I.J. Modern Education and Computer Science, **2019, 10, 14-24**
Published Online October 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2019.10.02

Identification of Trainees Enrollment Behavior and Course Selection Variables in Technical and Vocational Education Training (TVET) Program Using Education Data Mining

Rana Hammad Hassan
School of Systems and Technology, University of Management and Technology, Lahore – Pakistan
Email: rana_hammad@live.com / S2017288002@umt.edu.pk

Shahid Mahmood Awan
School of Systems and Technology, University of Management and Technology, Lahore – Pakistan
Email: shahid.awan@umt.edu.pk

Received: 01 August 2019; Accepted: 01 September 2019; Published: 08 October 2019

Abstract—Producing skilled workforce according to industry required skills is quite challenging. Knowledge of trainee's enrollment behavior and trainee's course selection variables can help to address this issue. Prior knowledge of both can help to plan and target right geographic locations and right audience to produce industry required skilled workforce. Globally Technical and Vocational Education Training (TVET) is used to provide skilled workforce for the industry. TVET is an educational stream which focus learning through more practicing with less theory knowledge.

In this article, we have analyzed TVET actual enrollment data of 2017 – 2018 session from a TVET training provider organization of Punjab, Pakistan. The purpose of this analysis is to understand trainee's enrollment behavior and course selection variables which plays an important role in TVET course selection by the trainees. This enrollment behavior and course selection variables can be used to monitor and control industry required and produced skilled TVET workforce. We developed a framework which contain series of steps to perform this analysis to extract knowledge. We used educational data mining techniques of association, clustering and classification to extract knowledge. The analysis reveals that central Punjab youth is getting more TVET education as compare to south and north Punjab, Pakistan. Similarly, trainee's 'age group', 'qualification', 'gender', 'religion' and 'marital status' are potential variables which can play important role in TVET course selection. By controlling these variables and integrating TVET training provider institutes, funding agencies and industry, we can smartly produce TVET skilled workforce required for industry nationally and internationally.

Index Terms—TVET Data Mining, Educational Data Mining, TVET Planning & forecasting, TVET Data Analytics

I. Introduction

The Technical and Vocational Education Training (TVET) is formal or informal education and training which enable trainee's to get employable skill to get a job or to start small business as an entrepreneur. TVET is an important education stream like school education and higher education but it is different from both with respect to it provide employable skills. TVET courses are normally based on 80 % for practical work and 20 % for theoretical knowledge [1]. The proportion of this equation is due to the fact that most labor work is done by hand and tools. This equation is entirely different from traditional education. TVET plays an important role in poverty alleviation and sustainable developing because it not only enhance pathways of career growth but also produces required skilled workforce for the industry [2]. Economic growth is based on multiple factors like human resource, natural resources, capital formation, technological development and social and political factors. The quality of human capital which is most important, is measured by the years of schooling, expertise and level of education [3]. Economic development cannot take place without the development of human resources therefore well-qualified professionals must be trained through TVET to raise competitiveness of companies, countries and regions [4].

Pakistan is the 5th largest young country, comprising 53% population aged between 15 and 33 years [5]. Providing education and jobs opportunities to youth is a

major challenge for Pakistan. It has been suggested for Pakistan's sustainable economy to maintain a good balance between TVET work force demand and trained labor supply [6] . This factor lead towards the need of proper TVET planning and supply of skilled workforce according to the industry requirements and job forecasting. Pakistan National TVET Policy 2018 [7] highlight that only 0.7 % of people aged between 15 - 24 years have access to TVET trainings in public institutes. A Pakistan TVET sector comparative analysis [8] estimates that there is a need to train 950,000 skilled workers annually whereas entire Pakistan TVET sector training capacity is 350,000 trainees per annum. These statistics demand that there is immense need to not only enhance TVET training opportunities but also use existing TVET training capacity smartly so that industry required skilled workforce can be produced in an economical and efficient way.

Data mining in education is known as Education Data Mining (EDM). EDM deals with the application of data mining tools, methods and techniques on the data related to education [9]. Data science is Knowledge Discovery in Databases (KDD) and Data Mining (DM) is its sub-topic. KDD is a process of finding knowledge in data emphasizing high level data mining methods [10]. KDD is big building block of fostering interfacing between different fields with the goal to identifying knowledge in data. A 10 years survey [11] of EDM concludes that EDM has extensively been used to explore in (i) student modeling (ii) decision support systems (iii) adaptive systems (iv) evaluations (v) scientific inquiry.

Research has successfully been focused in higher education and school education to explore knowledge in recent past. There is need to use EDM to explore knowledge in TVET education as well. EDM helps in discovering patterns and relations using machine learning, statistics and database systems. EDM techniques of association, clustering and classification are applied in this case study to explore knowledge of trainee's enrollment behavior and course selection variables which further can be used in TVET workforce planning, monitoring and forecasting.

Dataset in this case study has been used from actual enrollment data of TVET training provider of Punjab, Pakistan. An analytical framework based on Cross-Industry Standard Process for Data Mining (CRISP) is used for KDD. This framework include processes of business understanding, data preparation / understanding, input preprocessed data, apply EDM algorithms to extract knowledge of TVET course selection variables and enrollment behavior. Initially data was extracted from Academic System database and after preprocessing it was exported into single Microsoft Excel file for applying EDM techniques. Data visualizations techniques of EDM were used to explore insight of the dataset. After getting insight of dataset using data visualization, "age" numerical variable was converted into categorical variable of "age groups". The categorical variable "age group" was further used in applying EDM techniques. After preprocessing of data, EDM technique of association using frequent itemset and association rules algorithms are applied. Later, hierarchical clustering was applied to explore similar groups of the dataset. Finally, EDM classification techniques were used through Decision Tree, Random Forest, Neural Network and Naïve Bayes algorithms to find groups / classes of the dataset. EDM algorithm results have been shown using appropriate data visualizations techniques like graph, bar chart, scatter plot, boxplot, nomogram and Pythagorean. Based on data mining techniques applied and knowledge extracted enrollment behavior was learned that more graduates in Central Punjab are getting TVET education as compare to north and south Punjab. Similarly, trainee's 'age group', 'qualification', 'gender', 'religion' and 'marital status' are potential variables which play important role in TVET course selection.

This paper is organized as follows: Literate Review section include details about related work. The Materials & Method section brief about analytical framework used in this research and series of steps of analytical framework applied to extract knowledge of TVET enrollment behavior and course selection variables. Finally, conclusion and discussion section brief about summary of the research performed, results, limitations and future directions.

II. Literature Review

In this section, novel works from existing similar literature have been summarized. The objective of literature review is to study similar education research / case studies who have deployed EDM techniques to extract knowledge and this knowledge has been further used to improve teaching learning environment, to understand and improve student performance, used in making predictions, and extracting various variables and factors to improve systems and process. Data mining is a business analysis process of applying descriptive techniques, reporting and business rules on large dataset to find anomalies, patterns and correlations [12]. A study [13] reveals that EDM methods can help in deep quantitative and qualitative analysis for graspable and actionable knowledge for learners and teachers. A 10 years survey [11] concludes that EDM exploration can be categorized into five groups. These groups includes (i) student modeling (ii) decision support systems (iii) adaptive systems (iv) evaluation and (v) scientific inquiry. Literature review is categorized in education streams of school education, higher education, and TVET education.

EDM has successfully been used in higher education to get insight knowledge to resolve educational and administrative problems [14,15,16]. Knowledge in these case studies have been extracted using EDM techniques of association, clustering, and classification through various algorithms. Results have helped to improve student's performance, instructor's lectures delivery mechanism, and student's behavior. A 17 years review and synthesis from year 2000 to year 2017 [17] concludes that EDM in higher education has successfully been used in four dimensions. These dimensions includes computer

supported learning analytics, computer supported predictive analytics, computer supported behavioral analytics and computer supported visualization analytics. Thus EDM has helped in higher education successfully.

Applying EDM on school education dataset has also helped to learn various aspects of school students. In Mexican school case study [18], 670 middle school student's data has evaluated using classification and decision tree EDM techniques and found that predicting school failure is a difficult task because of multifactor problems like personal, social, economic and family influence issues. Similarly, EDM has also been applied in TVET education to extract knowledge. In Brazil [19], to rationalize educational needs, TVET clustering using EDM techniques have been completed. The result of this case study shows that top three industrial cluster group to focus for TVET educational needs in Brazil are manufacturing, communication & informatics and constructions. In Turkey [20], EDM has been used to find failure factors for Vocational High School for girls. It can be concluded that school education and TVET education has equally benefitted by applying EDM.

We studied educational all three streams of higher education, school education and TVET education with respect to application of EDM techniques. It is found that EDM has successfully been used to extract knowledge to improve systems and process. In this paper, we applied EDM techniques on TVET dataset to understand TVET trainee's enrollment behavior and trainee's course selection variables. We developed an analytical framework and applied it through series of steps on TVET data to extract knowledge. Acquired knowledge can be used to not only produce quality skilled workforce according to the industry requirement and job market demand but also understand to various aspects of behavior, trends and analysis.

III. Materials & Methods

An analytical framework based on traditional Cross-Industry Standard Process for Data Mining (CRISP-DM) [21] is developed. We selected CRISP-DM for application of its steps using EDM to extract TVET course enrollment behavior and course selection variables. This CRISP-DM base analytical framework for TVET is known as "TVET Knowledge Discovery in Databases (TVETKDD)". TVETKDD is then used to analyze TVET dataset for KDD. TVETKDD is illustrated in Fig. 1 and described hereafter:-

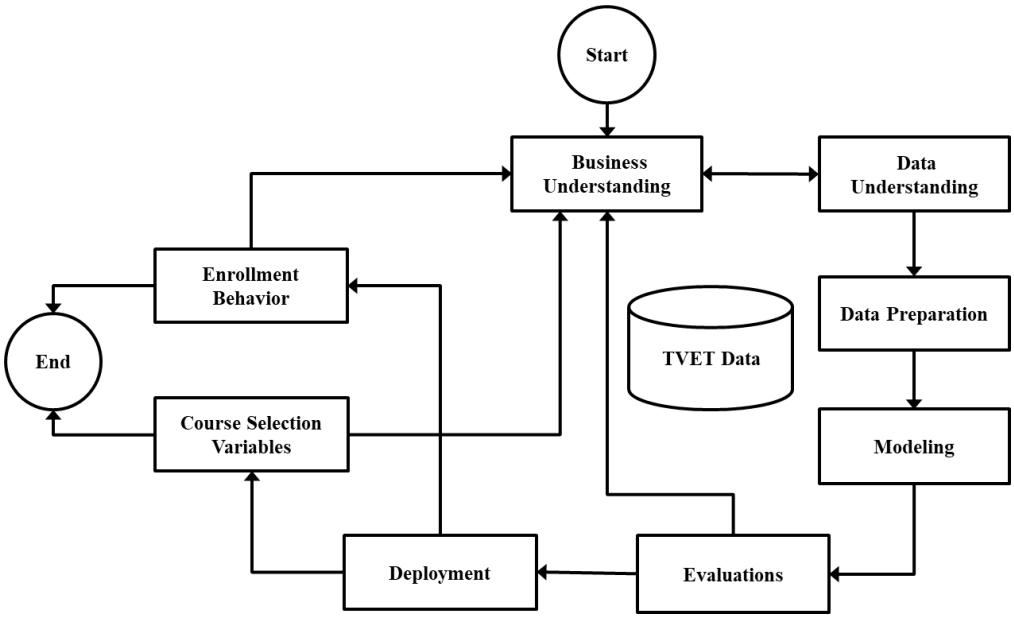

Fig. 1. Analytical Framework TVETKDD

1. Business Understanding

This initial phase help to understand business and study objectives. The objective of the study is understand TVET enrollment behavior and course selection variables of the TVET dataset. In this case study, we have used data from Punjab Vocational Training Council (PVTC) Academic System [22] for 2017 - 2018 sessions enrollment. PVTC [23] is an autonomous body and provides vocational training to youth across the province of the Punjab through its Vocational Training Institutes (VTIs) in 100+ different vocational trades.

Administratively PVTC operations for Punjab province of Pakistan are divided into 03 Regional offices (i) Central (ii) North and (iii) South. Each Regional offices is further divided into 3-4 Areas offices. Each Area office contains 20 to 50 VTI's from 2 – 5 districts of the Punjab. Administratively Punjab is divided into 36 districts [24]. These Regional and Area offices has been used define geographic locations of the trainees. Central Region contains Area Offices of Lahore, Sahiwal, Faisalabad and Gujranwala. North Region contains Area Offices of Sialkot, Rawalpindi and Sargodha and similarly South Regional contains Area Offices of Dera Ghazi Khan,

Vehari & Bahawalpur. Each Area Office comprising districts of Punjab province are listed below:-

1. Lahore Area: District Lahore & Kasur
2. Sahiwal Area: District Sahiwal, Okara & Pakpattan
3. Faisalabad Area: District Faisalabad, Jhang, Chiniot & Toba Tek Singh
4. Gujranwala Area: District Gujranwala, Sheikhpura, Hafizabad & Nankana Sahib
5. Sialkot Area: District Sialkot, Narowal, Gujrat, Mandi Bahuddin & Jehlum
6. Rawalpindi Area: District Rawalpindi, Chakwal & Attock
7. Sargodha Area: District Sargodha, Khushab & Mianwali
8. Dera Ghazi Khan Area: District Dera Ghazi Khan, Bhakkar, Layyah, Muzaffargarh & Rajanpur
9. Vehari Area: District Vehari, Khanewal, Bahawalnagar & Lodhran
10. Bahawalpur Area: District Bahawalpur, Rahim Yar Khan & Multan.

2. Data Understanding

This phase of analytical framework understand data and prepare it for further processing. To get better insight of the data during applying data mining techniques, we perform few preprocessing of the collected dataset. PVTC academic system is an online web application which is developed using Microsoft ASP.NET and Microsoft SQL Server as backend database system. The system is centralized and hosted in datacenter environment for secure access to more than 450 remote users through internet. Data was extracted from PVTC Academic System database of Microsoft SQL Server tables. All relevant information was preprocessed transferred into single Microsoft Excel file called dataset "D" for further analysis. Single Excel file was used for KDD. Dataset "D", used in this case study has been clustered using "Computer Applications" trade category / industry group. PVTC uses 19 different trades category / industry group to segregate various vocational courses of TVET. The reason for selecting this trade category / industry group is that it contain enrollment data across Punjab in majority of training institutes. Table 1 shows summary of the dataset "D".

Table 1. DATASET "D"

Description of Data	*Values*
No. of Categories / Industry Groups	01
No. of Trainees Admitted	7,426
No. of VTI's	167
No. of Trades	04

As shown in Table 1 that this data spread over 167 Vocational Training Institutes with the enrollment of 7,426 in four different trades. Table 2 shows data dictionary of dataset "D". As shown in the Table 2, certificate id (certid) is primary key for the dataset and enrollment data contain all regions and areas enrollment. It also contain both genders, all qualifications, marital status and religion in it. Dataset "D" data dictionary is shown in Table 2.

Table 2. Data Dictionary of Dataset "D"

Attributes	Description	Values
ZOName	Region Name	Central, North, South
AreaName	Area Name	Faisalabad, Gujranwala, Lahore, Sahiwal, Rawalpindi, Sargodha, Sialkot, Bahawalpur, D-G Khan, Vehari
VTIName	VTI Name	167 Names of the various Vocational Training Institutes across Punjab
Certid	Certificate ID	Unique ID for each trainee for admission process
Gender	Trainee Gender	Male, Female, Others
Religion	Trainee Religion	Islam, Christian, Hindu
MaritalStatus	Trainee Marital Status	Single, Married, Divorced, Widow
Qualification	Trainee Qualification	Primary, Middle, Matric, Intermediate, Bachelor, Master

3. Data Preparation

In this phase, data is prepared for further analysis. It was found that dataset "D" needs to be further explored based on trainee's qualification, gender, trade selection and age. To explore data distribution of dataset "D" function graph of continuous variables "Certid" was plotted. Data mining has been done using Orange Data Mining tool widgets. Orange is freeware data mining and machine learning tool [25]. We can see through precision bar of Fig. 2 that trainee's Certificate ID (Certid) is available in dataset "D" in all 10 area offices of PVTC

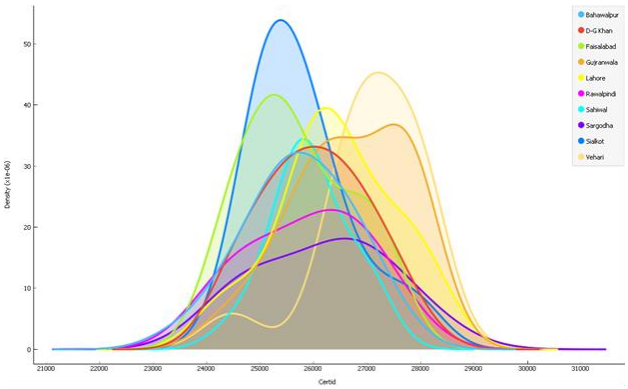

Fig. 2. Data Distribution of CertID across 10 Area Offices

of 03 regions. Fig. 2 shows that data dispersion of trainee's dataset is across all administrative areas of PVTC. This widespread will help us to get more accurate results during application of data mining techniques.

After initial data analysis it was found that attribute trainee's "age" is varying between 15 years to 35 years. Attribute trainee's "age" was further explored using data

visualization boxplot which has been shown in Fig. 3. Boxplot is standard way to display distribution of data on five number summary: minimum, first quartile, median, third quartile, and maximum [10].

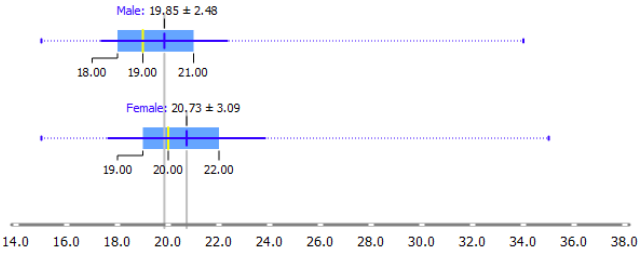

Fig. 3. Boxplot of Trainee's Gender Wise Age

Trainee's age shown in Fig. 3 numerical values "age" was converted into categorical "Age Groups" to get better insight of the data. Age Groups categorical information rather than Age numerical value has been used in further data mining process. Age groups distribution has been shown in Table 3.

Table 3. Age Groups

Age Group Name	*Age Group Description*
G1	Up to 15 Years
G2	Between 16 Years to 20 Years
G3	Between 21 Years to 25 Years
G4	Between 26 Years to 30 Years
G5	Above 30 Years

Using data mining's data visualization technique we get some primary knowledge about the data. By using scatter plot on dataset "D", it has been found that majority of trainee's falls under age group "G2" and "G3". Fig. 4 scatter plot shows the detail of trainee's age groups gender wise. Similarly, Fig. 5 bar chart shows trainee's age group trade wise, Fig. 6 bar chart shows trainee's age groups zone wise, Fig. 7 bar chart shows trainee's age groups qualification wise. Scatter plot and bar charts are data visualization techniques for viewing data through graphical means [10]

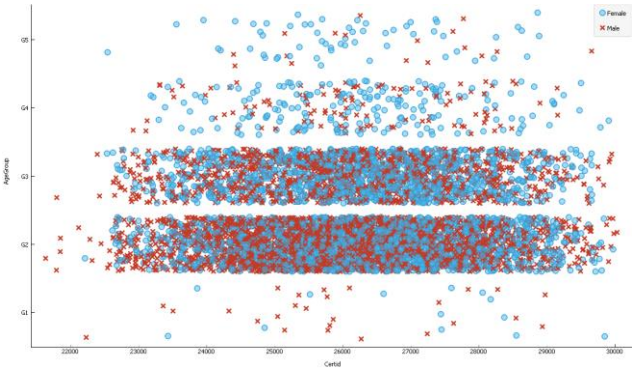

Fig. 4. Scatterplot of Trainees Age Groups Gender Wise

Fig. 4 scatterplot shows male and female population across 05 age groups. As we can see that "G2" and "G3" age group contains majority of male and females population and contains more number of "certid" as compare to other groups. A small blue circle represent "Female" and red small cross represent "Male" population. Age Group "G4" is the third largest group and similarly age group "G5" and "G1" are fourth and fifth respectively age groups of dataset "D" scatterplot in Fig. 4.

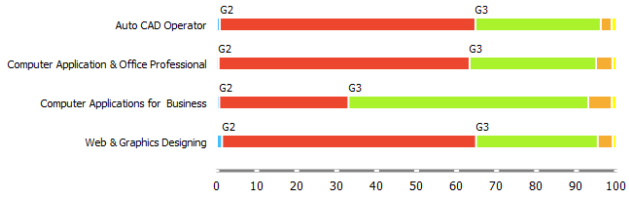

Fig. 5. Bar chart of Trainee's Age Group Trade Wise

Fig. 5 bar chart shows the dataset "D" 04 trades age group's distribution. Age group "G1", "G4" and "G5" has low population. Age "G2" and "G3" population contains the major portion of in all four trades. In "Computer Application for Business" trade, age group "G3" has more population than age group "G2". Fig. 6 bar chart shows Region wise Age Group population. In Central Region, age group "G2" is more dominant as compare to other two regions. Similarly Age group "G4" is dominant in south regions. The overall trend of bar chart shows that in central Punjab majority of trainees join Vocational Training Institute (VTIs) between 14 to 20 years of age whereas in south Punjab majority of trainees join VTIs between 20 to 30 years of age.

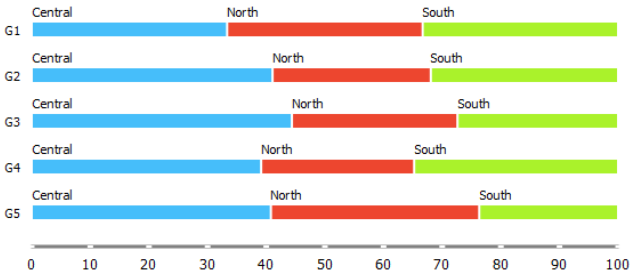

Fig. 6. Bar chart of Trainee's Age Group Wise Zone Wise

Fig. 7 bar chart shows the dataset "D" 06 qualifications age group distribution. As shown in the graph that all middle and primary education population falls in age group "G2". This shows that majority of school drop-outs who joined VTI's are between 15 to 20 years of age. As education level goes high like matric, intermediated, bachelor and master age groups vary. In master level of education, majority of trainees falls in age group "G3" & "G4". Fig. 7 shows the all details.

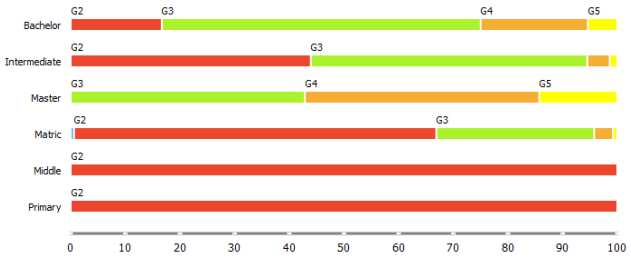

Fig. 7. Bar chart of Trainee's Trainee's Age Group Qualification Wise

4. Modeling

In this phase we applied data mining techniques to discover interesting patterns and relationships The most common data mining tasks in Education Data Mining are association, classification, clustering [14]. In the following sections, these three techniques have been applied on TVET dataset "D" and results have been shown accordingly.

4.1 Association

Association is data mining descriptive analysis technique. In association data mining we discover frequent items patterns and association rules [26]. In this case study initially we will find frequent items and then association rules for dataset "D". Frequent item patterns are used to check the frequency of items in transitions matrix. To find frequent items of dataset "D", following parameters has been set:-

1. Minimal Support : 2 %
2. Maximum no of itemsets: 10,000

Minimal support is minimal ratio of data instances that must support the itemset for it to be generated. Maximum number of instances are the limits for upward quantity of generated items. Itemset has been collapsed all. Fig. 8 shows the frequent itemset of dataset "D". As shown in Table 4 that frequent itemset support is available between 2.28 % – 85.48 % for qualification, trade name, Age Group, Area Name and Gender parameters.

To generate strong association rules in transition matrix "minimum support", "maximum number of itemsets" and "minimum confidence" are defined. Support is an indication of how frequently the itemset appears in the dataset. Confidence is an indication of how often the rule has been found to be true. To generated association rules for dataset "D", minimum support is set to 5%, minimum confidence is 50% and maximum number of itemsets are defined as 10,000. Association rules shown in Table 4, which has been sorted based on lift metrics. The most popular metrics to check the accuracy, correctness and interestingness of an association rule are support, confidence and lift [27].

Support, confidence and lift metrics are used in association mining [28] for data evaluation. Support is an indication of how frequent an item has appeared in the dataset. Confidence can be defined as the conditional probability among the items of an association rule on either-sides. Lift is ratio of support observed collectively to support being expected individually of items in an association rule; implies that items are independent in nature.

Table 4. Frequent Itemsets for Dataset "D"

Itemsets	Support	%
Qualification=Matric	6348	85.48
TradeName=Computer Application & Office Professional	6313	85.01
AgeGroup=G2	4621	62.23
Gender=Male	4079	54.93
Gender=Female	3347	45.07
AgeGroup=G3	2391	32.2
AreaName=Sialkot	922	12.42
AreaName=Gujranwala	897	12.08
AreaName=Faisalabad	884	11.9
Qualification=Intermediate	856	11.53
AreaName=Lahore	839	11.3
AreaName=Vehari	810	10.91
AreaName=D-G Khan	759	10.22
AreaName=Bahawalpur	697	9.386
AreaName=Rawalpindi	598	8.053
TradeName=Web & Graphics Designing	535	7.204
AreaName=Sargodha	520	7.002
AreaName=Sahiwal	500	6.733
TradeName=Auto CAD Operator	408	5.494
AgeGroup=G4	299	4.026
Qualification=Bachelor	204	2.747
TradeName=Computer Applications for Business	170	2.289

Table 5. Association Rules for Dataset "D"

Antecedent	Supp.	Lift	Conf.
AgeGroup=G2, Qualification=Matric, AreaName=Vehari	0.059	1.18	1
TradeName=Computer Application & Office Professional, Qualification=Matric, AreaName=D-G Khan	0.056	1.17	0.64
AgeGroup=G2, Qualification=Matric, AreaName=Faisalabad	0.055	1.16	0.99
TradeName=Computer Application & Office Professional, Qualification=Matric, AreaName=D-G Khan	0.062	1.14	0.71
TradeName=Computer Application & Office Professional, Qualification=Matric, AreaName=Faisalabad	0.055	1.13	0.71
TradeName=Computer Application & Office Professional, AgeGroup=G2, AreaName=Sialkot	0.055	1.13	0.97
TradeName=Computer Application & Office Professional, Qualification=Matric, Gender=Male	0.29	1.13	0.7
TradeName=Computer Application & Office Professional, Qualification=Matric, AreaName=Vehari	0.059	1.11	0.69
AgeGroup=G2, Qualification=Matric, Gender=Female	0.203	1.11	0.94
TradeName=Computer Application & Office Professional, AgeGroup=G2, Gender=Male	0.29	1.11	0.94
AgeGroup=G2, Qualification=Matric, AreaName=D-G Khan	0.062	1.1	0.94
TradeName=Computer Application & Office Professional, AgeGroup=G2, AreaName=D-G Khan	0.062	1.1	0.94
Qualification=Matric, Gender=Male, AreaName=D-G Khan	0.056	1.1	0.93

TradeName=Computer Application & Office Professional, AgeGroup=G2, AreaName=Gujranwala	0.056	1.09	0.93
AgeGroup=G3, Qualification=Matric, Gender=Female	0.104	1.08	0.92
TradeName=Computer Application & Office Professional, Gender=Male, AreaName=D-G Khan	0.056	1.08	0.92
TradeName=Computer Application & Office Professional, AgeGroup=G2, Qualification=Matric	0.29	1.07	0.59
TradeName=Computer Application & Office Professional, AgeGroup=G2, Gender=Female	0.203	1.05	0.9
TradeName=Computer Application & Office Professional, AgeGroup=G2, AreaName=Faisalabad	0.055	1.03	0.88
TradeName=Computer Application & Office Professional, AgeGroup=G2, AreaName=Vehari	0.059	1.02	0.87
TradeName=Computer Application & Office Professional, Qualification=Matric, AreaName=Sialkot	0.055	1.01	0.63

4.2 Clustering

Clustering is data mining technique which help in grouping similar objects [29]. It partition data based on similarity of data. Clustering method is divided into two groups: hierarchical clustering and partitional clustering [30]. We have applied hierarchical clustering on dataset "D" because we want to combine data object into bigger cluster to understand various data group available in dataset "D". Hierarchical clustering is mapped on scatter plot graph to view various cluster of dataset. Fig. 8 shows the PVTC Area office wise age group cluster using "certid" data field. As shown in Fig. 8 that clustering is available in all 10 area offices with the majority of the trainees fall into age group "G2" & "G3". Third largest group cluster is "G4". Area office Sahiwal shows no enrollment in age group "G5". There are very few enrollment in age group "G1". Fig. 9 shows trade wise age group clusters. Fig. 9 shows that "Computer Application & Office Professional" trade is heavily filled with trainees age groups "G2" and "G3" whereas other three trades have mixed enrollments.

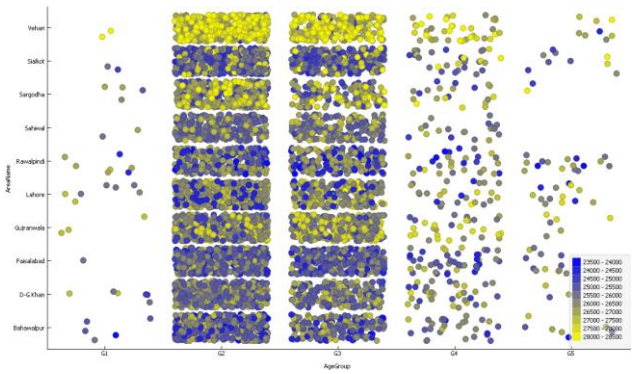

Fig. 8. Scatterplot of hierarchical clustering for Area Wise Age Group Certid Cluster

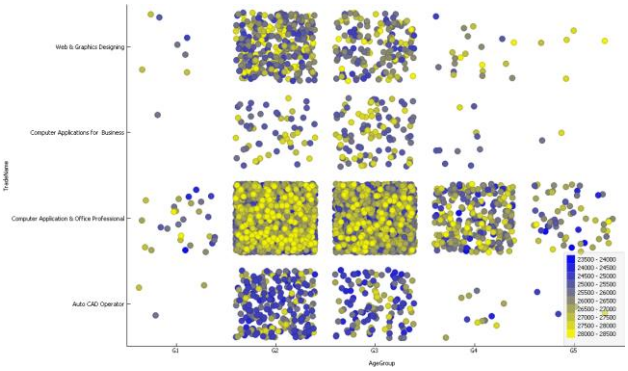

Fig. 9. Scatterplot of hierarchical clustering for Trade Wise Age Group Certid Clusters

Fig. 10 shows scatter plot graph cluster based on trade and qualification wise. As shown in Fig. 10 that majority of trainees qualification fall in all four trades are matric level and above. There are few trainees which are master level education in "Computer Application and Office Professional" and "Computer Application for Business trade". Looking at "Computer Application for Business trade" we can see that majority of trainees qualification is intermediate rather than matric. In "Web & Graphics Designing" trade trainee's fallen between matric and Bachelor qualification. Trade "Computer Application & Office Professional" cluster population shows that this trade is popular amongst all qualification age groups.

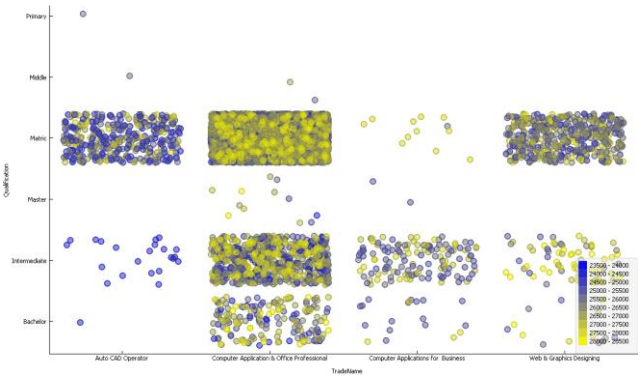

Fig. 10. Scatterplot of hierarchical cluster for Tradewise qualification wise.

4.3 Classification

Classification is data mining technique that help to find groups or classes into the dataset [31]. To apply classification dataset "D" was sliced into test and training data. From Dataset "D" total 7,426 records, 20% i.e. 1,486 records were taken as test data whereas 5,914 records were taken as training data. After splitting test and train data, classification algorithms Decision Tree, Random Forest, Neural Network and Naïve Bays were applied on dataset to test and train data. All four algorithms were applied using data mining tool Orange "Test and Score" widgets. Table 6 shows score results of "Test data" and Table 7 shows score results of "Training data".

Table 6. Score Result of Test Data

Method	AUC	CA	F1	Precision	Recall
Tree	0.618	0.848	0.779	0.747	0.848
Random Forest	0.629	0.848	0.780	0.830	0.848
Neural Network	0.603	0.848	0.779	0.849	0.848
Na ïve Bayes	0.589	0.848	0.779	0.812	0.848

Table 7. Score Result of Training Data

Method	AUC	CA	F1	Precision	Recall
Tree	0.618	0.848	0.779	0.747	0.848
Random Forest	0.627	0.848	0.779	0.762	0.848
Neural Network	0.601	0.848	0.779	0.849	0.848
Na ïve Bayes	0.589	0.848	0.779	0.812	0.848

As shown in Table 6 and Table 7 that Random Forest algorithm has shown high value of classifier as compared to other algorithms in Area Under Cover (AUC). AUC is used to measure the accuracy under the Receiver Operating Characteristic (ROC) curve. AUC predicted value is between 0 and 1. Higher value of AUC shows better performance of the classifier. Classification Accuracy (CA) is the proportion of correctness of the classified. In our case study, CA for test data & training data for all algorithms accuracy is the same. F1 score is the weighted average of precision and recall. F1 score for out model both dataset is almost the same. Precision is the ratio of correctly predicted positive. Recall is the ratio of correctly predicted positive observations to the all observations in the actual class – yes. In our case study all algorithms values for test and training data are same positive.

5. Evaluation

In this phase, we evaluated classifier model which we have built and tested during modeling phase. We used four algorithms of our dataset to test our classifier. We map four algorithms i.e. Decision Tree, Random Forest, Neural Network and Na ïve Bays on ROC curve to evaluate. As shown from Fig. 11 to Fig. 14 that all for four different trades, ROC for Computer Application for Business trade is True Positive. ROC True positive toward upper left corner shows the higher overall accuracy [32]. These overall results of all four trade's shows that our classifier accuracy is true positive but vary in all trades. Details are shown from Fig. 11 to Fig. 14.

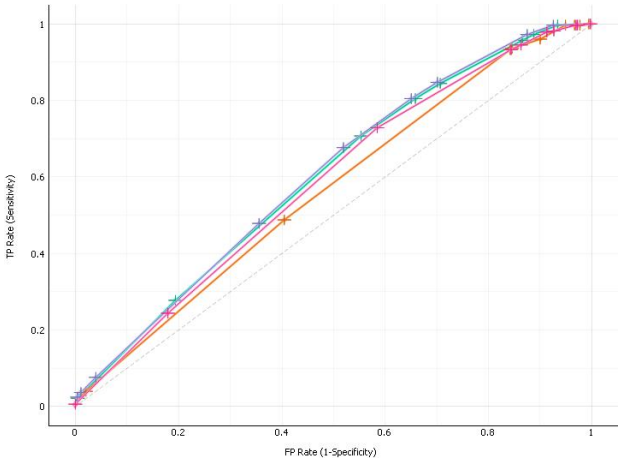

Fig. 11. ROC Curve for AutoCad Trade

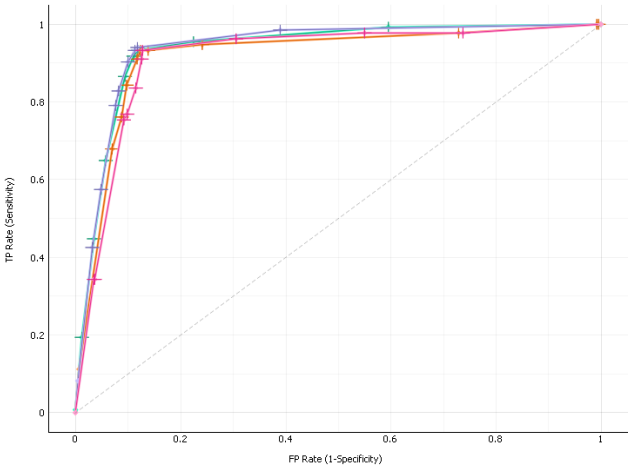

Fig. 12. ROC Curve for Computer Application for Business Trade

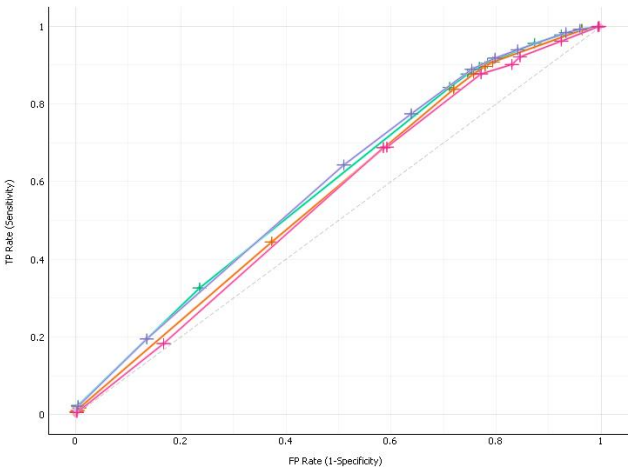

Fig. 13. ROC Curve for Computer Application Office Professional Trade

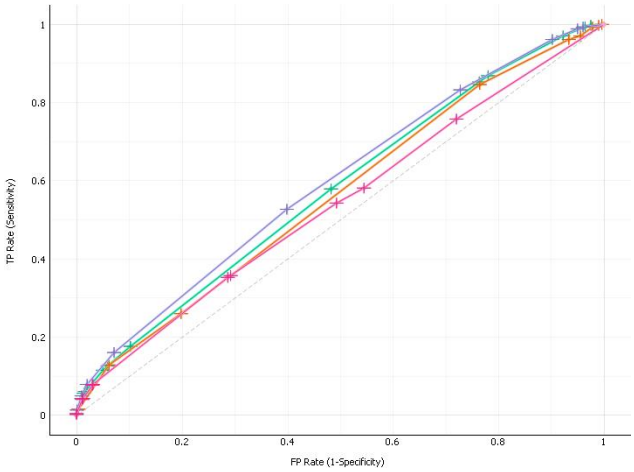

Fig. 14. ROC Curve for Web Graphics Designing Trade

Confusion matrix is another measure for performance of classifier model [33]. Confusion matrix for corrective predictions is shown in Table 8. As shown in the Table 8, classifier has accurately predicted for Computer application for business trade.

Table 8. Confusion Matrix

	AutoCAD Operator	**Computer Application & Office Professional**	**Computer Applications for Business**	**Web & Graphics Designing**
Auto CAD Operato	0	337	0	0
Computer Application & Office Professional	0	5029	0	0
Computer Applications for Business	0	141	0	0
Web & Graphics Designing	0	433	0	0

6. Deployment

In this phase of framework, we concluded our analysis results which we have acquired during different phases of analytical framework TVETKDD. Results are shown in two sections. Section 1 contain course selection variables which plays important role in TVET course selection. We can use these variables to monitor and control TVET required and produced skilled workforce. Section 7 contains trainee's enrollment behavior. This enrollment behavior can be used to not only understand existing TVET enrollment trend but also for equal distribution of resources and opportunities. Both sections details are given hereafter:-

7. Course Selection Variables

During applying frequent itemset of association mining, we found that "Age Groups", "Qualification", "Gender", "Religion" and "Marital Status" are potential variables which can plays important role in TVET course selection for Punjab youth. Table 3 shows that TVET course selection variables, their influencing factors and factor % ages. Details are shown in Table 9.

Table 9. TVET Course Selection Variable & Factors

Variable Name	**Factors**	**Factor % age**
Age Groups	G2 (16 – 20 Years)	62.23 %
	G3 (21 – 25 Years)	32.20 %
	G4 (26 -30 Years)	4.07 %
Qualification	Matric	85.48 %
	Intermediate	11.53 %
	Bachelor	2.74 %
Gender	Male	54.93 %
	Female	45.07 %
Trade Name	Computer Application & Office Professional	85.01 %
	Web Graphics Designing	7.24 %
	AutoCAD Operator	5.49 %
	Computer Application for Business	2.29 %
Religion	Muslim	99.93 %
	Non-Muslim	0.07 %
Marital Status	Single	99.13 %
	Married / Others	0.87 %

8. Enrollment Behaviors

In this section, we have concluded TVET trainee's enrollment behavior which we have acquired during phases of analytical framework TVETKDD. Application of association rules help us to find that AgeGroup = "G2" and Qualification = "Matric" has strongest association in our dataset. Similarly, Computer Application & Office Professional trade and Matric qualification has strong association in Area Faisalabad and D-G Khan. In clustering application, we found that G2 & G3 age group has strong cluster in all four trades and Area Offices of

Table 10. Geographic Enrollement Trends

Punjab – Pakistan Districts 10 Groups	**% of Enrollment**
District Faisalabad, Jhang, Chiniot & Toba Tek Singh	11.90 %
District Gujranwala, Sheikhpura, Hafizabad & Nankana Sahib	12.08 %
District Lahore & Kasur	11.30 %
District Sahiwal, Okara & Pakpattan	6.73 %
District Rawalpindi, Chakwal & Attock	8.05 %
District Sargodha, Khushab & Mianwali	7.00 %
District Sialkot, Narowal, Gujrat, Mandi Bahuddin & Jehlum	12.42 %
District Bahawalpur, Rahim Yar Khan & Multan.	9.39 %
District Dera Ghazi Khan, Bhakkar, Layyah, Muzaffargarh & Rajanpur	10.22 %
District Vehari, Khanewal, Bahawalnagar & Lodhran	10.91 %

the dataset. Summary of the extracted demographic enrollment behavior is shown Table 9. As shown the Table 9 that dataset "D" enrollment of 42.01 % is from central Punjab, 27.47 % is from north Punjab and 30.51 % is from south Punjab. Segregation of Punjab in Table 10 is done on the basis PVTC Regional structure.

IV. Conclusion And Discussion

In this paper we have analyzed TVET actual enrollment data of TVET training provider organization of Punjab, Pakistan. The aim of this analysis was to understand TVET course selection variables and enrollment behavior so that these variables and enrollment behavior can be used monitor and control required and produced TVET skilled workforce. We developed an analytical framework TVETKDD, which we applied in series of steps to extract knowledge. We found that 99.75 % TVET workforce minimum education level is Matric or above. 94.43 % age group of TVET trainees are between 16 – 25 years of age. This is very healthy sign that youth is engaged in learning TVET skills to support their families. 42.01 % enrollments in TVET courses is from Central Punjab. This shows that more efforts are needed in north and south Punjab – Pakistan to promote TVET education. 54 % Male and 45 % Females ratio of TVET trainees including remote areas of Punjab is positive indicator as far as gender is concerned.

In conclusion, we can sum up that trainee's 'age group', 'qualification', 'gender', 'religion' and 'marital status' are potential variables which can play important role in TVET course selection. Similarly, overall TVET enrollment behavior indicates that youth is engaged in TVET training and seeking opportunities for employment and entrepreneurship. Now, it is the duty of provincial & federal government authorities to not only provide employment opportunities to them but also ensure equal distribution of resources to provide training opportunities.

We have evaluated TVET data for one industrial group of TVET training organization. Developed analytical framework TVETKDD can be used to apply on bigger TVET dataset to get more accurate results. EDM has successfully been used in higher education and school education previously. Now applying EDM on TVET dataset has helped to improve coordination between funding organizations, TVET training provides and industry. Extracted enrollment behavior and course selection variables knowledge can be used to controlled supply of skilled workforce for national and international market by funding organization like government agencies, skilled enhancement projects and international donors like UNICEF, GIZ, USAID by funding only industry required skilled TVET courses.

References

[1] O. Q. Essel, E. Agyarkoh, M. S. Sumaila, and P. D. Yankson, "TVET stigmatization in developing countries: reality or fallacy," *European Journal of Training and Development Studies*, vol. 1, no. 1, pp. 27–42, 2014.

[2] UNESCO-UNEVOC, "Technical and Vocational Education and Training (TVET) Challenges and Priorities in Developing Countries - Google Search," Mar-2018. [Online]. Available: http://www.unevoc.unesco.org/e-forum/TVET_Challenges_and_Priorities_in_Developing_Countries.pdf. [Accessed: 22-Mar-2018].

[3] P. Descy and M. Tessaring, *The value of learning: evaluation and impact of education and training: third report on vocational training research in Europe: synthesis report*. Office for official publications of the European Communities, 2005.

[4] "GIZ Expertise. Start." [Online]. Available: https://www.giz.de/expertise/html/index.html. [Accessed: 23-Apr-2018].

[5] R. Carr and K. Scarim, "Delivering a world where every pregnancy is wanted every childbirth is safe and every young person's potential is fulfilled," p. 52.

[6] S. Nooruddin, "Technical and Vocational Education and Training for Economic Growth in Pakistan," *Journal of Education and Educational Development*, vol. 4, no. 1, pp. 130–141, May 2017.

[7] "Navttc – National Vocational & Technical Training Commission (NAVTTC), Pakistan," 2018. .

[8] "Comparative Analysis of TVET Sector in Pakistan." [Online]. Available: http://www.skillingpakistan.org/. [Accessed: 29-Mar-2019].

[9] R. A. Huebner, "A Survey of Educational Data-Mining Research.," *Research in higher education journal*, vol. 19, 2013.

[10] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[11] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, "Educational data mining applications and tasks: A survey of the last 10 years," *Educ Inf Technol*, vol. 23, no. 1, pp. 537–553, Jan. 2018.

[12] G. Shmueli, P. C. Bruce, I. Yahav, N. R. Patel, and K. C. Lichtendahl Jr, *Data mining for business analytics: concepts, techniques, and applications in R*. John Wiley & Sons, 2017.

[13] M. Berland, R. S. Baker, and P. Blikstein, "Educational Data Mining and Learning Analytics: Applications to Constructionist Research," *Tech Know Learn*, vol. 19, no. 1–2, pp. 205–220, Jul. 2014.

[14] A. El-Halees, "Mining students data to analyze e-Learning behavior: A Case Study," 2009.

[15] M. Agaoglu, "Predicting instructor performance using data mining techniques in higher education," *IEEE Access*, vol. 4, pp. 2379–2387, 2016.

[16] V. Gramoli *et al.*, "Mining autograding data in computer science education," in *Proceedings of the Australasian Computer Science Week Multiconference*, 2016, p. 1.

[17] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics and Informatics*, vol. 37, pp. 13–49, Apr. 2019.

[18] C. Márquez-Vera, C. R. Morales, and S. V. Soto, "Predicting school failure and dropout by using data mining techniques," *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 8, no. 1, pp. 7–14, 2013.

[19] P. N. Marra, "Identifying educational needs by rationalising TVET industry clusters in Brazil," p. 19.

[20] O. Deperlioglu and F. S. Birtil, "Analysis of Girls Vocational High School Students' Academic Failure Causes with Data Mining Techniques," *The Anthropologist*, vol. 23, no. 3, pp. 505–512, Mar. 2016.

[21] R. Wirth, "CRISP-DM: Towards a standard process model for data mining," in *Proceedings of the Fourth*

International Conference on the Practical Application of Knowledge Discovery and Data Mining, 2000, pp. 29–39.
[22] PVTC MIS Department, "PVTC MIS System," *PVTC MIS System*, Mar-2018. [Online]. Available: http://119.159.229.99/Default.aspx. [Accessed: 20-Mar-2018].
[23] PVTC, "PVTC Official Website," *Punjab Vocational Training Council Official Website*, Mar-2018. [Online]. Available: http://www.pvtc.gop.pk/. [Accessed: 20-Mar-2018].
[24] "Districts | Punjab Portal." [Online]. Available: https://punjab.gov.pk/districts. [Accessed: 28-Aug-2019].
[25] J. Demšar *et al.*, "Orange: data mining toolbox in Python," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2349–2353, 2013.
[26] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables," in *Acm Sigmod Record*, 1996, vol. 25, pp. 1–12.
[27] S. H. Park, S. Y. Jang, H. Kim, and S. W. Lee, "An association rule mining-based framework for understanding lifestyle risk behaviors," *PloS one*, vol. 9, no. 2, p. e88859, 2014.
[28] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.
[29] P. Berkhin, "A Survey of Clustering Data Mining Techniques," in *Grouping Multidimensional Data*, Springer, Berlin, Heidelberg, 2006, pp. 25–71.
[30] Y. Rani and D. H. Rohil, "A Study of Hierarchical Clustering Algorithm," p. 8.
[31] S. Jain, R. Raghuvanshi, and M. Ilyas, "A Survey Paper on Overview of Basic Data Mining Tasks," 2017.
[32] K. H. Zou, A. Liu, A. I. Bandos, L. Ohno-Machado, and H. E. Rockette, *Statistical evaluation of diagnostic performance: topics in ROC analysis*. Chapman and Hall/CRC, 2016.
[33] "An improved method to construct basic probability assignment based on the confusion matrix for classification problem - ScienceDirect." [Online]. Available: https://www.sciencedirect.com/science/article/pii/S00200255160004 4X. [Accessed: 01-Aug-2019].

Authors' Profiles

Rana Hammad Hassan is student of Ph.D. Computer Sciences in University of Management & Technology, Lahore Pakistan. His research focus includes Big Data, Data Analytics and Machine Learnings. He has 15+ years industry experience and currently working as Manager Management Information Systems with Punjab Vocational Training Council, Lahore Paksitan. He is the winner of Intel Education Award 2010-2011 for Punjab – Pakistan initaives. He is member of Technical Advisory Group (TAG) of National Skills Information System (NSIS) of Pakistan. He is certified as Oracle Certified Expert Consultant, Microsoft Certified Solution Developer, LRQA Quality Management System Internal Auditor. He is trained for PMP, CCNA, OCP-DBA and JAVA applications

Shahid Mahmood Awan has received Ph.D. in Computer Science (Machine Learning) from University of Engineering and Technology, Lahore in 2015. He has 14 years of research, teaching and software development experience. His research interests include: Big Data Analytics, Machine Learning, Deep Learning, Natural Language Processing, and Smart Environments. He is currently working as Assistant Professor at University of Management and Technology, Lahore. He is active member of IEEE Computer Society and Industrial Electronics Society.

How to cite this paper: Rana Hammad Hassan, Shahid Mahmood Awan, " Identification of Trainees Enrollment Behavior and Course Selection Variables in Technical and Vocational Education Training (TVET) Program Using Education Data Mining", International Journal of Modern Education and Computer Science(IJMECS), Vol.11, No.10, pp. 14-24, 2019.DOI: 10.5815/ijmecs.2019.10.02