STUDENT CAREER PREDICTION

Vidyapriya.C,Vishhnuvardhan.R.C

Email:vidyapriya.1601266@srec.ac.in

ABSTRACT

As students are going through their academics and pursuing their interested courses, it is very important for them to assess their capabilities and identify their interests so that they will get to know in which career area their interests and capabilities are going to put them in. This will help them in improving their performance and motivating their interests so that they will be directed towards their targeted career and get settled in that. Here we used three algorithm logistic regression, Gaussian navie bayes, svm(support vector machine)for classification and prediction to predict career hobbies,intrest,sports,acheivements,academic performance will be collected based on that answer the career of the student will be predicted for the school students.

Keywords: capabilities, logistic regression, svm , classification.

1.INTRODUCTION

Academic performance of students has always been a major factor for determining the student's career and the prestige of the Institutions. So as to compete and reach the goal, students need to be planned and organized from initial stages of their education. So it is very important to constantly evaluate their performance, identify their interests and evaluate how close they are to their goal and asses whether they are in the right path that directs towards their targeted. This helps them in improving themselves, motivating themselves to a better career path if their capabilities are not up to the mark to reach their goal and pre evaluate themselves before going to the career peek point.

Machine Learning is a technique where the machines are trained in such a way that it gains the ability to respond to a particular input or scenario based on the previous inputs it has learnt. Simply it the giving computers the ability to learn by using statistical techniques. Machine learning helps the computers to act without explicitly being programmed. This aims at reducing the human intervention in the machine dependable problems and scenarios. This helps in solving very complex tasks and problems very easily and without involving much human labor. Various applications of machine learning include classification,prediction, image recognition, medical diagnosis, algorithm building, self driving cars and much more. Majority of problems in machine learning can be solved using supervised andunsupervised learning. And if the final output classes and sets are not known and it is done by identifying the similarity between data point and their characteristics and finally they are made into groups based on these characteristics then it is called un-supervised.

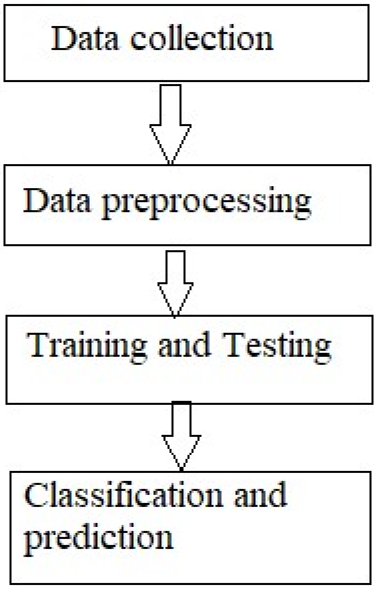
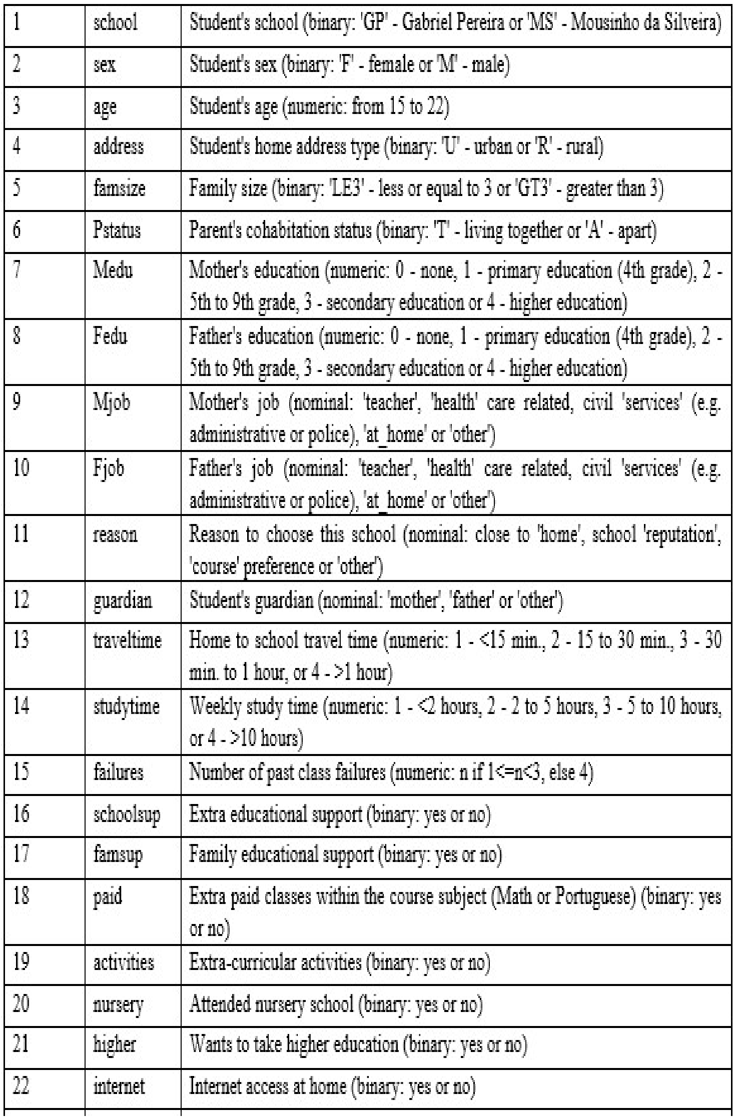


Figure 1: process flow diagram

2. IMPLEMENTATION

2.1 DATA COLLECTION

Collection of data is one of the major and most important tasks of any machine learning projects. Because the input we feed to the algorithms is data. So, the algorithms efficiency and accuracy depends upon the correctness and quality of data collected. So as the data same will be the output. For student career prediction many parameters are required like students academic scores in various subjects, specializations, programming and analytical capabilities, memory, personal details like relationship, interests, sports, competitions, workshops, certifications, books interested and many more. As all these factors play vital role in deciding student's progress towards a career area, all these are taken in-to consideration. Data is collected in many ways. Some data is collected from employees working in different schools, some amount of data is randomly generated and other from school database.



1	school	Student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2	sex	Student's sex (binary: 'F' - female or 'M' - male)
3	age	Student's age (numeric: from 15 to 22)
4	address	Student's home address type (binary: 'U' - urban or 'R' - rural)
5	famsize	Family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6	Pstatus	Parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7	Medu	Mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
8	Fedu	Father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
9	Mjob	Mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10	Fjob	Father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11	reason	Reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12	guardian	Student's guardian (nominal: 'mother', 'father' or 'other')
13	traveltime	Home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14	studytime	Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15	failures	Number of past class failures (numeric: n if 1<=n<3, else 4)
16	schoolsup	Extra educational support (binary: yes or no)
17	famsup	Family educational support (binary: yes or no)
18	paid	Extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19	activities	Extra-curricular activities (binary: yes or no)
20	nursery	Attended nursery school (binary: yes or no)
21	higher	Wants to take higher education (binary: yes or no)
22	internet	Internet access at home (binary: yes or no)

2.2 DATA PRE-PROCESSING:

Collecting the data is one task and making that data useful is an-an-another vital task. Data collected from various means will be in an unorganized format and there may be lot of null values, in-valid data values and unwanted data. Cleaning all these data and replacing them with appropriate or approximate data and removing null and missing data and replacing them with some fixed alternate values are the basic steps in preprocessing of data. Even data collected may contain completely garbage values. It may not be in exact format or way that is meant to be. All such cases must be verified and replaced with alternate values to make data meaning meaningful and useful for further processing. Data must be kept in a organized format.

2.3 TRAINING AND TESTING:

After processing of data and training the very next task is obviously testing. This is where performance of the algorithm, quality of data, and required output all appears out. From the huge data set collected 80 percent of the data is utilized for training and 20 percent of the data is reserved for testing. Training as discussed before is the process of making the machine to learn and giving it the capability to make further predictions based on the training it took. Whereas testing means already having a predefined data set with output also previously labeled and the model is tested whether it is working properly or not and is giving the right prediction or not. If maximum number of predictions is right then model will have a good accuracy percentage and is reliable to continue with otherwise better to change the model.

2.4 CLASSIFICATION AND PREDICTION:

The classification is done with the help of logistic regression, naïve Bayes, support vector machine (SVM) with the parameters of the student intrest, hobbies and their performance in the academic. the prediction is done with the help of data collected from various students in the respective way.

3.MACHINE LEARNING ALGORITHM:

3.1 Logistic Regression:

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. Logistic regression has become an important tool in the discipline of machine learning.Logistic regression can also play a role in data preparation activities by allowing data sets to be put into specifically predefined buckets during the extract, transform, load (ETL) process in order to stage the information for analysis shown in figure 2

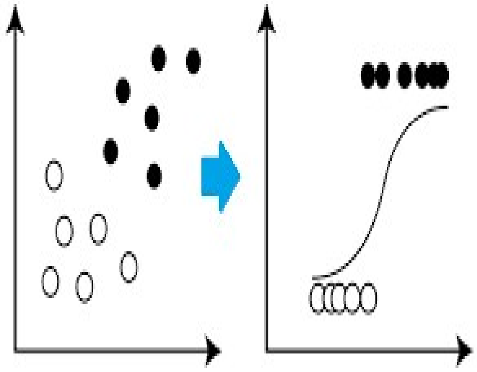


Figure 2: Logistic regression

Logistic regression can also be used in

- Healthcare to identify risk factors for diseases and plan preventive measures.
- Weather forecasting apps to predict snowfall and weather conditions.
- Voting apps to determine if voters will vote for a particular candidate.
- Insurance to predict the chances that a policy holder will die before the term of the policy expires based on certain criteria, such as gender, age and physical examination.
- Banking to predict the chances that a loan applicant will default on a loan or not, based on annual income, past defaults and past debts.

3.2 NAIVE BAYES:

It is a classification technique based on Naïve Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.Naive Bayes is a simple but surprisingly powerful algorithm for predictive modeling. The model is comprised of two types of probabilities that can be calculated directly from your training data: 1) The probability of each class; and 2) The conditional probability for each class given each x value. Once calculated, the probability model can be used to make predictions for new data using Bayes Theorem.

THEOREM:P(c/x)=P(x/c)P(c)/P(x)

3.3 SVM:

SVM denotes Support Vector Machine. It is a supervised machine learning algorithm which is generally used for both regression and classification type of problems. The typical procedure of the algorithm is first each data item is plotted in a n dimensional space, where n is the number of features and the value of each feature being the value of that particular coordinate shown in figure1.1.3

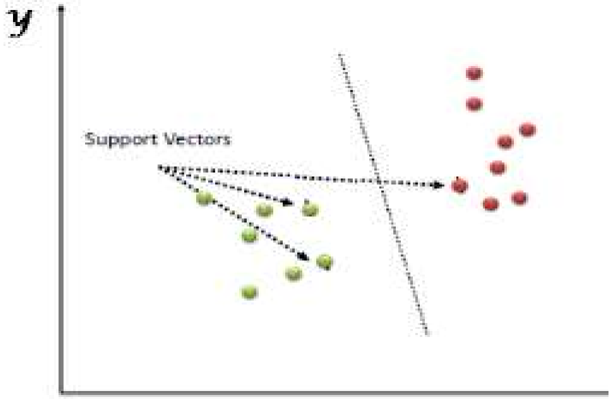


Figure 3: Support Vector Machines Example

SVM algorithms practically are implemented using kernels. There are three types of SVM's and in linear SVM hyperplane is calculated or found by transforming the problem using linear algebra. The insight is that SVM can be rephrased by using the inner product of two observations. The sum of the multiplication of each pair of inputs is called inner product of two vectors. The equation for dot product of a input xi and support vector xi is:

$$f(x) = B0 + sum(ai * (x,xi)). \tag{1}$$

Instead of using the dot-product, a polynomial kernel can be used, for example:

$$K(x,xi) = 1 + sum(x * xi)^d \tag{2}$$

And not only that a more complex radio kernel is also there. The general equation is:

$$K(x,xi)=exp(-gamma*sum((x– xi^2)) \tag{3}$$

4. RESULT :

The data is trained and tested with all three algorithms and out of all logistic regression gave more accuracy with 85.3 percent .As logistic regression gave the highest accuracy, all further data predictions are chosen to be followed with logistic regression. So, finally a web application is made to give the input parameters of the student and the final prediction is generated and displayed.The background algorithm being used is logistic regression and the new prediction are keep on adding to the dataset for further more accuracy.

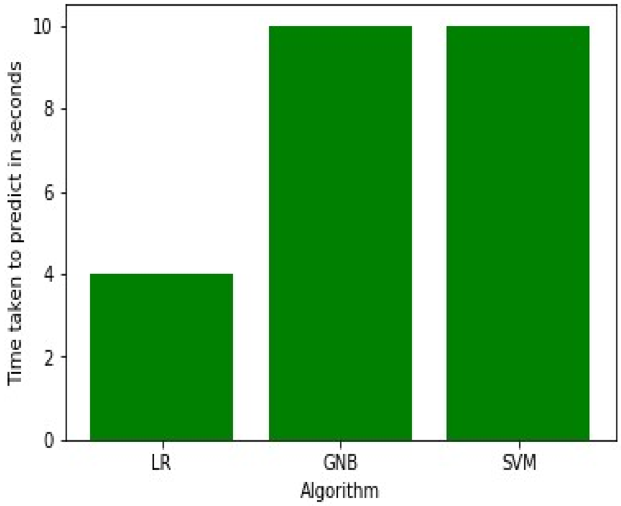


Figure 4:Result Graph

5. REFERENCES:

[1] Ali Daud, Naif Radi Aljohani, “Predicting Student Performance using Advanced Learning Analytics”, 2017 International World Wide Web Conference Committee (IW3C2). Volume 13-No.6 July 2016.

[2] Anuj Karpatne, Gowtham Atluri, “Theory- Guided Data Science: A New Paradigm for Scientific Discovery from Data”, IEEE trans-actions on knowledge and data engineering, vol.29, no. 10, october 2017.

[3] Bo Guo , Rui Zhang, “Predicting Students Performance in Educa-tional Data Mining”, International Symposium on Educational Technology Vol.143, No.8 Augest 2016.

[4] P.KaviPriya, “A Review on Predicting Students’ Academic Performance Earlier, Using Data Mining Techniques”, International Jour-nal of Advanced Research in Computer Science and Software Engineering Volume 3-No .9, May 2017.

[5] Mahendra Tiwari ,Manmohan Mishra, “Accuracy Estimation of Classification Algorithms with DEMP Model”, International Jour-nal of Advanced Research in Computer and Communication Engineering Vol. 2, No.11, November 2016.

[6] Marium-E-Jannat,SaymaSultana,Munira Akther, “A Probabilistic Machine Learning Approach for Eligible Candidate Selection”, Inter-national Journal of Computer Applications (0975 – 8887)Volume 144 – No.10, June 2016.

[7] Nikita Gorad ,Ishani Zalte, “Career Counselling Using Data Mining”, International Journal of Innovative Research in Computer and Communication Engineering. Vol.6,No.18 Jan2015.

[8]Roshani Ade,Dr. P. R. Deshmukh, “An incremental ensemble of classifiers as a technique for prediction of student’s career choice”, First International Vol. 5,No.13 Dec 2015.

[9] Rutvija Pandya Jayati Pandya , “C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning“, Inter-national Journal of Computer Applications (0975 – 8887) Volume 117 – No. 16, May 2015.

[10] Sudheep Elayidom, Dr. Sumam Mary Idikkula, “Applying Data mining using Statistical Techniques for Career Selection”, International Journal of Recent Trends in Engineering, Vol. 1, No. 1, May 2009.