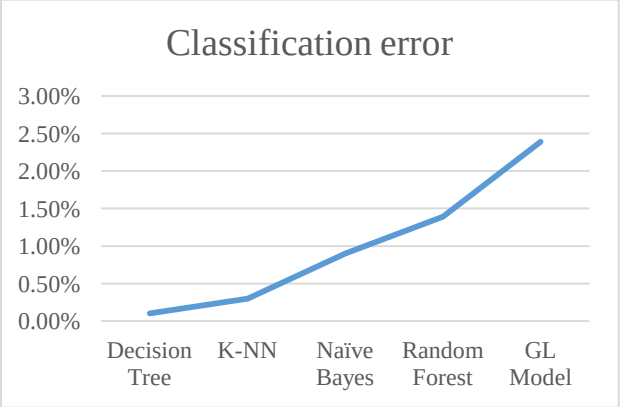

Fig 6: Classification error graph

Classification error rate defines the percentage of incorrect predictions. The graph shows the decision tree with least classification error rate of 0.10, as it achieved high accuracy. Other classifiers have also less error rate of less than 1%; where only random forest has given little high error rate as it achieved low accuracy.

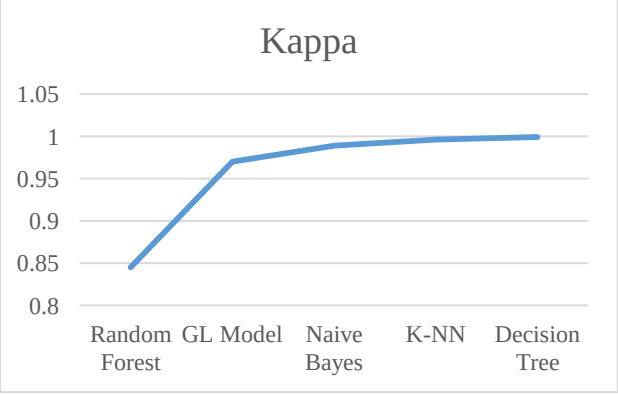

Fig 7: Kappa Graph

Kappa statistics is a measure to check how correct the predictions are made. It is considered more robust as it also considers correct predictions made by 'chance'. It compares the accuracy of the system with a random system. The analysis shows that decision tree has gained highest Kappa value of 0.999 which is very close to 1. Other classifiers have also achieved high Kappa values.

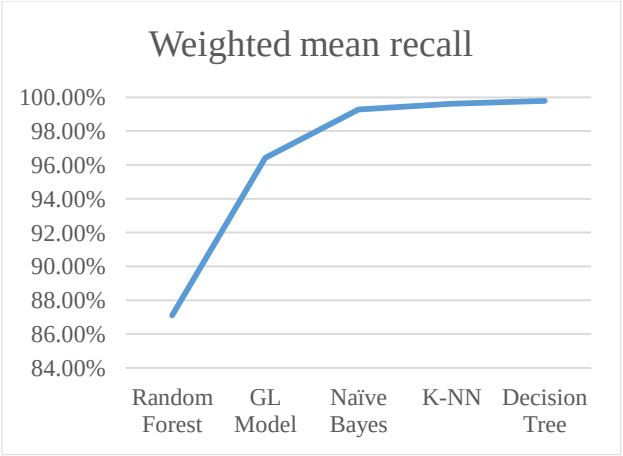

Fig 8: Weighted Mean Recall Graph

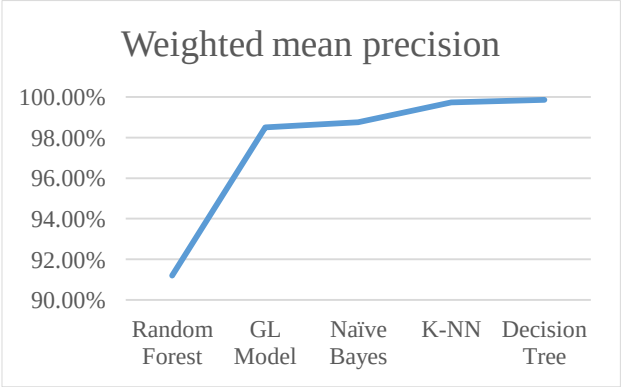

Fig 9: Weighted Mean Precision Graph

Recall and Precision are based on how relevant the result is. The weighted mean for each individual class is calculated. The analysis shows that decision tree has achieved high mean value of 99%. Other classifiers have also obtained high recall and precision values.

5. CONCLUSION

In this paper, the analysis is performed on the problems of Students which they face in their engineering/college life. It is important to get rid of these problems so that they can have a better career; as it's been said that "Today's Students Are Tomorrow's Nation Builders". It was noticed that there are some common problems such as Library and WIFI issues where books and e-books are not provided to students on time which affects their results. There are several other issues which must be addressed by the concerned authorities.

In this work, a popular tool called RapidMiner, which is a Data Mining software; is used to analyse huge set of data. Classification is a popular data mining technique (which classifies items into different classes) is applied to the CMS data to analyse the Students' problems. Naive Bayes Classifier is used to classify Students' problems into different classes and to train the system using 1000 records to correctly

predict the class for students' problems. The accuracy of the system is tested using 51 students' records which can be further increased.

In future, this research work can be further extended to more Students' records from other fields than engineering such as medical or from other streams. The research work can also be extended to students from school or university. In this work, Naive Bayes classifier has been used; but there are other Data Mining classification techniques such as Support Vector Machine, Artificial Neural Network, and ID3; which can also be applied to the dataset in future.

6. REFERENCES

[1] 80% of engineers in India unemployable http://www.thehindubusinessline.com/economy/over-80-engineering-graduates-in-indiaunemployablestudy/article8147656.ece

[2] M. V. K. M. Xin Chen, "Mining Social Media Data for Understanding Students' Learning Experiences," IEEE Transactions on Learning Technologies, vol. 7, no. 3, pp. 246 - 259, 06 January 2014.

[3] C. Yuan, "Data mining techniques with its application to the dataset of mental health of college students," in Advanced Research and Technology in Industry Applications (WARTIA), 2014 IEEE , Ottawa, ON, Canada, 29-30 Sept. 2014.

[4] A. P. A. Banumathi, "A novel approach for upgrading Indian education by using data mining techniques," in Technology Enhanced Education (ICTEE), 2012 IEEE International Conference, Kerala, India, 01 June 2012.

[5] R. Z. X. S. Y. Bo Guo, "Predicting Students Performance in Educational Data Mining," in Educational Technology (ISET), 2015 International Symposium , Wuhan, China, 27-29 July 2015.

[6] N. M. Nyalleng Moorosi, "Privacy in mining crime data from social Media: A South African perspective," in Information Security and Cyber Forensics (InfoSec), 2015 Second International Conference , Cape Town, South Africa, 15-17 Nov. 2015.

[7] Data Mining Concepts. Classification: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#DMCON004

[8] Naive Bayes Archives - Analytics Vidhya https://www.analyticsvidhya.com/blog/2015/09/naivebayes-explained

[9] RapidMiner www.rapidminer.com

[10] Content Management System (CMS) http://searchcontentmanagement.techtarget.com/definition/content-management-system-CMS