International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1952300

Student Future Prediction Using Machine Learning

*Dileep Chaudhary, Harsh Prajapati, Rajan Rathod, Parth Patel, Rajiv Kumar Gurjwar
Department of Computer Science and Engineering, Parul University, Vadodara, Gujarat, India

ABSTRACT

Selecting an appropriate career is one of the most important decisions and with the increase in the number of career paths and opportunities, making this decision have become quite difficult for the students. According to the survey conducted by the Council of Scientific and Industrial Research's (CSIR), about 40% of students are confused about their career options. This may lead to wrong career selection and then working in a field which was not meant for them, thus reducing the productivity of human resource. Therefore, it is quite important to take a right decision regarding the career at an appropriate age to prevent the consequences that results due to wrong career selection. This system is a web application that would help students studying in high schools to select a course for their career. The system would recommend the student, a career option based on their personality trait, interest and their capacity to take up the course.

Keywords : Career Prediction, Machine Learning, Academic Performance, Linear Regression, Decision Tree Regression

I. INTRODUCTION

With the increase in research and exploration in various domains, there are many new career options in every field. This creates more confusion to the students studying in tenth or twelth grade to select one career option. The reasons for this confusion could be unawareness of self-talent and self-personality trait, unawareness of the various options available, equal interests in multiple fields, less exposure, market boom, assumed social life, peer-pressure etc. Due to theses confusions, the student 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.

Thus, there should be proper counseling of the student's Academic Performance, interest and their capacity to work in a particular field.

II. RELATED WORK

There are various websites and web apps over the internet which helps students to know their suitable career path. But most of those systems only used personality traits as the only factor to predict the career, which might result in an inconsistent answer. Similarly, there are few sites that suggest career based on only the interests of the students. The systems did not use the capacity of the students to know whether they would be able to survive in that field or not.

The paper by [1] Beth Dietz-Uhler & Janet E. Hurn suggest the importance of learning analytics in predicting and improving the student's performance which enlightens the importance of student's interest, ability, strengths etc. in their performance. According to the paper by [2] Lokesh Katore , Bhakti Ratnaparkhi , Jayant Umale , the career prediction

accuracy was determined using 12 attributes of students and different classifiers with c4.5 having the highest accuracy of 86%. [3] Another paper by Roshani Ade, P.R.Deshmukh suggested incremental ensemble of classifiers in which the hypothesis from number of classifiers were experimented and by using 'Majority voting rule', the final results was determined. The proposed ensemble algorithm gave an accuracy of 90.8% [4].The paper by Mustafa Agaoglu suggested the importance of different attributes in evaluating the performance of faculty. It also showed the comparison of different classifiers proved that the most accurate classifier was c5.0 which has the maximum attribute usage compares to other classifiers like CART, ANN-Q2H, SVM etc. Also, the suggestions provided by the system are very much generalized and not specific to a university or country/state. The suggestion for course is also generalized. For example, the results of few systems were a group of courses like data analyst, accountant, law etc. Thus, if a student gets such a recommendation then he/she might again get confused as the above specified course belongs to different streams.

III. OVERVIEW

The project was to develop a web application that can be used by any student who needs help in selecting the career path. The system displays questionnaire to the students which the student will have to answer. The three set of questionnaires provided to the students based on Academic Performance, interests and capacity.

3.1 Academic's Information: Academic Information are includes student's percentage, subject marks, extra curricular activities and so on.

- SSC Medium
- HSC Medium
- SSC Percentage
- HSC Percentage
- Subject Marks

3.2 Interest: Interest in this context implies that how much a student likes a subject and is keen to learn about it. Here the student will be first asked about the basic 3 streams i.e. Arts, Science and Commerce. The system then checks in which stream the student is more interested and then the further questions are comparison based questions which compare the subject of the selected stream and at the end determines one particular subject that the student is interested in.

- interested in games
- Interested Type of Books

3.3 Capacity: Capacity implies that how efficiently a student can learn their interested subject and survive in that particular career path. For this purpose, the student will get questions that they had in their school curriculum and each question will have 4 options and a timer associated with it. Here the system asses not only the correctness of the answer but also the speed of the student to answer the question. This helps in knowing the memory, ability to solve and grasping capacity of the student.

- self-learning capability
- memory capability score
- Hours working per day etc.

IV. ALGORITHM USED

4.1 Data Collection

Data collection is the systematic approach to gathering and measuring information from a variety of sources to get a complete and accurate picture of an area of interest. Data collection enables a person or organization to answer relevant questions, evaluate outcomes and make prediction about future probabilities and trends.

4.2 Data Pre-Processing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

4.3 Exploratory Data Analysis

Scatterplot and Histograms

We will start by creating a scatterplot matrix that will allow us to visualize the pair-wise relationships and correlations between the different features. It is also quite useful to have a quick overview of how the data is distributed and whether it contains or not outliers.

Correlation Matrix

We have created a correlation matrix to quantify and summarize the relationships between the variables. This correlation matrix is closely related with covariance matrix, in fact it is a rescaled version of the covariance matrix, computed from standardize features. It is a square matrix (with the same number of columns and rows) that contains the Person's r correlation coefficient.

4.4 Algorithm's

4.4.1 Regression

Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

4.4.2 Linear Regression

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. The red line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation shown below.

$y = a\_0 + a\_1 * x$ ## Linear Equation

The motive of the linear regression algorithm is to find the best values for a_0 and a_1. Before moving on to the algorithm, let's have a look at two important concepts you must know to better understand linear regression.

4.4.3 Decision Tree Regression

The decision tree is a simple machine learning model for getting started with regression tasks. A decision tree is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node. Different kinds of models have different advantages. The decision tree model is very good at handling tabular data with numerical features, or categorical features with fewer than hundreds of categories. Unlike linear models, decision trees are able to capture non-linear interaction between the features and the target.

4.4.4 Random Forest Algorithm

The Random Forest is one of the most effective machine learning models for predictive analytics, making it an industrial workhorse for machine learning. The **random forest** model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of

models as: where the final model is the sum of simple base models. Here, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called **model ensembling**. In random forests, all the base models are constructed independently using a **different subsample** of the data.

V. IMPLEMENTATION

Step 1. The first step of implementation was to collect data from students studying in different courses.For this purpose, an online survey was conducted using Google forms. The questions asked in the survey are based on personality traits. Based on the answers the personality type of the student will be determined which will be one among the 16 types. Then it was found that which courses were selected by which personality trait.

```
df.head()
```

	NAME	GENDER	ADMISSION TYPE	CASTE	NATIONALITY	RELIGION	BLOOD GROUP	IS_HANDICAPED	SSC MEDIUM	HSC MEDIUM	...	ECONOMICS	BUS.STUD
0	JAHNVI PATEL	F	ACPC	OPEN	INDIAN	CHRISTIAN	A+	NO	GUJARATI	GUJARATI	...	NaN	I
1	CHAUDHARI MILAN JAYESHBHAI	M	ACPC	ST	INDIAN	CHRISTIAN	B+	NO	GUJARATI	ENGLISH	...	NaN	I
2	ACHARYA PRATIK BHASMANG	M	MQ	OPEN	INDIAN	HINDU	A+	NO	GUJARATI	ENGLISH	...	NaN	I
3	SHAH SHREY AMIYA	M	NRI	OPEN	INDIAN	HINDU	O+	NO	HINDI	HINDI	...	NaN	I
4	SINHA VAIBHAV BIRENDRAKUMAR	M	ACPC	SC	INDIAN	HINDU	A+	NO	GUJARATI	GUJARATI	...	NaN	I

5 rows × 43 columns

Figure 1. Collected Dataset

Step 2. Then data obtained from the survey had to pre-processed and consolidated into a common format as required by the system.

```
In [56]: from sklearn.preprocessing import LabelEncoder, OneHotEncoder

In [57]: labelencoder = LabelEncoder()

In [59]: for i in range(2,14):
             data[:,i] = labelencoder.fit_transform(data[:,i])
         data[:5]

Out[59]: array([[1, 150000000000.0, 414, 699, 468, 1, 0, 1, 0, 0, 1, 4, 0, 2,
                 9423645461, 203422, 2013, 96, 2014, 42, 67, 85, 62, 57, 62.6,
                 2015, 49007, 122, 335766],
                [2, 150000000000.0, 673, 465, 686, 1, 0, 0, 0, 0, 0, 2, 1, 0, 0,
                 9084037710, 163239, 2013, 76, 2015, 41, 91, 97, 65, 68, 72.4,
                 2015, 42727, 135, 677588],
                [3, 150000000000.0, 299, 766, 353, 0, 0, 0, 0, 0, 3, 3, 0, 0,
                 9975379122, 437269, 2013, 78, 2015, 42, 71, 93, 92, 92, 78.0,
                 2015, 33089, 238, 582633],
                [4, 150000000000.0, 649, 685, 701, 0, 0, 1, 0, 2, 1, 0, 0, 1,
                 9753272232, 467265, 2012, 92, 2015, 43, 86, 61, 56, 78, 64.8,
                 2015, 36929, 277, 565494],
                [5, 150000000000.0, 182, 511, 191, 1, 0, 0, 0, 1, 1, 1, 0, 1,
                 9211564257, 166387, 2013, 79, 2015, 77, 36, 46, 68, 54, 56.2,
                 2015, 34599, 318, 882713]], dtype=object)

In [60]: data[15:-1]

Out[60]: array([[16, 150000000000.0, 129, ..., 20246, 143, 889346],
                [17, 150000000000.0, 683, ..., 7680, 177, 631040],
                [18, 150000000000.0, 651, ..., 43824, 60, 833228],
                ...,
```

Figure 2. Pre-processed Data

Step 3. The dataset was then used to derive the decision tree for various courses. Using the C5 package in R, the algorithm was applied on the dataset for different courses. This figure 2 shows the people with thinking and perceiving personalities can take engineering.

This way, the algorithm was applied other courses and the personality types that can take up those courses were determined.

Step 4. To improve the accuracy of the prediction, adaptive boosting was applied on the algorithm.

Step 5. After completing the entire backend decision tree generations, the next step was to develop the web application. The application was developed using python. The database was stored and processed using the MySQL server.

VI. CONCLUSION

This system will help student used to predict the suitable course. This system helps to minimize the failure ratio and to take acceptable action for career. This system can facilitate the students, as it will guide them to take appropriate decision while choosing the stream as his/her career. This system will also help the college to analyze the admissions branch and take the necessary actions depending upon the result.

VII. REFERENCES

[1]. UD Beth, HE Janet, Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective. Journal of Interactive Online Learning 2013; 12:17-26.

[2]. KS Lokesh, RS Bhakti, et al. Novel Professional career prediction and recommendation method for individual through analytics on personal traits using C4.5 algorithm. IEEE Communication Technology (GCCT) on 3 December 2015.

[3]. A Roshani, PR Deshmukh, An incremental ensemble of classifiers as a technique for prediction of student's career choice. IEEE Networks & Soft Computing (ICNSC) on 25 September 2015.

[4]. A Mustafer, Predicting Instructor performance using data mining technique in higher education. IEEE 2016; 4:2379-2387.

[5]. C Ling, R Dymitr, et al. Big Data: Opportunities for Big Data Analytics. IEEE Digital Signal Processing (DSP) on 10 September 2015.

[6]. M Yannick, X Jie, et al. Predicting Grades. IEEE transactions on signal processing on 15 February 2016

[7]. M Jiri, S Jan, AdaBoost. Centre for Machine Perception, Czech Technical University, Prague.

Cite this article as :

Dileep Chaudhary, Harsh Prajapati, Rajan Rathod, Parth Patel, Rajiv Kumar Gurjwar, "Student Future Prediction Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 1104-1108, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT1952300

Journal URL : http://ijsrcseit.com/CSEIT1952300