Survey on Techniques for Predictive Analysis of Student Grades and Career

Sagar More[1], Pravin Bhagwat[2], Indrajeet Karande[3], Sayaji Dhandge[4], Sachin Shinde[5]

[1]Student, Department of Computer Engineering, PDEA's College of Engineering, Pune, Maharashtra, India
[2]Student, Department of Computer Engineering, PDEA's College of Engineering, Pune, Maharashtra, India
[3]Student, Department of Computer Engineering, PDEA's College of Engineering, Pune, Maharashtra, India
[4]Student, Department of Computer Engineering, PDEA's College of Engineering, Pune, Maharashtra, India
[5]Assistant Professor, Department of Computer Engineering, PDEA's College of Engineering, Pune, Maharashtra, India

---***---

Abstract - *In recent years, predictive analytics has seen a surge in popularity, with many organizations using it to make decisions about everything from product development to marketing campaigns. The education sector is no exception, with many schools and universities using predictive analytics to identify at-risk students and improve retention rates. This survey paper reviews state of art in predictive analytics for student grades and career outcomes. The survey begins by discussing the different data types that can be used for predictive modeling, including demographic data, academic performance data, and social media data. Then this article reviews a few techniques used for predictive modeling in the education domain, including logistic regression, decision trees, and neural networks. Finally, this article discusses some of the challenges associated with predictive analytics in education and suggests future directions for research.*

***Key Words*: student grade, career, machine learning, survey, svm, knn, j48, naïve bayes, linear regression, random forest, gradient boosting technique, xg boost, bayesian ridge regression**

1. INTRODUCTION

In recent years, predictive analytics has become an increasingly popular tool for educators, administrators, and policymakers to use to make data-driven decisions about students' grades and careers. Predictive analytics is data mining that uses statistical techniques to predict future events or outcomes. In education, predictive analytics has been used to forecast everything from student retention and success rates to job placement and earnings. Various techniques can be used for predictive analytics, and the choice of method depends on the type of data available and the specific question being asked. Some standard methods include regression analysis, decision trees, and artificial neural networks. This survey paper will review the literature on predictive analytics in education, focusing on techniques for predicting student grades and career outcomes. We will first provide an overview of the history and applications of predictive analytics in education. Next, we will discuss some of the most used methods for predictive analytics. Finally, we will discuss some challenges and limitations of predictive analytics in education.

2. LITERATURE SURVEY

[1] Siti Dianah Abdul Bujang, Ali Selamat, Roliana Ibrahim, Ondrej Krejcar, Enrique Herrera-viedma, Hamido Fujita, And Nor Azura Md. Ghani (2021): In this article, the authors propose a multiclass prediction model with six predictive models to predict final students' grades. The model is based on the previous students' final examination results of the first-semester course. The article does a comparative analysis of combining oversampling SMOTE with different FS methods to evaluate the performance accuracy of student grade prediction.

[2] Arati Yashwant Amrale, Namrata Deepak Pawshe, Nikita Balu Sartape, Prof. Komal S. Munde (2022): This article proposes a counseling system that uses artificial intelligence to help with career guidance.

[3] Vidyapriya.C, Vishhnuvardhan.R.C: In this article, the authors trained and tested three algorithms: logistic regression, Naive Bayes, and Support Vector Machine. They found that logistic regression had the highest accuracy compared to the other two algorithms.

[4] Prathamesh Gavhane, Dhanraj Shinde, Ashwini Lomte, Naveen Nattuva, Shital Mandhane (2021): In this article, authors have analyzed most machine learning algorithms for student career prediction. They found that combining new hybrid algorithms like SvmAda, RfcAda and SvmRfc showed excellent results.

[5] N. Vidyashreeram, Dr. A. Muthukumaravel: In this article, authors have used machine learning approaches such as Adaboost, SVN, RF, and DT to predict students' careers and have found that RF produces the best results in terms of accuracy.

[6] Zafar Iqbal, Junaid Qadir, Adnan Noor Mian, And Faisal Kamiran: In this article, authors have discussed the use

of Collaborative Filtering (UBCF), Matrix Factorization (MF), and Restricted Boltzmann Machine (RBM) techniques for predicting a student's Grade Point Average (GPA). They have used the RBM machine learning technique to predict a student's course performance. Empirical validation on a real-world dataset shows the effectiveness of the RBM technique.

[7] Hana Bydžovská: In this article, the author has presented two different approaches. The first approach used widely used classification and regression algorithms, with SVM reaching the best results. This approach can be beneficially utilized for the grade prediction of courses with a small number of students. The second novel approach used collaborative filtering techniques and predicted grades based on the similarity of students' achievements. The advantage of this approach was that each university information system stores the data about students' grades needed for the prediction, unlike the data about students' social behavior.

[8] Prakash, Mr. Sachin Garg (2021): In this article, after evaluating all the algorithms like Linear Regression, Random Forest, Gradient Boosting Regression, Bayesian Ridge Regression, etc., on different parameters, the authors have proposed a model that can predict the grade more accurately using the gradient boosting regression algorithm.

[9] K. Sripath Roy, K.Roopkanth, V.Uday Teja, V.Bhavana, J.Priyanka (2018): In this article, the authors have trained and tested three algorithms - SVM, XG Boost and Decision Tree - for the career prediction of students. They found that SVM gave more accurate predictions than XG Boost. As SVM gave the highest accuracy, all further data predictions are chosen to be followed with SVM.

[10] Anitha K, Bhoomika C, J Andrea Kagoo, Kruthika K, Aruna Mg (2022): In this article, the authors have proposed a multiclass prediction model with six predictive models to predict the final student's grades based on the previous student's final examination results of the first-semester course. The six predictive models used in this study were Decision Tree, Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbour (kNN), Logistic Regression (LR), and Random Forest (RF).

[11] Anooja S K, Dileep V K (2020): In this article, the authors have found that the machine learning and data mining techniques used in their paper for student career prediction had an accuracy of 87% or lower. They also found that there was a high possibility of misprediction.

3. TECHNIQUES FOR STUDENT GRADE AND CAREER PREDICTION

1) J48

The J48 algorithm is a machine learning decision tree classification algorithm based on the Iterative Dichotomiser 3 algorithm. It is conducive to examining data sets containing both categorical and continuous data.[1]

2) K-Nearest Neighbors

K-nearest neighbours (KNN) is a supervised learning algorithm used for regression and classification purposes, but it is mainly used for the latter. Given a dataset with different classes, KNN tries to predict the correct test class 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. [1,4]

3) Naïve Bayes

Naive Bayes is a classification technique that predicts the class of a new feature set using the Bayesian theorem. This theorem states that the probability of an event occurring is based on its prior probability and the evidence present. In other words, a Naive Bayes classifier assumes that a particular feature in a class is unrelated to the presence of any other feature. [1,3,4]

4) Support Vector Machine

SVM is a supervised machine learning algorithm that uses the concept of support vectors to do the linear separation. It has a clever way of reducing overfitting and can use many features without requiring much computation. It is generally used for both regression and classification types of problems. The main applications of this can be found in various classification problems. The typical procedure of the algorithm is as follows; first, each data item is plotted in an n-dimensional space, where n is the number of features, and the value of each feature is the value of that coordinate. The next step is to classify by getting the hyper-plane that separates the two classes very finely. [1,3,4,9]

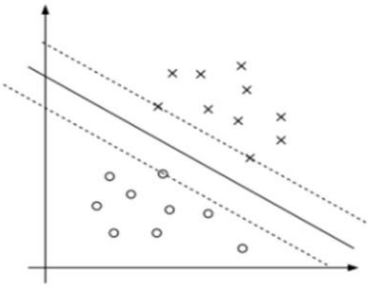

Fig -1: Support Vector Machine [4]

5) Linear Regression

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 into specifically predefined buckets during the extract, transform, load (ETL) process. [1,3,4]

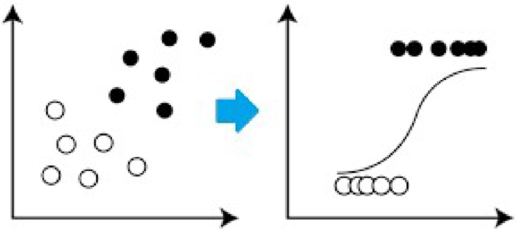

Fig -2: Linear Regression [3]

6) Random Forest

Random Forest is a bagging technique that uses the Decision tree as its base model. It applies feature sampling and row sampling with replacement to feed the data to the base models. [1,4,8]

7) Gradient Boosting Regression

Gradient boosting is a machine learning algorithm where models are trained consecutively. Each new model uses the Gradient Descent approach to reduce the entire system's loss function (y = axe + b + e, where 'e' is the error component). The learning process fits new models in a sequence to provide a more precise estimate of the response variable. Gradient boosting can be used for both regression and classification problems. [8]

8) Bayesian Ridge Regression

Bayesian Regression is a regression algorithm that is useful when there needs to be more data in a dataset or when the data is not evenly distributed. In contrast to traditional regression techniques, which produce an output based on a single attribute value, the Bayesian Regression model's output is derived from a probability distribution. A normal distribution (where mean and variance are parameters estimated from the data) generates the output. [8]

9) Decision Tree

Decision trees are a popular machine learning algorithm, especially for classification problems. They are relatively simple to implement and easy to understand, which makes them a good choice for many applications. Decision trees also form the basis for more advanced algorithms like bagging, gradient boosting, and random forest. [4,9]

10) XG Boost

XGBoost refers to eXtreme Gradient Boosting. It implements gradient boosting algorithms available in many forms, such as a tool, library, etc. XGBoost mainly focuses on model performance and computational time. It can significantly reduce the time and improve the performance of the model. Its implementation has the features of scikit-learn and R implementations and newly added features like regularization. [9]

Table -1: Input Features for Student Grade Prediction

Input Features for Student Grade Prediction			
Attribute	Type	Values	Description
StudID	Nominal	S1s61	Student identification
Year	Numeric	[2016,2019]	Year of student intake
Class	Nominal	DDT1A, DDT1b	Class of student
Session	Nominal	DEC, JUNE	Session of student intake per year
Credit Hour	Numeric	[3]	A credit hour for each course
Course Code	Nominal	[CSA, ICS]	Course ID of 2 courses
Total Marks	Numeric	[38, 91]	Student final marks obtained from the final exam and course assessment
Grade Pointer Average	Numeric	[0.00,4.00]	Student course grade pointer
Grade	Nominal	[A+, A, A-, B, B+, B-, C+, C, C-, D+, D, E, F]	Student's final grade for each course
Group	Nominal	EXCEPTIONAL, EXCELLENT, DISTINCTION, PASS, FAIL	Category of student academic performance

Table -2: Input Features for Student Career Prediction

Input Features for Student Career Prediction			
Attribute	Type	Values	Description
School	Binary	'GP' – Gabriel Pereira	Name of School
Sex	Binary	['F', 'M']	Gender of student
Age	Numeric	15	Age of student
Address	Binary	'U' – Urban, R – Rural	Address of student
FamSize	Binary	'LE3' – Less than equal to 3	Family size of a student
PStatus	Binary	'T' – Together, 'A' – Apart	Parent's cohabitation status
MEdu	Numeric	0 – None, 1 – Primary Education, 2 - 5th to 9th Grade, 3 – Secondary Education, 4 – Higher Education	Mother's education
FEdu	Numeric	0 – None, 1 – Primary Education, 2 - 5th to 9th Grade, 3 – Secondary Education, 4 – Higher Education	Father's Education
MJob	Nominal	'Teacher', 'Heath Care', 'Civil Service', 'At_Home', 'Other'	Mother's job
FJob	Nominal	'Teacher', 'Heath Care', 'Civil Service', 'At_Home', 'Other'	Father's job
Reason	Nominal	'Close to Home', 'School Reputation'	Reason to choose this school
Guardian	Nominal	'Mother', 'Father', 'Other'	Student's Guardian
TraveTime	Numeric	1 - <15 min, 2 – 15 to 30 mins, 3 – 30 mins to 1 hour, 4 - >1 hour	Home to school travel time
StudyTime	Numeric	1 - <15 min, 2 – 15 to 30 mins, 3 – 30 mins to 1 hour, 4 - >1 hour	Weekly study time
Failures	Numeric	N if 1 <=N<3 else 4	Number of past classes failures
SchoolSup	Binary	'Yes' or 'No'	Extra educational support
FanSup	Binary	'Yes' or 'No'	Family educational support
Paid	Binary	'Yes' or 'No'	Extra paid classes within Couse subject
Activities	Binary	'Yes' or 'No'	Extra-curricular activities
Nursery	Binary	'Yes' or 'No'	Attended nursery school
Higher	Binary	'Yes' or 'No'	Wants to take higher education
Internet	Binary	'Yes' or 'No'	Internet access at home
Attribute	Type	Values	Description
School	Binary	'GP' – Gabriel Pereira	Name of School

Table -3: Student Grade Prediction Techniques

Student Grade Prediction Techniques		
Technique	Accuracy	Reference Article
J48	98.9% [1]	[1,8]
K-Nearest Neighbor	98.5%	[1]
Naïve Bayes	98.4%	[1]
Support Vector Machine	98.4%	[1]
Logistic Regression	98.4%	[1]
Random Forest	98.9% [1], 74% [8]	[1,8]
Gradient Boosting Regression	79%	[8]
Bayesian Ridge Regression	69%	[8]

Table -4: Student Career Prediction Techniques

Student Career Prediction Techniques		
Technique	Accuracy	Reference Article
Logistic Regression	85.3% [3], 95.24% [4], 98.41% [4]	[3,4]
Naïve Bayes	98.15% [4]	[3.4]
Support Vector Machine	98.41% [4], 90.3% [9]	[3,4,9]
Decision Tree	97.24% [4]	[4,9]
K-Nearest Neighbor	98.15%,100%	[4]
Random Forest	98.50%	[4]
XG Boost	96.59% [4], 88.33% [9]	[4,9]

4. CONCLUSION

As per the survey conducted on various machine-learning techniques for predicting student grades and careers, based on all observations, this article concludes that RF, kNN, and J48 are effective techniques for predicting student grades with an accuracy of 98.9%, 98.8%, and 98.9%, respectively. Additionally, this article finds that LR and SVM are effective techniques for predicting student careers with an accuracy of 98.41%.

REFERENCES

[12] Siti Dianah Abdul Bujang, Ali Selamat, Roliana Ibrahim, Ondrej Krejcar, Enrique Herrera-viedma, Hamido Fujita , And Nor Azura Md. Ghani, "Multiclass Prediction Model For Student Grade Prediction Using Machine Learning", Received May 26, 2021, Accepted June 12, 2021, Date Of Publication June 30, 2021, Date Of Current Version July 13, 2021, Ieee Access

[13] Arati Yashwant Amrale, Namrata Deepak Pawshe, Nikita Balu Sartape, Prof. Komal S. Munde, "Student Career Prediction Using Machine Learning", Volume:04/Issue:03/March-2022, E-ISSN: 2582-5208, International Research Journal Of Modernization In Engineering Technology And Science

[14] Vidyapriya.C,Vishhnuvardhan.R.C, "Student Career Prediction"

[15] Prathamesh Gavhane, Dhanraj Shinde, Ashwini Lomte, Naveen Nattuva, Shital Mandhane, "Career Path Prediction Using Machine Learning Classification Techniques", Volume 8, Issue 3, May-june-2021, International Conference - Innovation-2021-innovation-2021, International Journal Of Scientific Research In Computer Science, Engineering And Information Technology Issn : 2456-3307

[16] N. Vidyashreeram, Dr. A. Muthukumaravel, "Student Career Prediction Using Machine Learning Approaches"

[17] Zafar Iqbal, Junaid Qadir, Adnan Noor Mian, And Faisal Kamiran, "Machine Learning Based Student Grade Prediction: A Case Study"

[18] Hana Bydžovská, "A Comparative Analysis Of Techniques For Predicting Student Performance"

[19] Prakash, Mr. Sachin Garg, "Performance Analysis And Prediction Of Student Result Using Machine Learning", Volume:03/Issue:12/December-2021, E-issn: 2582-5208, International Research Journal Of Modernization In Engineering Technology And Science

[20] K. Sripath Roy, K.Roopkanth, V.Uday Teja, V.Bhavana, J.Priyanka, "Student Career Prediction Using Advanced Machine Learning Techniques", International Journal Of Engineering & Technology, 7 (2.20) (2018) 26-29

[21] Anitha K , Bhoomika C, J Andrea Kagoo, Kruthika K, Aruna Mg, "Student Grade Prediction Using Multiclass Model", August 2022, Ijirt , Volume 9 Issue 3 , Issn: 2349-6002

[22] Anooja S K, Dileep V K, "A Study On Student Career Prediction", Volume: 07 Issue: 02 , Feb 2020, International Research Journal Of Engineering And Technology (Irjet), E-issn: 2395-0056