International Conference - Innovation-2021-Innovation-2021
International Journal of Scientific Research in Computer Science, Engineering and
Information Technology ISSN : 2456-3307 (www.ijsrcseit.com)
Volume 8, Issue 3, May-June-2021

Career Path Prediction Using Machine Learning Classification Techniques

Prathamesh Gavhane*, Dhanraj Shinde, Ashwini Lomte, Naveen Nattuva, Shital Mandhane
Department Computer Engineering, Dr. D. Y. Patil School of Engineering, Pune, Maharashtra, India

ABSTRACT

In today's era, choosing the right career option is a challenging task [5]. Starting at the early stage of life students usually fail to grasp the idea of which career to pursue as they lack maturity and the experience related to that field. Furthermore, students suffer greatly in deciding which career would result the highest benefit. Students do not have sufficient knowledge to take the decision on their own which may lead to complications in future. In order to avoid future complications students should make a proper decision in selecting a highest benefit career for them. Selecting a wrong career which is not meant for them will end up with work in which they are not interested or they do not have that much knowledge in that field. As students lack in decision making, they reach fortune tellers hoping that they will guide them on the right path for a bright future [3]. Instead of relying on fortune tellers to make the best prediction for the future. By considering all these things in this work we will scientifically and systematically study the feasibility of career path prediction from the survey data. This model will recommend students a career choice according to their abilities and qualities with respect to their field. If students end up having good abilities and qualities in their respective field, they can select that field otherwise they have to drop that field and choose another one. This paper presents a career path prediction using machine learning which will help students to select the appropriate career for their bright future. As career recommendations are a unique approach, we feel it should be an interactive platform. So, while building the application we presented an interactive framework which will allow students interactively perform the task and get results. The present work has 15 different types of career options. Experiments have been done using machine learning supervised classification techniques like Logistic Regression, Decision Tree, KNN, Naïve Bayes, SVM, Random Forest, Stochastic Gradient Descent, AdaBoost, XgBoost, and some hybrid algorithms using stacking like SvmAda, RfAda and KnnSgd.

Keywords: Machine Learning, supervised classification algorithms, Career prediction.

I. INTRODUCTION

Machine Learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications. Machine learning helps the computers to act without explicitly being programmed. Simply it is giving computers the ability to learn by using statistical techniques [11]. This helps in solving very complex tasks and problems very easily and without involving much human labour.

Nowadays students are getting confused about their right career path. Because of this confusion they are ending up with the career in which they are least interested. In today's world competition is heavily increasing day by day. Mainly it is too heavy in present day's technical society. Students need to be firmly well organized and planned from the initial stages of their education, so as to reach the goal. To help them in improving themselves, motivating themselves to a better career path. So, it is very important to constantly evaluate their performance, identify their interests and evaluate how close they are to their goal and whether they are on the right path that directs towards their targeted [11].

There are many new career opportunities in every field, because with the increase in research and exploration in various domains. The reasons for this confusion could be unawareness of self-talent and self- personality traits, unawareness of the various options available, equal interests in multiple fields, less exposure, market boom, assumed social life, peer-pressure etc. This creates more confusion for the students to select one career option. There should be proper counselling of the student's psychology, interest and their capacity to work in a particular field [7]. Otherwise, students may select a wrong career option and the consequences of this wrong decision could be work dissatisfaction, poor performance, anxiety and stress, social disregard etc.

There are career counselling services which are helping students to find their career goals, which is the reason counselling centres have been established. These counselling centre's help students to know the wide variety of options available for them. Now students can choose the best path for them provided by the counsellor.

In this paper we are going to provide a machine learning model which will give you the career path prediction. To the best of our knowledge, there is no available benchmark dataset suitable for career path modelling [3]. We thus created new datasets by crawling fifteen popular career paths, namely engineer, doctor, pharmacist, lawyer, archaeologist, financial advisor, motivational speaker, chartered accountant, hotel management, wedding planner, writer, photographer, cabin crew, journalist and translator. For each career path we have an individual dataset. We have collected the dataset by forwarding the google form links to the students, which were containing some questions to answer. Likewise, we successfully completed the collection of datasets.

II. LITERATURE REVIEW

[1]. Roshani Ade & P. R. Deshmukh (2014). In this paper for classification of students using psychometric tests. They used incremental naive bayes algorithm. And the results were TP-Rate_0.896, FP Rate_0.01, Precision_0.903, Recall_0.896, F-Measure_0.893 and ROC-Area_0.99. In future naïve bayes algorithm can be used as a weak classifier in the ensemble concept for incremental learning.

[2]. Ahmad F. Subahi (2018). He proposes a data collection strategy to build the required career path prediction dataset for a promising data-driven system. A new artificial neural network (ANN) approach for career path prediction was used.

[3]. Ye Liu, ET AL (2016). They have created a career path prediction model for career path instead of going to the fortune tellers. They have collected the information from various social networks. And the future work is to extend the model to consider the source descriptiveness and learn the source confidence adaptively.

[4]. Beth Dietz-Uhler & Janet E. Hurn (2013). So, they have used a learning analytics to predict student success through a perspective of faculty. In this paper, they defined about learning analytics, how educational institutions has been used it, what learning analytics tools are available and how faculty can make use of data in their courser to improve the performance of students.

[5]. Min Nie, ET AL (2020). In past, professional career appraisers used questionnaires to suggest the best career path for a student, instead of that they have created a career choice prediction based on campus big data mining the potential behavior of college students. Algorithm used is XGBOOST (ACCBOX). Accuracy of ACCBOX was 0.638.

[6]. Amer Al-Badarenah & Jamal Alsakran (2016). As we know that there are recommendation systems for the recommendation purpose while online shopping, movies, songs, etc. In that way they have created an automated recommender system for course selection which will be easy for students to choose the right subject for them.

[7]. Nikita Gorad, ET AL (2017). Keeping in mind that selecting the right career is one of the important decisions. Some students end up selecting wrong decision. For that purpose, they have created a career counselling model using data mining. They used adaptive boosting algorithm which gave around 94% of accuracy.

[8]. Dileep Chaudhary, ET AL (2019). For selecting an appropriate career path, they have created a student future prediction model using machine learning. Algorithms used were linear regression, decision tree and random forest, to improve accuracy they used adaptive boosting over the algorithms.

[9]. Vivek Kumar Mourya, ET AL (2020). They have created a career guide application using machine learning. Through this application students can easily choose a best career path for them. The machine learning algorithm used for predicting is a clustering algorithm named as K-means algorithm.

[10]. Lakshmi Prasanna & DR.D.Haritha (2019). Keeping recommender system in mind, they have created a smart career guidance and recommendation system. This paper proposes feasible predictions for student's field selection based on their marks and choice of interest. Ten to eleven machine learning algorithms were used for the predictions. In which logistic regression gave 82% accuracy. In future we can use clustering methods for better understanding.

[11]. K. Sripath Roy, ET AL (2018). They have created a student career prediction model using advanced machine learning techniques. Algorithms used are support vector machine (SVM), xgboost and decision tree. SVM gave more accuracy with 90.3 percent and then the XG Boost with 88.33 percent accuracy.

[12]. Mubarak Albarka Umar (2019). A case study of student academic performance prediction using artificial neural networks was presented. This study presents a neural network model capable of predicting student's GPA using students' personal information, academic information, and place of residence. Thus, the model correctly predicts 73.68% of student performance and specifically, 66.67% of students that are likely to dropout or experience delay before graduating.

[13]. Ezenkw.C.P, ET AL (2017). In this paper, an Automated Career Guidance Expert System (AC-GES) has been developed using case-based reasoning (CBR) technique. AC-GES is to assist high school students in choosing career paths that best suit their abilities based on their previous performances in some selected subjects, using Nigerian students as a case study.

[14]. Leaf Abu Amirah, ET AL (2016). They have used data mining technique in educational data to predict student's academic performance using

ensemble methods. They have used bagging, boosting and random forest (RF) and set of classifiers such as artificial neural network, naïve Bayesian and decision tree. The obtained results reveal that there is a strong relationship between learner's behaviors and their academic achievement.

[15]. Sudheep Elayidom, ET AL (2009). They have applied data mining on dataset using statistical techniques for career selection. This will help the students in a great way in deciding the right path for them for a bright future. The software developed is simple to use besides being reasonably accurate. Moreover, the user-friendly interface used in this project turns out to be easy to handle and avoid complications.

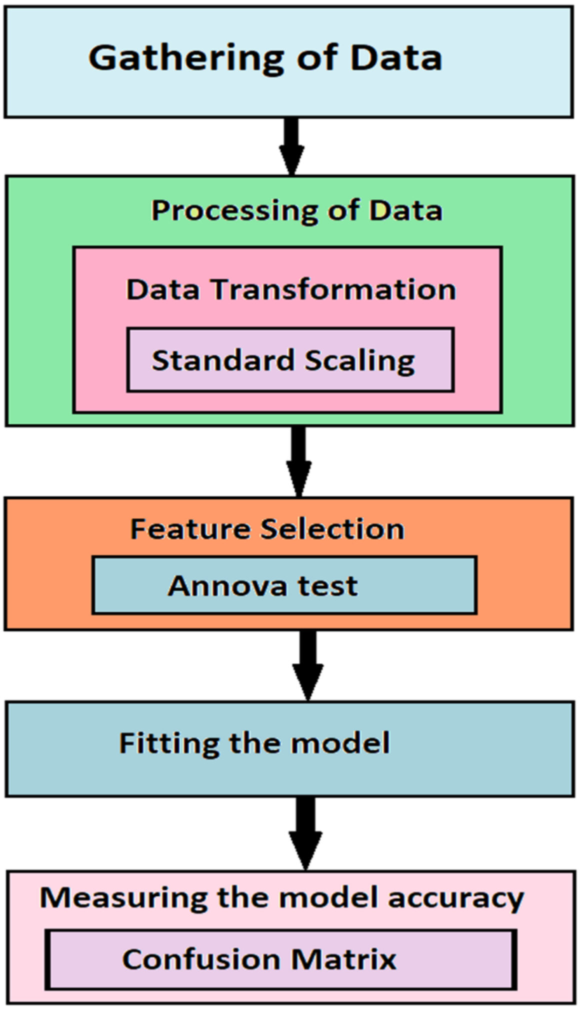
[16]. Maha Nawaz, ET AL (2014). In this paper they have created an automated career counseling system for students using case-based reasoning (CBR) and J48. This model presents an automated system that copies a one-to-one meeting with a professional career counselor. Out of the two algorithms tested, CBR gave the highest accuracy and Decision tree J-48 gave the lowest accuracy. The results indicate that the system is capable of correctly proposing majors with approximately 80% accuracy when presented with sufficient data and features.

III. RESEARCH METHODOLOGY

The proposed work is done with the use of data that is collected from survey forms. There are 15 types of datasets namely Engineering, Doctor, Chartered accountant, Cabin crew, Journalist, Photographer, Lawyer, Pharmacist, Archaeologist, Motivational speaker, Writer, Wedding planner, financial advisor, Hotel management, Translator. The various classification methods have been used to predict the class of the careers.

The whole procedure of analysis is divided into the following steps:

1. Gathering the data
2. Pre-processing of the data
3. Feature Selection
4. Fit the model
5. Measure the model accuracy



IV. PREPROCESSING OF DATA

Pre-processing is a step-in data science to clean, transform and reduce the data in order to better fit the model.

There are various methods of data pre-processing:

1) Data Cleaning

   a) Missing Data
   b) Noisy Data
2) Data Transformation
   a) Normalization
   b) Attribute selection
3) Data Reduction
   a) Aggregation
   b) Dimensionality Reduction

In this work, standard scaling has been done to transform the data

The standard score of a sample x is given as

Z = (x- μ) s

Where μ is the mean of samples and s is the standard deviation.

V. FEATURE SELECTION

Feature selection is a method to select the best features that are able to contribute more in the prediction of the output.

The methods that are used in the feature selection are:

1) Filter Method
   i) Pearson Correlation
   ii) Linear Discriminant Analysis
   iii) Analysis of Variance (ANOVA)
   iv) Chi Square Test
2) Wrapper Method
   i) Forward Selection
   ii) Backward Elimination
   iii) Recursive Feature Elimination
3) Embedded Method

In the present work, Anova Test is used to do the feature selection. The general formula for Anova test is:

F = MST/MSE

Where,

F = Anova Coefficient

MST = Mean sum of squares between the groups

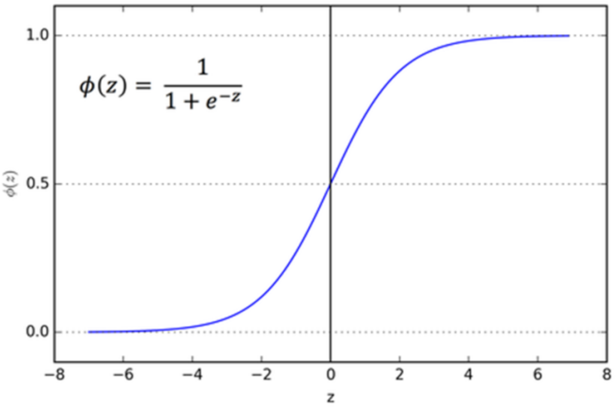
MSE = Mean sum squares due to error

VI. FITTING THE MODEL

The detailed steps of the machine learning classification techniques that has been used in the proposed work are discussed here one by one

LOGISTIC REGRESSION

Logistic regression is one such regression algorithm which can be used for performing classification problems. It calculates the probability that a given value belongs to a specific class. If the probability is more than 50%, it assigns the value in that particular class; else if the probability is less than 50%, the value is assigned to the other class. Therefore, we can say that logistic regression acts as a binary classifier. We use the sigmoid function as the underlying function in Logistic regression as shown in fig below



DECISION TREE

The decision tree is an easy technique to reach a conclusion following some conditions. A decision tree contains two types of nodes: a) Decision Node and b) Leaf Node. The decision node tells the condition that which attribute has to be selected and the leaf node tells the class.

The primary decision node is known as the root node. Each decision node is selected on the basis of the two different popular methods:

1. Information Gain Method
2. Gini Index Method

Now for obtaining the result we have calculated the Entropy.

Entropy is the measure of randomness in the data. In other words, it gives the impurity present in the dataset.

Entropy is given the formula:

E = -p*log2(p) – q*log2(q)

Therefore, Information Gain is given by,

Gain (T, X) = Entropy(T) – Entropy (T, X)

Where,

T = Parent node before split.

X = split node from T.

Gini Index is given by,

Gini = 1 - Σ(pi)2

Where,

P = Probability of particular class.

Stochastic Gradient Descent Classifier:

Stochastic Gradient Descent (SGD) is an efficient technique for linear classification problems under the convex loss functions such as (linear) support vector machine and logistic regression.

SGD is merely an optimization technique and does not correspond to a specific set of machine learning algorithms. The advantage of stochastic gradient descent is efficiency and ease of implementation.

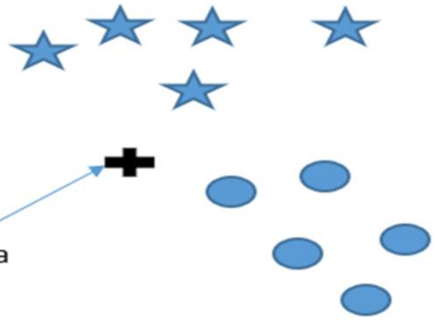
In this work, a linear support vector machine is used as a classifier and a gradient descent algorithm is applied to optimize the result.

K-NEAREST NEIGHBORS

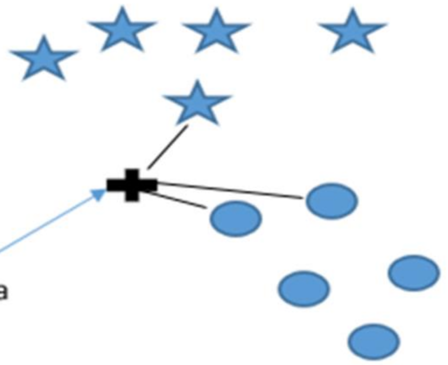
K-nearest neighbors (KNN) is a type of supervised learning algorithm which is used for both regression and classification purposes, but mostly it is used for the later. Given a dataset with different classes, KNN tries to predict the correct class of test data by calculating the distance between the test data and all the training points. It then selects the k points which are closest to the test data. Once the points are selected, the algorithm calculates the probability (in case of classification) of the test point belonging to the classes of the k training points and the class with the highest probability is selected. In the case of a regression problem, the predicted value is the mean of the k selected training points.

Let's understand this with an illustration:

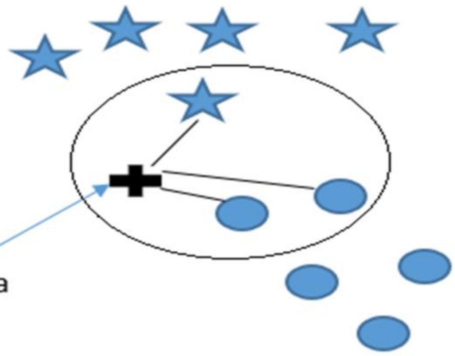
1) Given a training dataset as given below. We have a new test data that we need to assign to one of the two classes.



2) Now, the k-NN algorithm calculates the distance between the test data and the given training data.



3) After calculating the distance, it will select the k training points which are nearest to the test data. Let's assume the value of k is 3 for our example.



4) Now, 3 nearest neighbors are selected, as shown in the figure above Number of Oval class values = 2

Number of Star class values = 1 Probability (Oval) = 2/3 Probability (Star) = 1/3 Since the probability for oval class is higher than Star, the k-NN algorithm will assign the test data to the oval class.

NAIVE BAYES:

Naïve Bayes is a classification technique that uses the Bayesian theorem to predict the class for the new feature set.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Where,

P(A|B) = Probability of A occurring given evidence B has already occurred.

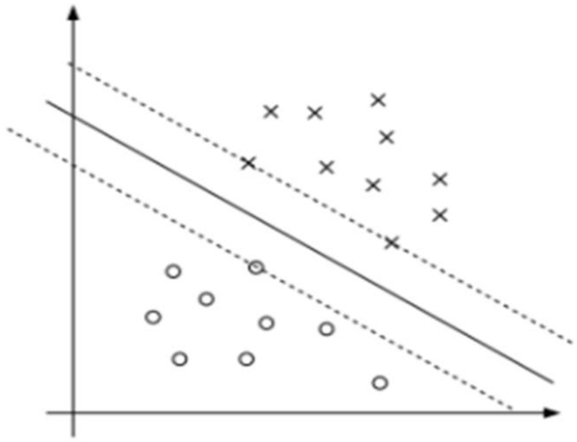
P(B|A) = Probability of B occurring given evidence A has already occurred.

P(A) = Probability of A occurring.

P(B) = Probability of B occurring

SUPPORT VECTOR MACHINE

Support Vector Machines are the one of the most effective algorithms to solve the linear problems. Although, Logistic regression also classifies the linear problems but SVM uses the concept of support vectors to do linear separation. It has a clever way to reduce overfitting and can use many features without requiring too much computation.



Consider the above fig is a hypothetical example of a dataset which is linearly separable and a decision boundary is drawn as a solid line as a plane with two dotted lines as positive and negative planes. The stars in green colour are considered as positive point while circle in orange colour is negative point

The equation of positive plane is given as $w^Tx + b$ +1, equation of negative plane is given as $w^Tx + b = -1$ and equation of hypothesis plane is given as $w^Tx + b = 0$.

The positive and negative plane changes as the hypothesis plane changes. For better understanding, here points have been taken that are linear separable but in the real world the data can be non-separable. The distance from the hypothesis plane to the nearest positive point or plane is given by d+. and distance from the hypothesis plane to the nearest negative point or plane is given by d-

Hence the equation for the positive plane is given as $w^Td+ + b = +1$ and that of the negative plane is given as $w^Td- + b = -1$

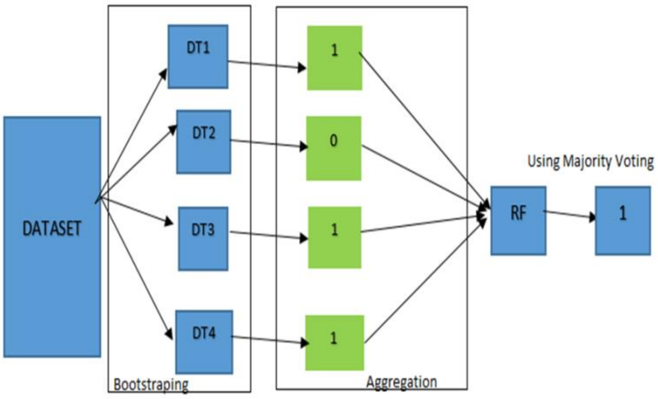
Random Forest Classifier

Random Forest is an ensemble technique. Ensemble technique is an aggregation or combination of several base models. The ensemble technique is of two types: Bagging and Boosting

Bagging is also called as bootstrap aggregation. Feeding the data to the base models by row sampling with replacement and predicting the classes is called as bootstrapping and aggregation is the result based on majority vote of base models on the test data.

Random Forest is a bagging technique that uses the Decision tree as its base model. It applies both feature sampling and row sampling with replacement to feed the data to the base models.

The figure given below is an example showing the Random Forest classification.



Suppose training dataset which is being classified into 0 or 1 that is binary classification is given to different decision tree models with the feature sampling and row sampling with replacement then the results by the decision trees are given as shown in the figure. Now when a test dataset is passed then the results of the decision trees aggregates using the majority voting method to predict the final class.

AdaBoost:

Adaboost is an ensemble technique. It is a boosting algorithm. It combines the weak learners or classifiers to improve the performance. Each learner is trained with a simple set of training samples. Each sample has a weight and the sample of all the weights are adjusted iteratively. Adaboost iteratively trains each learner and calculates a weight for each one, and this weight represents the robustness of the weak learner. Here the decision tree is used as a base learner.

The Adaboost algorithm has three main steps:

- **Sampling step:** In this step, some samples Dt are selected from the training set, where Dt the set of samples in the iteration t.
- **Training step:** In this step, different classifiers are trained usingDt, and the error rates (ϵi) for each classifier is calculated.
- **Combination step:** Here all trained models are combined.

Stacking is an ensemble technique. It combines the predictions of many machine learning models on the same dataset.

1. The stacking uses the two-level architectures:

Base Level (0- Level):

2. The base level architecture consists of the base machine models of which some features are to be combined.

Meta Level (1-Level):

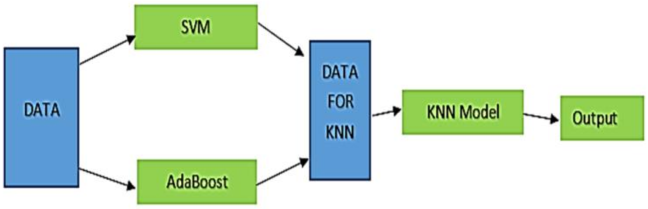
The Meta level architecture consists of the machine learning model that learns how to best combine the predictions of the base model.

The base models are trained on the training dataset while the Meta model is trained based on the predictions of the base model. Hence the output of the base model works as an input to the Meta model.

In this work, the stacking is done in order to produce three different new classifiers. In all the algorithms, KNN is used as the Meta model while the base models are changed. Let's discuss all the models one by one:

SVMAda:

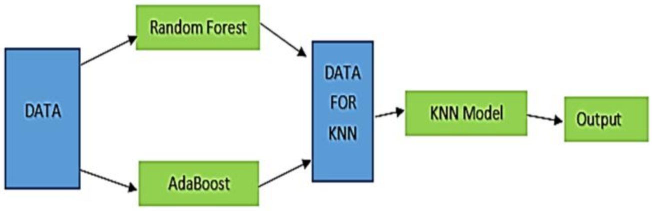
This classifier consists of the two base classifiers linear support vector machine and Adaboost. The outputs of the base models are then given to the KNN to predict the final output. The model consists of the two base models SVM (Support Vector Machines) and Adaboost hence it is named as SvmAda.



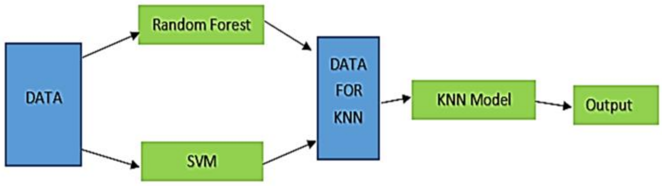
RfcAda:

This classifier consists of the two base classifiers Random Forest and Adaboost. The outputs of the base models are then given to the KNN to predict the final output. The model consists of the two base models

RFC (Random Forest Classifier) and Adaboost hence it is named as RFCAda.

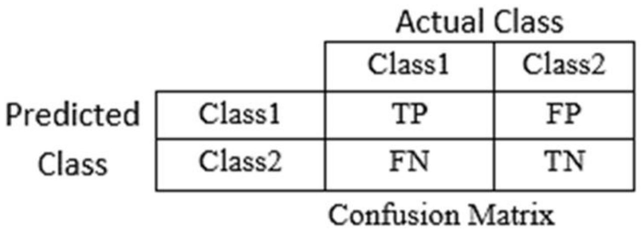


SvmRfc:
This classifier consists of the two base classifiers linear support vector machine and Random Forest Classifier. The outputs of the base models are then given to the KNN to predict the final output. The model consists of the two base models SVM (Support Vector Machines) and Rfc(Random Forest Classifier) hence it is named as SvmRfc.



CALCULATION OF ACCURACY

The accuracy is calculated on the basis of the confusion matrix. The confusion matrix is a table that is used to calculate the accuracy.



Actual Class

		Class1	Class2
Predicted Class	Class1	TP	FP
	Class2	FN	TN

Confusion Matrix

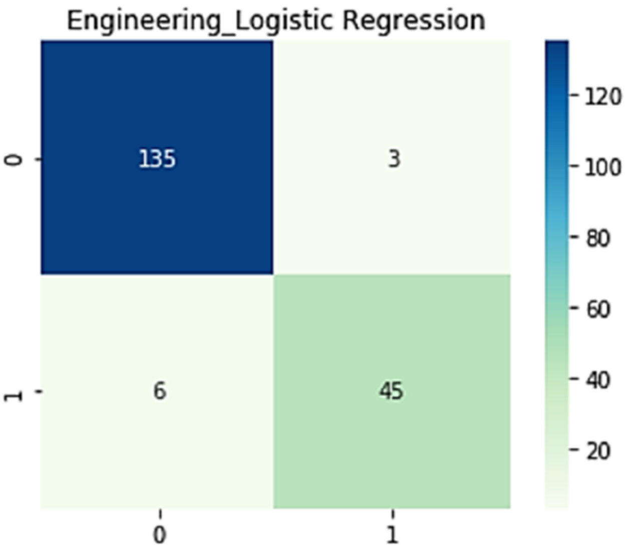
Accuracy = TP+TN/TP+FP+TN+FN

Where:
TP: True Positive
TN: True Negative
FN: False Negative
FP: False Positive

VII. RESULT AND ANALYSIS

Each algorithm has been implemented using some common steps. The feature selection is done with the Anova test.

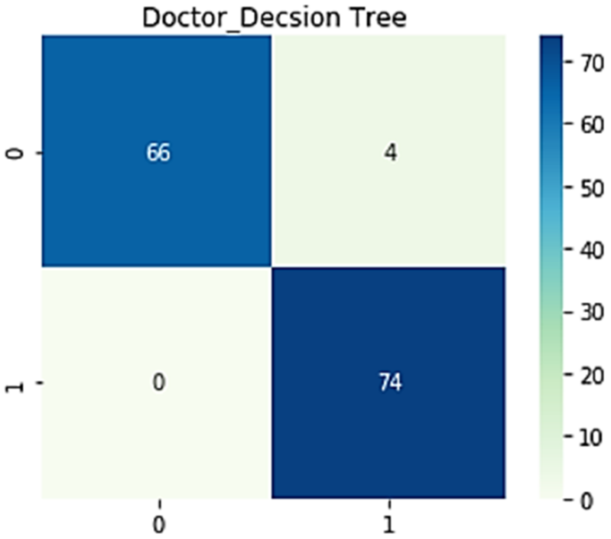
Engineering (Logistic Regression):
Logistic Regression classifier classifies the data based on the sigmoid function. The accuracy is calculated on the basis of selecting the features. Evaluating the accuracy, it was 95.24%. Accuracy was calculated using the Confusion matrix as shown in Fig below.



Accuracy = 135 +45/135+3+6+45
= 95.24

Hence the accuracy obtained is 95.24%.

Doctor (Decision Tree):
Initially the model was trained with an imbalanced dataset. Due to which model was biased. Later data was balanced using oversampling. Decision tree was used for model building and it achieved the accuracy of 97.22%. Accuracy was calculated using a confusion matrix as shown in fig. below



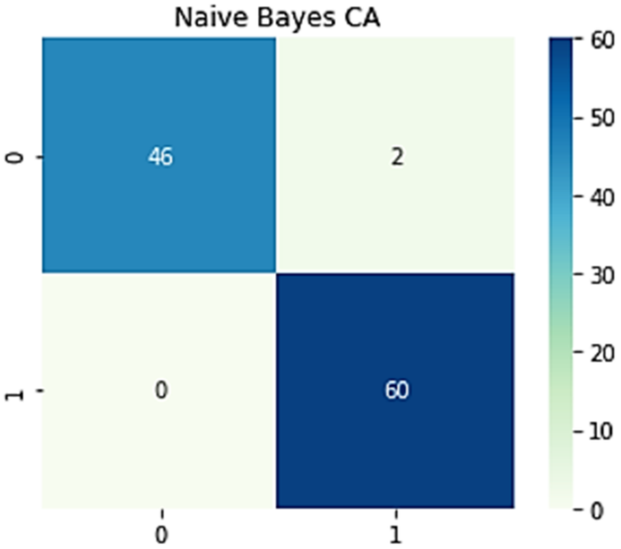
Accuracy = 66+74/66+4+0+74

= 97.22

Hence the accuracy obtained is 95.24%.

CA (Naïve Bayes):

Naïve Bayes algorithm performed well when all the features were used for classifying the career data the overall accuracy achieved is calculated using confusion matrix shown in Fig. (19)



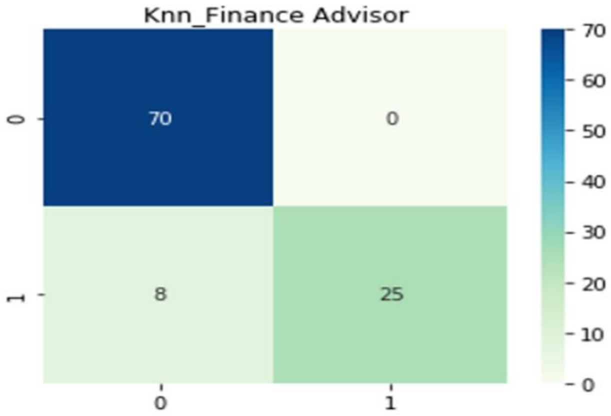
Accuracy = 46+60/46+2+0+60

= 98.15

Hence the accuracy obtained is 98.15%.

Financial Advisor (Knn):

To prepare the k-NN model dataset is divided into a training set (80%) and test set (20%). On evaluating the performance of model by considering the all features the accuracy achieved was 89.39% this is calculated using confusion matrix which is shown in Fig. (16)



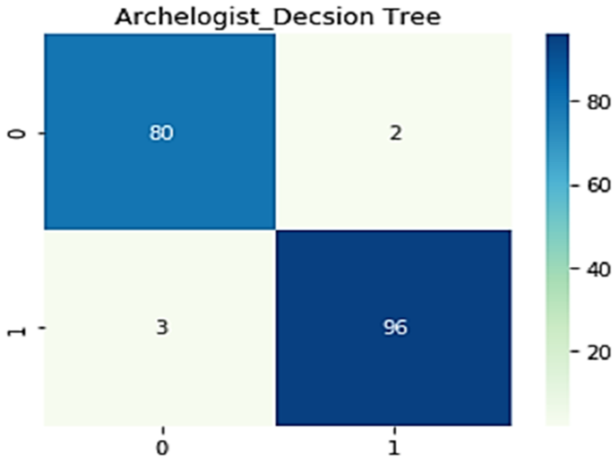
Accuracy = 46+60/46+2+0+60

= 98.15

Hence the accuracy obtained is 98.15%.

Archaeologist (Decision tree classifier):

Initially the archaeologist's dataset was passed through various algorithms. It was found out that the model is performing good with the decision tree. It achieved an accuracy of 97.24%. Accuracy was calculated using a confusion matrix. As shown below.



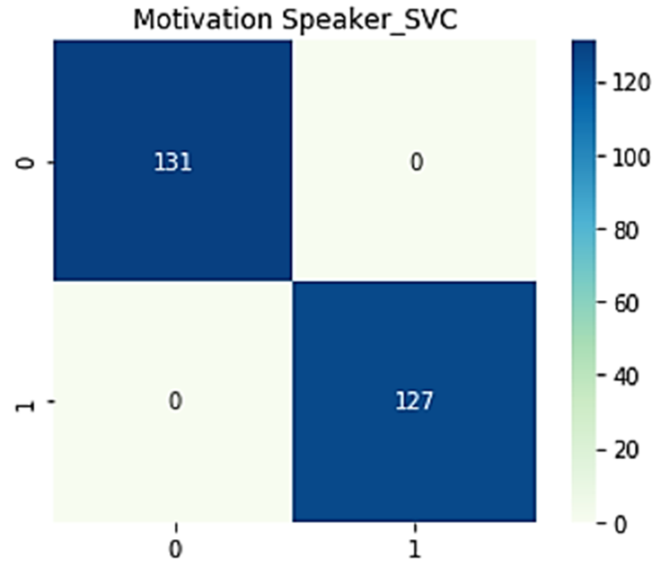
Accuracy = 80+96/80+2+3+96

= 97.24

Hence the accuracy obtained is 97.24%.

Motivational Speaker (Support Vector Classifier):

Firstly, Model was trained using the support vector machine by considering five best features. On evaluating the performance of the algorithm, accuracy achieved was 88.64%. Further the features were added according to their importance. Model outperformed when all the features were considered for classifying the data. The confusion matrix is shown in fig. (21)



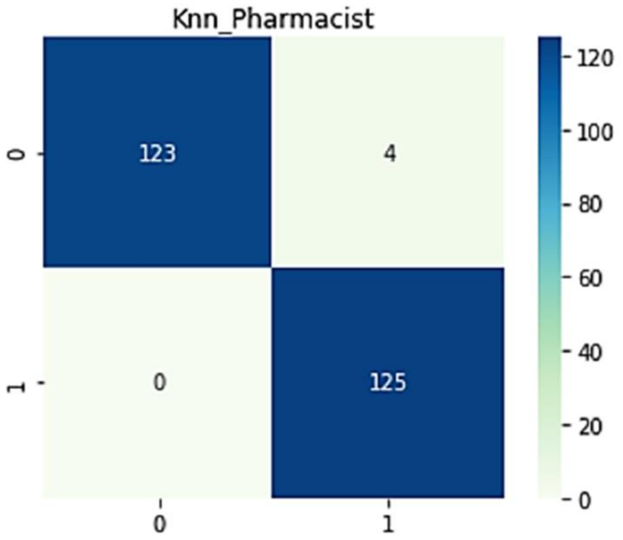
Accuracy = 131+127/131+0+0+127

= 100%

Hence the accuracy obtained is 100%.

Pharmacist (Knn):

Initially the model was trained with an imbalanced dataset. Due to which model was biased. Later data was balanced using oversampling. Decision tree was used for model building and it achieved the accuracy of 98.41%. Accuracy was calculated using a confusion matrix as shown in fig. below



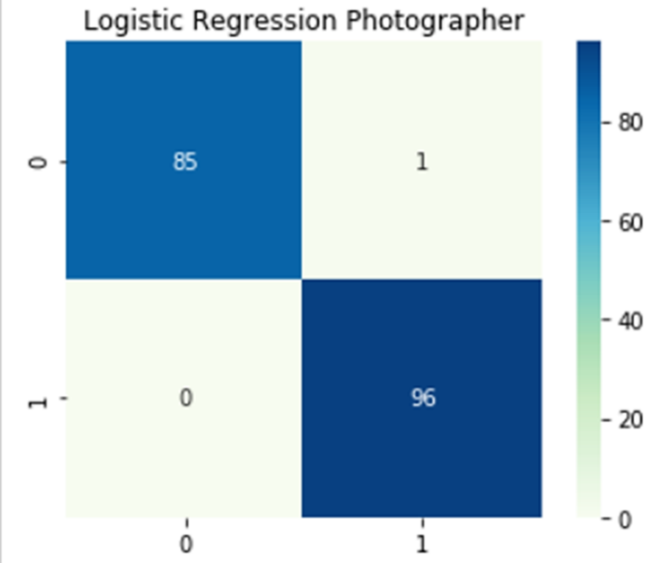
Accuracy = 123+125/123+4+0+125

= 98.41%

Hence the accuracy obtained is 98.41%.

Photographer (Logistic Regression):

Initially there were many outliers in the dataset, which led to drop down of accuracy. Later proper statistical analysis was done by quantile method. Hence the model outperformed and achieved the accuracy of 99.45%.



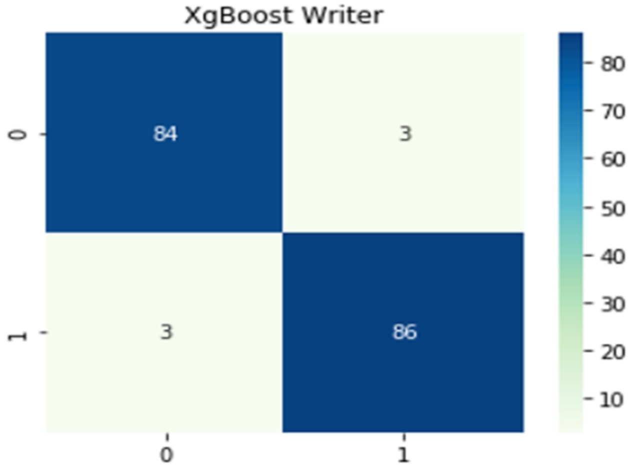
Accuracy = 85+96/85+1+0+96

= 99.45%

Hence the accuracy obtained is 99.45%.

Writer (XgBoost):

An Ensemble technique was used for model building of Writer. Boosting algorithms performed better than all other algorithms. Accuracy was calculated using confusion matrix as shown below



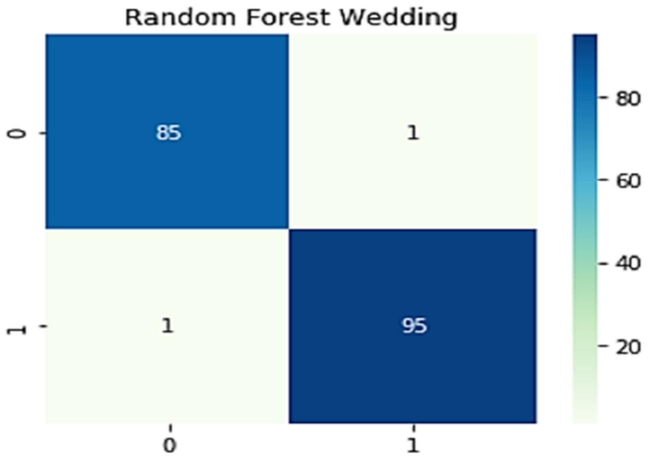
Accuracy = 84+86/84+3+3+86

= 96.59%

Hence the accuracy obtained is 96.59%.

Wedding Planner (Random Forest):

Random forest Classifier, an ensemble learning method was also trained to classify the customer data. Firstly, the model was trained using six best features and the accuracy achieved was 89.24%. Later all the features were added and Confusion matrix was drawn for calculating the accuracy is shown in Fig. (23)



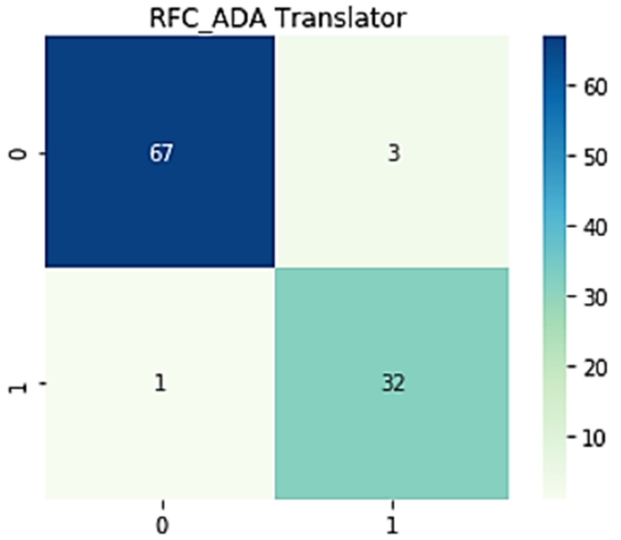
Accuracy = 85+95/85+1+1+96

= 98.90%

Hence the accuracy obtained is 98.90%.

Translator (RfcAda):

RfAda is a hybrid algorithm with Random Forest Classifier and AdaBoost as the base algorithm. Stacking an ensemble technique is used to combine the prediction from Random Forest and AdaBoost. On evaluating the performance of the algorithm, the accuracy obtained was 96.12.



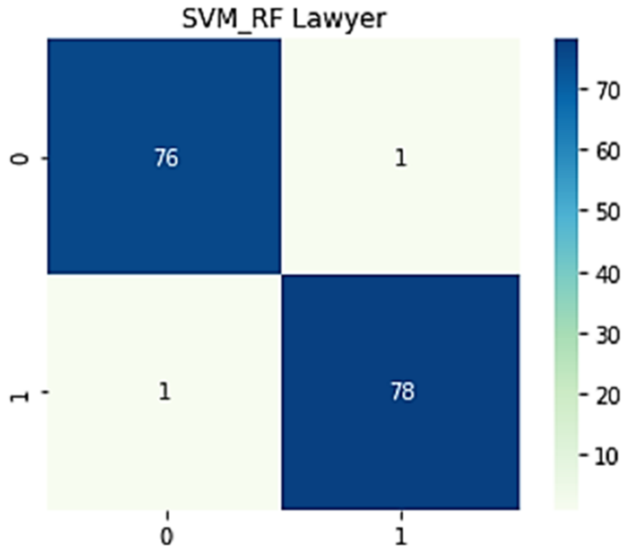
Accuracy = 67+32/67+3+1+32

= 96.12%

Hence the accuracy obtained is 96.12%.

Lawyer (Svm_Rfc):

Svm_Rfc is a hybrid algorithm with Random Forest Classifier and Support Vector Machine as the base algorithm. Stacking an ensemble technique is used to combine the prediction from Random Forest and Support Vector Machine. On evaluating the performance of algorithm, the accuracy obtained was 98.72



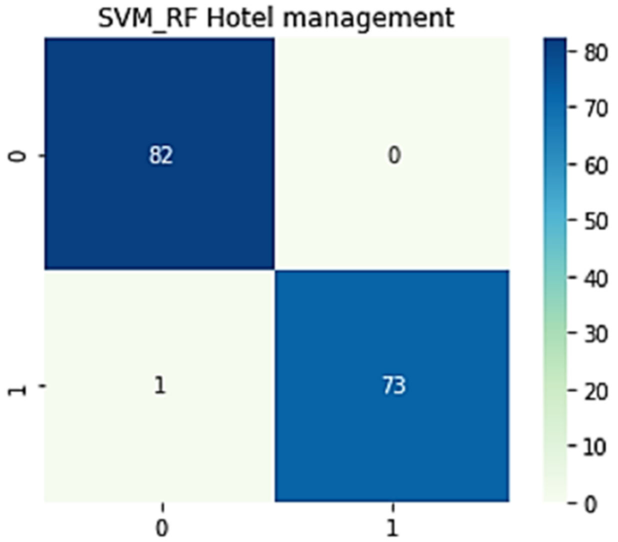
Accuracy = 76+78/76+1+1+78

= 98.72%

Hence the accuracy obtained is 98.72%.

Hotel Management:

Svm_Ada is a hybrid algorithm with Support Vector Machine and AdaBoost as the base algorithm. Stacking an ensemble technique is used to combine the prediction from Support Vector Machine and AdaBoost. On evaluating the performance of the algorithm, the accuracy obtained was 99.36%.



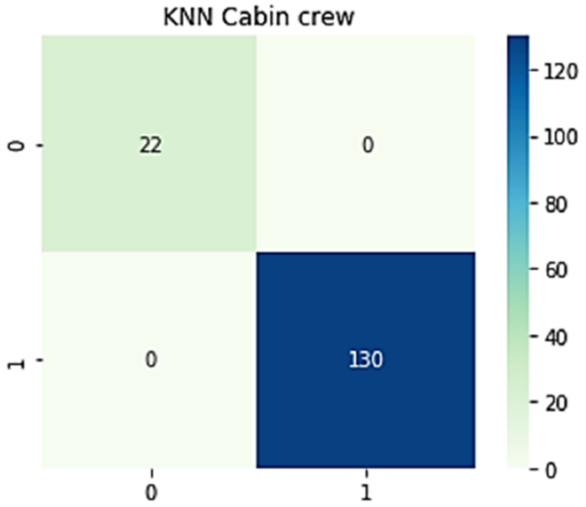
Accuracy = 82+73/82+0+1+73

= 99.36%

Hence the accuracy obtained is 99.36%.

Cabin Crew (KNN):

To prepare the k-NN model dataset is divided into a training set (80%) and test set (20%). On evaluating the performance of the model, it was found that the model achieved the accuracy of 100%. Accuracy was calculated using Confusion matrix as shown below:



Accuracy = 22+130/22+0+0+130

= 100%

Hence the accuracy obtained is 100%.

VIII. CONCLUSION

In this paper, we have studied students career choice based on their interest and most importantly the skillset they have. Furthermore, the study has offered several significant insights for improving the model. Choosing a right career option plays important role for an individual. So having the good skillset related to that career is very important. Observations showed the student having interest in particular career contribute only 50% of his success.

But having proper skillset and capacity to do that work contributes rest of the 50%. Although, the analysis is done on the most of the important machine learning algorithms but a combination of new hybrid algorithms like SvmAda, RfcAda and SvmRfc showed the great result. Observations showed new hybrid algorithms and new dataset with some more instances may create an impact in the enhancement of the model accuracy.

IX. REFERENCES

[1]. Ade R. and Deshmukh P. R. (2014). Classification of Students Using psychometric tests with the help of Incremental Naive Bayes Algorithm. International Journal of Computer Applications. (0975 – 8887) Volume 89 – No 14.

[2]. Subahi A., F. (2018). Data Collection for Career Path Prediction Based on Analyzing Body of Knowledge of Computer Science Degrees. Journal of Software. Volume 13.

[3]. Liu Y., Zhang L., Nie L., Yan Y., Rosenblum D. S. (2016). Fortune Teller: Predicting Your Career Path. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16).

[4]. Uhler B. D., Hurn J. E. (2013). Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective. Journal of Interactive Online Learning. Volume 12.

[5]. Nie M., Xiong Z., Zhong R., Deng W., Yang G. (2020). Career Choice Prediction Based on Campus Big Data—Mining the Potential Behavior of College Students. Applied science. a. Doi: 10.3390/app10082841.

[6]. Badarenah A. A., Alsakran J. (2016). An Automated Recommender System for Course Selection. International Journal of Advanced Computer Science and Applications, Vol. 7, No. 3.

[7]. Gorad N., Zalte I., Nandi A., Nayak D. (2017). Career Counselling Using Data Mining. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 5, Issue 4.

[8]. Chaudhary D., Prajapati H., Rathod R., Patel P., Gurjwar R. K. (2019). Student Future Prediction Using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. Volume 5, Issue 2.

[9]. Mourya V., Phatale S., Thakur S., Mane P. (2020). Career Guide Application using ML. International Research Journal of Engineering and Technology (IRJET). Volume: 07 Issue: 09.

[10]. Prasanna L., Haritha D. (2019). Smart Career Guidance and Recommendation System. International Journal of Engineering Development and Research. Volume 7, Issue 3.

[11]. Roy K. S., Roopkanth K., Uday V., Bhavana V., Priyanka J. (2018). Student Career Prediction Using Advanced Machine Learning Techniques. International Journal of Engineering & Technology.

[12]. Umar M. A. (2019). Student Academic Performance Prediction using Artificial Neural Networks: A Case Study. International Journal of Computer Applications (0975 – 8887) Volume 178.

[13]. Ezenkwu C.P., Johnson E.H., Jerome O.B. (2017). Automated Career Guidance Expert System Using Case-Based Reasoning Technique. Cisd ijournal. Volume 8, No. 1.

[14]. Amieh E. A., Hamtini T., Aljarah I. (2016). Mining Educational Data to Predict Student's Academic Performance using Ensemble Methods. International Journal of Database Theory and Application. doi.org/10.14257/ijdta.2016.9.8.13.

[15]. Elayidom S., Idikkula S. M., Alexander J. (2009). Applying Data mining using Statistical Techniques for Career Selection. International Journal of Recent Trends in Engineering, Vol. 1, No. 1.

[16]. Nawaz M., Adnan A., Tariq U., Salman J. F., Asjad R., Tamoor M. (2014). Automated Career Counseling System for Students using CBR and J48. Journal of Applied Environmental and Biological Sciences.

[17]. Anthony V., Naidoo. (1998). Career Maturity: A Review of Four Decades of Research. Educational Resources Information Centre (ERIC).

[18]. Buhlmann P. (2012). Bagging, Boosting and Ensemble Methods. ETH Zurich, Seminar fur Statistik, HG G17, CH-8092 Zurich, Switzerland.

[19]. Miškovic V. (2014). Machine Learning of Hybrid Classification Models for Decision Support. SINTEZA.

[20]. Vafeiadis T., Diamantaras K.I., Sarigiannidis G., Chatzisavvas K.CH. (2015). A comparison of machine learning techniques for customer churn prediction. Elsevier.

[21]. Dawood E.ABD.E., Elfakhrany E., Maghraby F.A. (2017). Improve profiling bank customer's behavior using machine learning. IEEE ACCESS.