VOL. 15, NO. 24, December 2020 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

CAREER PREDICTION OF IT APPLICANTS BY MINING EDUCATIONAL AND ALUMNI DATA

Nadia Ghaffar[1], Abdul Mateen[1], Saeed Ullah[1], Rubina Adnan[2], Muhammad Javed[3] and Kashif Rizwan[1]
[1]Department of Computer Science, Federal Urdu University of Arts, Science and Technology, Islamabad, Pakistan
[2]Department of Computer Science, COMSATS Institute of Information Technology, Islamabad, Pakistan
[3]Department of Computer Science, University of Peshawar, Peshawar, Pakistan
E-Mail: nadia@biit.edu.pk

ABSTRACT

Technical education is becoming more and more career-oriented. Therefore, most of the researchers and studies have contributed their part by predicting students' career. On the other hand, students' career after graduation has become a major factor in building the reputation of the institutes. Predicting their degrees and future direction beforehand can help to take timely action by institutions. Data mining is a process of finding patterns and correlations within large datasets to predict outcomes. When the data mining process is used to study the learning characteristics, behaviour and performance of the student, it is called Educational Data Mining (EDM). This study implements a supervised machine learning technique to predict career of IT applicants whether IT or Non-IT by analyzing the data of those students' who had graduated. The dataset is collected from the Course Management System (CMS) of Barani Institute of Information Technology (BIIT) PMAS, AAUR Rawalpindi and mapped with the alumni dataset which is collected from BIIT graduated students using Google forms to have one dataset for experimental use. The major concerns of this work is to develop an efficient student recommendation system for predicting career of BSCS/BSIT applicants' whether IT or Non-IT at the time of admission and enhance the performance of learning model by re-sampling of dataset with SMOTE algorithm. This work also highlights the impact of selection of most relevant set of features for accurate results. The system efficiency has been tested upon the data of 3327 students. The total number of 11 attributes have been considered for career prediction i.e. gender, current degree program, demography details, SSC board, SSC subject, SSC grade, HSSC board, HSSC subject, HSSC grade, final degree CGPA and alumni job. To improve accuracy of learning models, re-sampling of dataset technique is sused to handle class imbalance problem by applying SMOTE algorithm. In this research, Random Forest, C4.5(J48) and Support Vector Machine (SVM) models are used to determine the best predictive model for supervised machine learning. 10-fold cross validation and standard performance evaluation measures such as: accuracy, precision, recall, F-Measure are used to evaluate the classifiers results. During the experiments Random Forest obtained 96.86% accuracy measure which is better than all of the learning algorithms, i.e., C4.5(J48) 96.55% and SVM 96.15% after handling class imbalance problem with SMOTE. This recommendation system may be more helpful by predicting career of IT department applicants' at the time of enrollment.

Keywords: education, alumni, career prediction, accuracy, data mining.

1. INTRODUCTION

Nowadays, higher education has become more and more career-oriented. The main focus of students is a fast start into the job market by completing professional degrees as soon as possible and start earning for living. This trend affects all higher educational institutes to organize their curriculum to meet the requirement of a different range of students. A large number of students are getting enrolled in higher education with different mindsets. The interest of a student in a specific field is not enough for a quality career-oriented education. This is the responsibility of the higher educational institutes to provide proper guidance to their students for selecting appropriate education field timely (Arafath *et al.* 2018). Higher education must play a vital role in strengthening nation's economy like other bussiness industries and also support rest of the industries by providing skilled and trained workforce. The major concern for these institutes is to provide high quality education to the students, (Agarwal *et al.* 2019), minimize the rate of students' dropout, increase students' success rate and fill gaps in counseling to students in subject selection.

In the last few decades, the number of computer science educational institutions has grown rapidly. A large number of IT graduates are moving speedily to IT industry to earn a job. Therefore, the employment condition of IT graduates is becoming more serious. It is the responsibility of IT institutes to prepare their graduates according to the demands of industry with high quality education and IT skills. On the other hand, during the graduation program, it is difficult for IT institutes to prepare those students for job market who are not suitable for IT field. Those students which are not suitable for IT may not be. It is important for these institutes to suggest their students' if they are capable enough for IT field. Career counseling at right time can be beneficial for both institute and students. Institutes can minimize the failure and dropouts rate by proper counseling on initial stage. Student can save their money, efforts and time by choosing appropriate education field for their career. This paper aims to apply different machine learning algorithms to train model to predict IT applicants' career and to reveal best attributes affecting appropriateness of applicant in IT field.

Sometimes students are unable to decide about choosing the IT field after school for higher education.

Lack of computer knowledge in the past, becomes crucial for students to select IT for graduation. Can admission details of students' helpful in career selection? Students' demographic details would not be only for admission purpose but also predict whether they be suggested to adopt IT field or not? If the student has had poor grades in the past education or no computer background, then what would be the better option for next? Whether he be suggested to choose IT field or not? On the other hand, students with poor educational excellence history are causing to increase the ratio of academic failure and university dropouts. In education sector, this research provides useful knowledge for students' betterment. Administration can make more suitable decisions for students. Students' Educational history and their demography background imply great impact on their career. The main objectives of this work are to:

- Identify the required feature which comprehensively defines the mentioned problem of the work.
- Examine the best class balancing technique for better training of classifier.
- Explore the classification algorithm which performs better over the standard classification parameters such as: Accuracy, Precision, Recall, F-Measure, ROC Curve.

The dataset of computer science students has been obtained from the Course Management System (CMS) of Barani Institute of Information Technology PMAS, AAUR Rawalpindi, Pakistan. There was no student recommendation system that helps the administration to guide applicants. The aim of this research is to help administration to have precise understanding about their applicants by studying different academic, demographic and personal factors of the students and predicting estimated career whether IT or Non-IT at the time of admission. Ability of predicting students' career using the proposed techniques basically helped authorities of educational institutes in maintaining a healthy collaboration with IT industry by serving proper trained and skilled professionals to the industry. This can also help to institutes to make more affective policies for improvement of week students so that they can accomplish their degrees with good grades and reduce the ratio of failures and dropouts. This research will also enables to ensure proper counseling to those applicants who are unaware of their career.

The rest of the paper is organized as follows: Section 2 introduces some similar work and literature survey on the prediction related with educational data mining. Section 3 contains the proposed data mining process. Section 4 describes the implementation of data mining task and the discussion about the final result and analysis. Conclusion and future work are presented in Section 5.

2. Literature Review

Data mining is a process of finding patterns and correlations within large datasets to predict outcomes. Data mining is also known as Knowledge Discovery in Data (KDD). When data mining process is used to study the learning characteristics, behaviour and performance of the student, it is called Educational Data Mining (EDM). There are several useful KDD tools for extracting knowledge and this knowledge can be used to increase the quality of education (Silva et al. 2017).

Unsupervised learning is used to determine the hidden formats in data which is unlabeled. The most common unsupervised machine learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. Machine Learning (ML) is a data mining technique and used to train computer on a training dataset, which enables the machine how to make sense of data and then how to make predictions of new dataset. These algorithms check the input data and provide outputs accordingly. An optimized result will allow an algorithm to determine the correct system for unseen data (Hussain et al. 2018). Researchers have proposed various techniques for mining data so that required information can be retrieved from large dataset. Like, Association rule mining is a well-reputed and most researched method for extracting and discovering the hidden and interesting relation between various variables in huge databases. It identifies strong rules found in big databases by different interesting measures (Hussain et al. 2018). Classification is a very helping technique to predict values accurately (Hussain et al. 2018). Classification problems can be solved using techniques, called classifiers. Educational Data Mining (EDM) uses many classifiers like Decision Trees, Naive Bayes, C4.5(J48), KNN, Neural Networks, etc. Clustering or cluster analysis is the technique of combining multiple set of values in such a way that similar values become closer to each other. This set of similar values combined together is termed as a cluster.

Educational Data Mining (EDM) is most promising and emerging field nowadays. This field used to extract useful knowledge from educational data and use extracted information for the betterment of students to improve academic performance, student and domain modeling, analysis and visualization of educational data, reduction in academic failure rate and dropouts. The main focus of Educational Data Mining (EDM) is to focus on extracting, evaluating variables and identifying variables connected to the process of learning (Silva et al. 2017). An efficient model was constructed to predict undergraduate students' academic problems to enhance their academic performance. For this experimental research, undergraduate students' dataset, collected from a university in Beijing, China. The main goal of this research is to predict employment after graduation by applying academic variables (Yanchao et al. 2018). Educational Data Mining (EDM) helps students to improve their performance by indicating academic issues on initial stage (Widyahastuti et al. 2017). Researchers aimed to predict performance of students for final examination and improvement of students' achievement. Multilayer Perception (MLP) and Linear Regression algorithms are used to predict final results of a student.

Educational Data Mining is considered a more promising field and (Sivasakthi et al. 2017) predicted performance of students' on early stages of degree, to reduce ratio of failure and dropout students. Researchers collected BCA students' data from college of Madras. Demographic history, grades in programming in higher secondary level, grades in programming by test with 60 questions. Researchers applied multiple classification algorithms to compare the most accurate and best performance of classification algorithm between them. Multilayer Perception (MPL), Naive Bayes (NB), C4.5(J48), SMO, Reduce Error Pruning (REP) Tree classification algorithm were used for comparison and prediction. (Arafath et al. 2018), predict the career of CS students by extracting useful information from alumni data. Some important parameters considered like professional skills, interpersonal skills, and academic records for accurate prediction. (Takci et al. 2017) aimed to help students to choose the most appropriate field which leads to low academic failure. This study focuses the relationship between student ability and academic achievement. A placement prediction model (Tarekegn et al. 2016) was implemented to place students into different departments according to their choice. Researchers took students' data from Gondar University. 1496 attributes of placed students were in date in MS Excel sheet. A comparative study (Nie et al. 2018) conducted to design a data-driven framework to predict career on the basis of students' behavior during graduation. 4246 university students' data conducted to train and test this model. 5-fold cross validation applied on first three years' semester.

Table-1. List of related studies.

Author	Dataset	Research Objective	Techniques Used	Results	Limitations
Piad *et al*. 2016	515 Instances 9 Variables	To predict IT employability after graduation.	NB, C4.5(J48), CART, Logistic R, CHAID	Logistic R 78.4%	Smaller dataset
Widyahastuti *et al*. 2017	50 Instances 2 Variables	Indicating academic issues on initial stage to predict final result of students.	Linear-Regression, MLP	MLP 84%	Low dimensionality problem.
Sivasakthi, 2017	300 Instances 7 Variables	Predict performance to reduce ratio of failure students.	MLP, NB, C4.5(J48), SMO, REP-Tree	MLP 93.2%	All applied classifiers work well with large dataset. Smaller dataset.
Arafath *et al*. 2018	506 Instances 9 Variables	Predict career of CS final year students by relevant skill parameters.	ID3,CART,RF,SVM,MLP	MLP, CART 95.2%	Smaller dataset.
Vyas *et al*. 2017	8 Variables	Predict good or bad performer before final year result	CART	CART	Single model approach.
Takci *et al*. 2017	210 Instances	To examine important factors for students' career selection	C4.5,SVM,NB,MLP	C4.5, MLP	Smaller dataset.
Rustia *et al*. 2018	446 Instances	Identify weak performance students'	C4.5,SVM,NN,NB, Logistic R	C4.5 73.1%	Smaller dataset.
Rojanavasu, 2019	106 Instances 9 Variables	To predict job of students' whether IT or not after graduation.	Apriori, ID3	ID3 73.5%	Class balancing algorithm not applied to improve accuracy of algorithm. Don't consider background attributes such as demography, school history etc. Smaller dataset.
Tarekegn *et al*. 2016	1499 Instances 22 Variables	Students' placement into deferent departments in University	NB,C4.5(J48),RF	RF 95.6%	Don't consider background attributes such as demography, school history etc.
Khongchai *et al*. 2016	13541 Instances 7 Variables	Predict future salary based on student's skills, academic history, and job training experience.	RF, ID3, C4.5,	RF 90%	Original recent salary should consider for training model for better prediction.
Nie *et al*. 2018	4246 Instances 4 Variables	Predict career on the basis of students' behavior during graduation.	Logistic-Regression, SVM,D.T, RF	RF 65%	Less attributes for training of the learning model.
Agarwal *et al*. 2019	306 Instances 9 Variables	Predict student placed for job in Campus or not	KNN, RF	KNN 88%	Class balancing algorithm not applied to improve accuracy of algorithm.

Devasia *et al.* 2016	700 Instances 19 Variables	Predict performance to minimize the failure rate of students.	Logistic-Regression, NN, D.T, NB	NB 79%	Irrelevant features can decrease performance of model. Food habit, other habit, vehicle etc.
Saa, 2016	270 Instances 21 Variables	Examine students' performance for predicting final grades.	NB, ID3, C4.5, CART, CHAID	NB 36.4%	Naïve bayes not efficient with small size dataset.
Kumar *et al.* 2019	Not mentioned	Predict final grades of university students.	NB, ID3, C4.5, RF	NB 85%	Dependency among variables may cause low accuracy in Naïve Bayes.
Costa *et al.* 2017	262 Instances 16 Variables	Reduce failure students rate in Programming course	NB,C4.5(J48),NN, SVM	SVM 92%	Smaller dataset
Verma, 2018	658 Instances 13 Variables	To improve performance, help students' to choose efficient subjects during graduation	KNN, NB, NN, D.T, SVM, Logistic-Regression	SVM 88%	Less attributes for training of the learning model. Subject marks treated as independent observation.
Casuat *et al.* 2019	3000 Instances 9 Variables	To predict engineering students' employability	D.T, RF, SVM	SVM 91.2%	Don't consider background attributes such as demography, school history etc.
Dubey *et al.* 2019	195 Instances	Predict employability after high school whether hired or not hired	Logistic-Regression, SVM, RF, D.T, KNN	SVM 93%	Smaller dataset.
Gong *et al.* 2019	1215 Instances 13 Variables	To predict students' employability after graduation.	RF, Logistic-Regression, D.T, SVM	SVM 87.4%	Oversampling technique used for class balancing it has not been evaluated on other efficient class balancing techniques.

3. Career Prediction of IT Graduates by Mining Educational & Alumni Data

Machine learning classification algorithms are widely used to train models for prediction in Educational Data Mining (Silva *et al.* 2017). We performed some best machine learning algorithms on dataset in this research. To conduct experiments, we collected data of graduated students from the Course Management System (CMS) of Barani Institute of Information Technology (BIIT) and BIIT alumni data from Google Forms and merged both to have one dataset for experimental use. We have performed experiments over the data of BSCS and BSIT programs. After data collection, we cleaned and prepared data through preprocessing phase then performed different machine learning classification algorithms by using WEKA tool. SMOTE algorithm was used for rebalanced data and improve values of classifiers. The obtained results have been analyzed and discussed in the next chapter. Figure-1 describes the proposed methodology for this study.

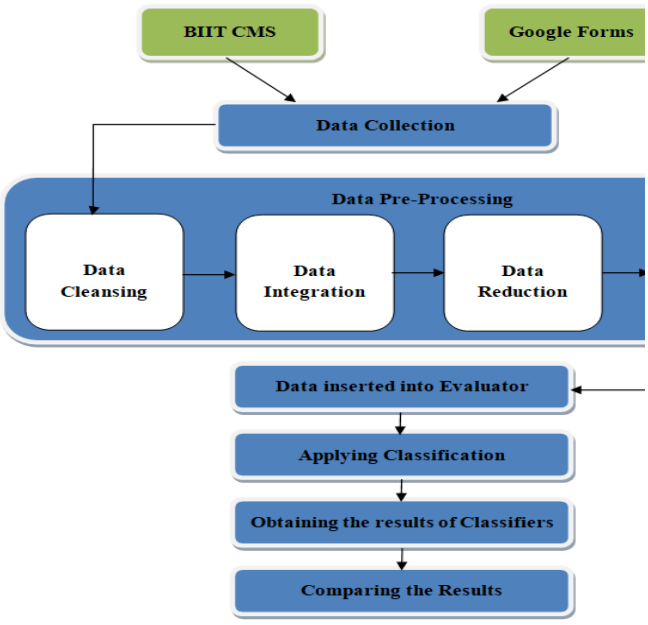

Figure-1. Proposed Approach.

3.1 Data Collection

This study examines the career of undergraduate students of BS (CS) and BS (IT) students of BIIT. Data collected from year 1998 to Fall 2015. The total instances were 7842. Then in next step, BIIT Alumni's data collected for career prediction. The university has no proper job data of its graduated students. Alumni's data collected through Google Forms with students' current job details. Alumni's data gathered from Google forms from graduated students of BIIT and mapped with BIIT CMS data of graduated students, to have one dataset for further use. Total 3327 instances collected after merging of both datasets.

3.2 Data Preprocessing (Transformation)

Data pre-processing is a data mining technique, used to convert raw data into understandable format. This phase used to identifying the missing values, falsie data and repeated information from the dataset. To handle missing and noisy data "Ignore the Tuples" technique is used for data cleansing (Shaleena *et al.* 2015); (Khongchai *et al.* 2016); (Kumar *et al.* 2019). Incomplete and missing records are deleted from the temporary table. Only those

records are saved in temporary table which has all required fields. An integration of two datasets such as students' degree dataset from BIIT CMS and alumni data of BIIT graduated students is also in the pipeline to have one dataset for further use (Shaleena *et al*. 2015); (Kumar *et al*. 2019); (Casuat *et al*. 2019). In data mining domain, for prediction of attributes there are two types of variables. Dependent and Independent variables; the dependent variable is resultant variable in which the researchers are concerned to monitor during the experiment whether it is affected or not. The dependent variable in our study was Alumni job status. The independent variables were collected from the categories of student gender, demographic information, pre-graduation grades, subject, and board and current academic program data. We have collected data from these categories because mostly researchers have collected data from these categories to conduct their experiments (Vyas *et al*. 2017); (Fernandes *et al*. 2019); (Shankhdhar *et al*. 2020). The independent attributes Matric and HSSC grade, gender, city and student CGPA. For example, CGPA, HSSC grade, gender, city, were used in previous research papers and play important role in prediction so that is why we have selected these attributes to check whether these attributes are well performing features or not. The description of selected attributes is shown in Table-2.

Table-2. List of Attributes.

Attribute	Description	Values and Transformation
Gender	Gender of students enrolled in the institute	M for Male, F for Female
Discipline	Discipline of students enrolled in the institute	BCS,BIT
SSC Board	SSC Board of enrolled students	SSC Board Name
HSSC Board	HSSC Board of enrolled students	HSSC Board Name
SSC Subject	SSC major subjects of enrolled students	Humanities, Science O-Level etc
HSSC Subject	HSSC major subjects of enrolled students	Humanities, Science, Commerce, Diploma, Computer etc
SSC Grade	SSC performance of enrolled students	A/B/C/D/E
HSSC Grade	HSSC performance of enrolled students	A/B/C/D/E
Address	Address of enrolled student where he belong	Full address string
Degree Completion CGPA	Degree completion CGPA of students enrolled in institute	1 for 3.5-4.00 CGPA 2 for 2.00-3.4 CGPA 3 for 2.5-2.9 CGPA 4 for below 2.5 CGPA
Alumni Job	Job of graduated student (output class)	YES for IT Field No for Non-IT Field

For the task of preprocessing, the collected data must be prepared in ARFF format which is compatible with the selected tool (Paid *et al*., 2016). The collected data were in raw form. We have prepared the entire data set in our required form.

3.3 Selection of Classifier

In order to achieve this objective, we performed a comprehensive literature review of the students' performance prediction, student career prediction techniques with respect to the classifiers used in those techniques for the purpose of the identifying the important factors behind the future performance and career prediction. During our literature review, we found out that the classifier that has proved to be the most useful with respect to the student career prediction are like Random Forest, C4.5(J48) and Support Vector machine (Nie *et al*. 2018); (Casuat *et al*. 2019); (Gong *et al*. 2019). Random Forest is a best predicting model algorithm among C4.5(J48), DT, Logistic Regression and Naïve Bayes (Tarekegn *et al*. 2016); (Khongchai *et al*. 2016); (Nie *et al*. 2018). Random Forest is a tree based machine learning algorithms with full of power of multiple decision trees. J48 is an extension of C4.5 decision tree in supervised machine learning, used to create binary trees for decisions. It transforms the problem of machine learning into a tree representation (Sivasakthi, 2017); (Tarekegn *et al*. 2016); (Costa *et al*. 2017). Support Vector Machine (SVM) is a supervised machine learning algorithm which is mostly used to solve classification problems in educational data mining (Costa *et al*. 2017); (Casuat *et al*. 2019).

4. Result Analysis and Evaluation

In order to achieve this particular research objective, we carried out a number of tasks such as conducting a comprehensive literature review, acquiring relevant datasets and merge them both for the purpose of the prediction of career and then performing a number of different experiments in order to predict the career of IT students at early stage of admission in higher education

institutions. For the experimental purpose, our selection of the classifiers and the test options has been based on the literature review. The majority of classification algorithms are designed to maximize the overall accuracy rate, which is independent of class distribution (Pristyanto *et al.* 2018). Data imbalance distribution problem occurs when the number of records is higher enough in one class then number of records in another class or classes. The performance of classifiers becomes less effective by imbalance class distribution. Handling of imbalanced class distribution is one of the important tasks to produce more accurate results and enhanced the efficiency of classifiers. There are many data balancing algorithms used in Educational Data Mining (EDM) like SMOTE, OSS, ROS and RUS etc. SMOTE algorithm is used widely to handling class imbalance distribution. This study will not only provide more accurate results by handling class imbalance distribution but also provide a comparison between learning algorithms with balanced and imbalanced class distribution. The study used three algorithms Random forest, C4.5(J48), SVM to build a prediction model for career of IT applicants. The analysis result shows that Random Forest algorithm performed well with balanced and imbalanced datasets and stood out as best predictor than the other applied for this work. The overall observations of models are represented in Figure-2.

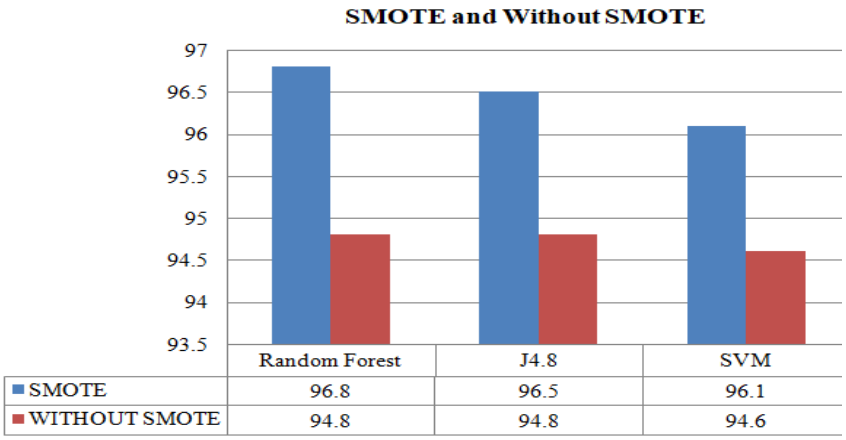

Figure-2. Accuracy comparison of Proposed Model with and without SMOTE.

Firstly, we use original dataset to build the models. The analysis result of different models with actual data showed the accuracy of Random Forest, C4.5(J48), SVM, 94.8%, 94.8% and 94.6% respectively. Then, we re-sample original dataset with SMOTE technique to balance the dataset and then we applied Random Forest, C4.5(J48), SVM classification algorithms to build the model. The accuracies of the models are 96.8%, 96.5% and 96.1% respectively. The accuracies are increased of classification algorithms about 2% after applying SMOTE on dataset. We noticed by the observations of each model that result Random Forest algorithm is most suitable and classical classifier for career prediction of IT applicants.

We have successfully applied Random Forest, C4.5(J48) and SVM machine learning models on actual and balanced class dataset after applying SMOTE and perform a comparative analysis between them and shown in Figure-2. It is clearly depicting that re-sampling data with SMOTE could produce more accuracy in all three standard classifiers. For efficient comparison of these machine learning models used in this study, some standard performance evaluation measures were selected to evaluate the performance the classifiers namely: Accuracy, Precision, Recall and F-Measure, ROC Curve. 10 fold cross validation approach was also applied for statistical analysis of the dataset. All the measures helped us to clearly picture out the behavior of classifiers. WEKA is one of the best tools in Educational Data Mining (EDM). It helped us to executing such kind of classifiers perfectly.

4.1 Performance Evaluation

Machine learning model's performance is measured on the basis of how many correct and incorrect predictions are made by the model on the testing dataset (A'rifian *et al.* 2019). We calculated accuracy of proposed machine learning models by accuracy metric, which is defined by the confusion matrix (Figure-3) as follows:

		Predicted Class	
		Predicted IT	Predicted Non-IT
Actual Class	Predicted IT	TP	FN
	Predicted Non-IT	FP	TN

Figure-3. Confusion Matrix for Binary Classes.

A. Accuracy

Accuracy shows the correctness of the model (Arafath *et al.* 2018). We performed K-Fold Cross Validation (K = 10) to find out the learning model accuracy on the data. Accuracy comparison of three classification algorithms, as shown in Figure-4 displays that model building with the best attributes and rebalanced data increase the accuracy of prediction for all classifiers. As we can observe that Random Forest (RF) is one of the best models to achieve highest prediction accuracy of 96.83%, C4.5(J48) is the second best classifier gives accuracy of 96.55%, SMO provides accuracy of 96.15% after applying SMOTE for re-sampling of dataset.

B. Precision

Precision is basically most important evaluation test which validates the quality of our output. If precision is high then it means false positive rate is low (Arafath *et al.* 2018), similarly if recall is high then it means false negative rate is low. If both precision and recall are high in output then it means the classifier will produce most accurate results. If precision is low then it means the classifier is predicting inaccurate results. But if only recall is low then it means most of the predicted set is correctly identified. We can easily understand by look closely into Figure-5 and we will be satisfied with results.

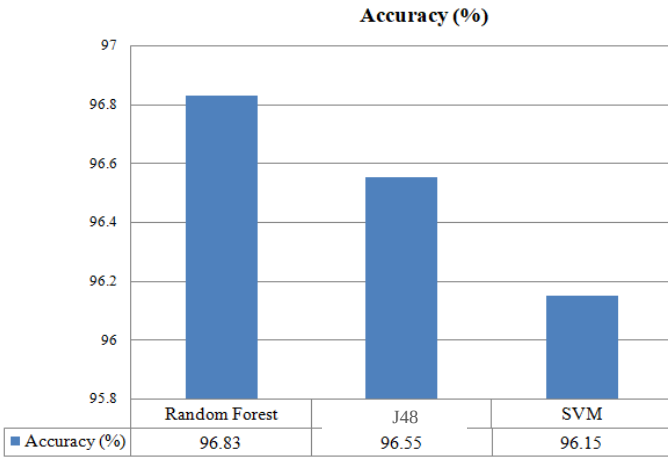

Figure-4. Accuracy Comparison of Proposed Models.

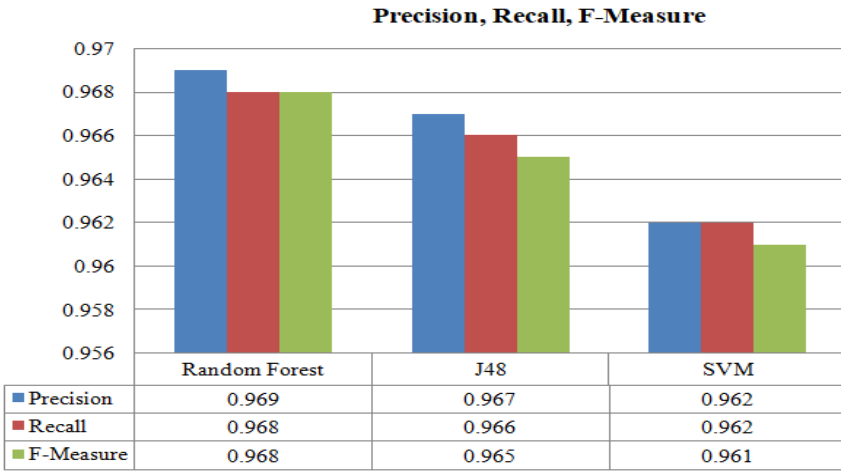

Figure-5. Comparison of Precision, Recall, F-Measure of Machine Learning Models.

C. Recall

Recall is the degree to determine the completeness of the model (Arafath *et al.* 2018). Best and worst case for recall are same as precision. From Figure-4, a consolidated comparison of recall for the three learning algorithms, Random Forest stood out as the best classifier after rebalanced data and best attributes used. There are multiple bars shown in graph where on the x - axis represents the percentage of the obtained performance for each model. In the y- axis, each section highlights analysis of a particular model which is experimented in this study. Recall is represented by red bar in this Figure-4. The best possible value for precision is 1 and worst is 0. From the chart, we can see that the difference between precision, recall and F-Measure in all three classifiers is minor but Random forest gave the highest recall score. C4.5(J48) produced less recall score than Random Forest and better than SVM.

D. F-Measure

F-Measure measures the test accuracy of model for binary classification (Arafath *et al.* 2018). It is the harmonic mean of precision and recall. In Figure-4, it is observed that Random Forest performed best in F-Measure among other classification algorithms after using best attributes for prediction and rebalanced data, shown in Figure-4.

E. ROC Curve

ROC curve or Receiver Operating Characteristic curve, as discussed in (Takci *et al.* 201)], this curve is basically a graphical representation of sensitivity (i.e. True

Positive Rate against fallout (i.e. False Positive Rate).Knowledge flow model for proposed work have been generated with the help of the knowledge flow application of WEKA tool. Three machine learning classifiers Random Forest, C4.5(J48), SVM passed to knowledge flow application to build ROC curve. Based on this knowledge flow model 'Model Performance Chart' has drawn and shown in Figure-6.

ROC curve of model performance chart showed excellent results in shown in Figure-6:

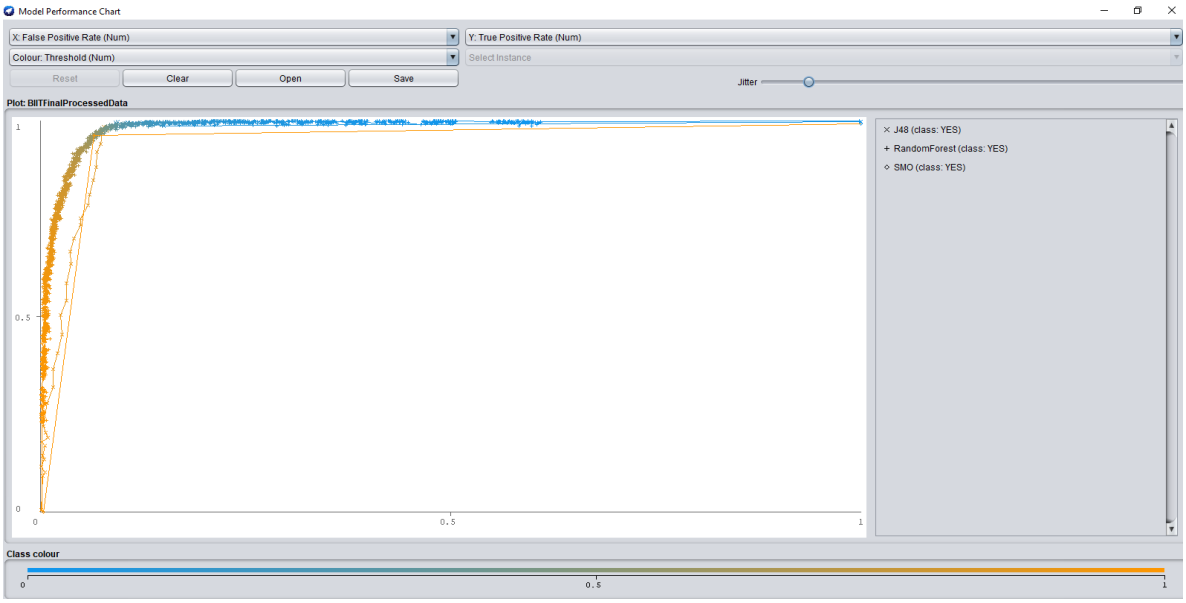

Figure-6. Model Performance Chart.

5. CONCLUSION and FUTURE WORK

Career counseling on right time can be beneficial for both institutes and students. Non suitable IT students may cause to maximize the academic failure and university dropouts' rate. Poor career counseling strongly impact to student's personal life by all means. As a result, the scientific community has been focused towards solving this problem by proposing data mining tools and techniques. From university perspective it is very costly and time consuming for non-suitable IT students to stay in the educational system. This issue also strongly impact on student's life.. It is critically observed that all career related predictions models were design to predict career after graduation for current degree students who were in middle or end of the degree. It is critically observed that all career related predictions models were design to predict career after graduation for current degree students who were in middle or end of the degree. The motivation behind this workis to create a prediction model to predict career of IT applicants' but on the time of enrollment using supervised machine learning approach. To conduct experiments, two datasets have been used for this study. One dataset of BSCS and BSIT students' has been collected from Course Management System (CMS) of Barani Institute of Information Technology (BIIT) and second dataset of BIIT alumni's' collected from Google Forms. For career classification, three best supervised machine learning algorithms were selected through comprehensive study of related literature. The proposed techniques applied on input data and the results reveal that Random Forest stood on top for predicting career with balanced and imbalanced dataset. Random Forest predicted more accurately whether applicants are suitable for IT or not, as compare with other classifiers used in this work.

This research was conducted in private institute of Pakistan. This can be extended in other public large scale universities in Pakistan. Secondly, this study was conducted on undergraduate computer science students. Further it can be extended for MS and PHD students and conducted on large datasets. Thirdly, this research was conducted on computer science discipline's dataset in future this research can be extended in other disciplines like management sciences and engineering.

REFERENCES

Piad K. C., Dumlao M., Ballera M. A. & Ambat S. C. 2016, July. Predicting IT employability using data mining techniques. In 2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC) (pp. 26-30). IEEE.

Silva C. & Fonseca J. 2017. Educational Data Mining: a literature review. Europe and MENA Cooperation Advances in Information and Communication Technologies. 87-94.

Hussain S., Atallah R., Kamsin A. & Hazarika J. 2018, April. Classification, clustering and association rule mining in educational datasets using data mining tools: A case study. In Computer Science On-line Conference (pp. 196-211). Springer, Cham.

Widyahastuti F. & Tjhin V. U. 2017, July. Predicting student's performance in final examination using linear regression and multilayer perceptron. In 2017 10th International Conference on Human System Interactions (HSI) (pp. 188-192). IEEE.

Sivasakthi M. 2017, November. Classification and prediction based data mining algorithms to predict students' introductory programming performance. In 2017 International Conference on Inventive Computing and Informatics (ICICI) (pp. 346-350). IEEE.

Arafath M. Y., Saifuzzaman M., Ahmed S. & Hossain S. A. 2018, September. Predicting career using data mining. In 2018 International Conference on Computing, Power and Communication Technologies (GUCON) (pp. 889-894). IEEE.

Vyas M. S. & Gulwani R. 2017, April. Predicting student's performance using CART approach in data science. In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA) (Vol. 1, pp. 58-61). IEEE.

Takci H., Gurkahraman K. & Yelkuvan A. F. 2017, September. Measurement of the appropriateness in career selection of the high school students by using data mining algorithms: A case study. In 2017 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 113-117). IEEE.

Shaleena K. P. & Paul S. 2015, March. Data mining techniques for predicting student performance. In 2015 IEEE international conference on engineering and technology (ICETECH) (pp. 1-3). IEEE.

Pristyanto Y., Pratama I. & Nugraha A. F. 2018, March. Data level approach for imbalanced class handling on educational data mining multiclass classification. In 2018 International Conference on Information and Communications Technology (ICOIACT) (pp. 310-314). IEEE.

Rustia R. A., Cruz M. M. A., Burac M. A. P. & Palaoag, T. D. 2018, February. Predicting Student's Board Examination Performance using Classification Algorithms. In Proceedings of the 2018 7th International Conference on Software and Computer Applications (pp. 233-237). ACM.

Rojanavasu P. 2019. Educational Data Analytics using Association Rule Mining and Classification. In 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, z Computer and Telecommunications Engineering (ECTI DAMT-NCON) (pp. 142-145). IEEE.

Tarekegn G. B. & Sreenivasarao V. 2016. Application of data mining techniques to predict student's placement in to Departments. International Journal of Research Studies in Computer Science and Engineering. 3(2): 10-14.

Khongchai P. & Songmuang P. 2016, January. Random forest for salary prediction system to improve students' motivation. In 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (pp. 637-642). IEEE.

Nie M., Yang L., Sun J., Su H., Xia H., Lian D. & Yan K. 2018. Advanced forecasting of career choices for college students based on campus big data. Frontiers of Computer Science. 12(3): 494-503.

Agarwal K., Maheshwari E., Roy C., Pandey M., & Rautray S. S. 2019. Analyzing Student Performance in Engineering Placement Using Data Mining. In Proceedings of International Conference on Computational Intelligence and Data Engineering (pp. 171-181). Springer, Singapore.

Fernandes E., Holanda M., Victorino M., Borges V., Carvalho R. & Van Erven G. 2019. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research. 94, 335-343.

Devasia T., Vinushree T. P. & Hegde V. 2016, March. Prediction of students' performance using Educational Data Mining. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE) (pp. 91-95). IEEE.

Saa A. A. 2016. Educational data mining & students' performance prediction. International Journal of Advanced Computer Science and Applications. 7(5): 212-220.

Kumar T. R., Vamsidhar T., Harika B., Kumar T. M. & Nissy R. 2019, February. Students Performance Prediction Using Data Mining Techniques. In 2019 International Conference on Intelligent Sustainable Systems (ICISS) (pp. 407-411). IEEE.

A'rifian N. I. N. B., Daud N. S. A. B. M., Romzi A. F. B. M. & Shahri N. H. N. B. M. 2019, November. A comparative Study on Graduates Employment in Malaysia by using Data Mining. In Journal of Physics: Conference Series (Vol. 1366, No. 1, p. 012120). IOP Publishing.

Costa E. B., Fonseca B., Santana M. A., de Araújo F. F. & Rego J. 2017. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior. 73, 247-256.

Verma R. 2018, November. Applying Predictive Analytics in Elective Course Recommender System while preserving Student Course Preferences. In 2018 IEEE 6th

International Conference on MOOCs, Innovation and Technology in Education (MITE) (pp. 52-59). IEEE.

Casuat C. D. & Festijo E. D. 2019, December. Predicting Students' Employability using Machine Learning Approach. In 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS) (pp. 1-5). IEEE.

Shankhdhar A., Agrawal A., Sharma D., Chaturvedi S. & Pushkarna M. 2020, February. Intelligent Decision Support System Using Decision Tree Method for Student Career. In 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC) (pp. 140-142). IEEE.

Dubey A. & Mani M. 2019, October. Using Machine Learning to Predict High School Student Employability–A Case Study. In 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 604-605). IEEE.

Gong H., Chen Y. & Li H. 2019, December. Modeling Graduates' High Quality Employment Based on Support Vector Machine. In 2019 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 2954-2958). IEEE.