BioGecko

Vol 12 Issue 02 2023

ISSN NO: 2230-5807

HARNESSING THE POWER OF MACHINE LEARNING FOR PREDICTING STUDENTS EMPLOYABILITY

Sarita Byagar
Department of Computer Science, Indira College of Commerce and Science, Pune, India.
Email: sarita.byagar@iccs.ac.in

Sarika Thakare
Department of Computer Science, Indira College of Commerce and Science, Pune, India.
Email: sarika.thakare@iccs.ac.in

Abstract
Being employable is an essential aspect of a student's academic career, as it determines their entry into the workforce. Predicting students' employability can help universities and colleges identify students who are likely to be successful in their chosen career paths and provide them with the necessary support to secure a job. In recent years, there has been a growing interest in using machine learning algorithms to predict students' employability. This research paper will explore the different approaches to predicting students' employability , the factors that influence employability, and the benefits and limitations of using machine learning algorithms for prediction. Machine learning have the ability to adapt and with the use of statistical models and algorithms they are able to draw inferences from patterns in data. Using ML algorithms forecasting can be done about the employability of students. Three ML algorithms viz, Naïve Bayes, Random Forest and Decision Trees are used to predict the employability of students and evaluation of the aforesaid algorithms are performed with respect to accuracy of the classifier.

Keywords:Machine Learning, Prediction, Naïve Bayes, Random Forest, Decision Tree, Placement, Forecasting,Employability

I. INTRODUCTION

In today's highly competitive job market, employability has become a critical factor in the success of students' careers. Employability refers to the ability of a student to obtain and maintain a job in their chosen field. Students' academic performance is closely related to their employability, as it is a crucial indicator of their skills and knowledge. With the increasing competition in the job market, it has become important to determine the factors that determine the employability of students. Machine learning (ML) algorithms can be used to analyze the data and predict the employability of students based on various factors such as academic performance, skills, and previous work experience. Employability is influenced by several factors, including academic performance, work experience, internships, extracurricular activities, and soft skills. Research suggests that academic performance is the most critical predictor of employability [1,2,3]. Employers consider a student's academic performance when evaluating their qualifications for a job. Therefore, students who perform well academically are more likely to be employed than those who do not. Work experience and internships are also essential predictors of employability. These experiences provide students with real-world skills and knowledge that are valued by employers. Students who have work experience or internships are more likely to be hired than those who do not [9]. Soft skills, such as communication, teamwork, problem-solving, and time management, are also critical predictors of employability. Employers look

for candidates who possess these skills, as they are essential for success in any job. Therefore, students who develop these skills are more likely to be employed than those who do not [4]

a) Machine Learning Methods

Machine Learning is a growing technology and as the name suggests makes the machine capable to learn and decide from past data and helps to make rational predictions and classifications. It is used for a variety of tasks viz. image processing, natural language processing, speech recognition, spam filtration etc. Supervised and Unsupervised learning are the two techniques of machine learning. But both the techniques are used in different scenarios and with different datasets. Supervised learning is a machine learning method in which models are trained using labeled data. Unsupervised learning is another machine learning method in which patterns inferred from the unlabeled input data. There are plethora of supervised and unsupervised algorithms available for classification and prediction of data.[1]

b) Factors Influencing Students Employability

Several factors can influence students' employability. Some of these factors includes academic Records, internship experience, communication skills, personality traits and many more [6]. Academic record indicates the skill and grading obtained in subjects of the studied curriculum. The student's academic record is a crucial factor in his/her employability. A good academic record can increase the chances of being placed in top companies.Internship Experience is the exposure obtained by internee in organization after completing graduate study.IT provides students necessary skills and experience required to secure a job.Communication Skills are the abilities of the student for effective communication.This is essential for employability. Students who can communicate well are more likely to be placed in top companies.Personality Traits are inter and external qualities of individual required for effective execution of the work. It includes leadership skills, team management, and adaptability can influence employability. [3]

II. RESEARCH OBJECTIVES

The following are some possible research objectives for "Students Employability Prediction using Supervised Machine Learning Algorithms":

- To identify the most effective supervised machine learning algorithms for predicting students employability based on historical data.
- To collect and preprocess data from various sources, including student performance data, skills etc, for use in training and testing the machine learning models.
- To develop and optimize machine learning models that accurately predict the likelihood of a student getting employable based on their academic performance, skills, and other relevant factors.
- To evaluate the performance of different machine learning models using various metrics, such as accuracy, precision, recall, F1 score, and ROC AUC, and compare them to identify the best-performing model.
- To explore the impact of different input features, such as grades, internships, extracurricular activities, and personal attributes, on the prediction accuracy of the machine learning models.
- To analyze the factors that influence students employability and identify actionable insights that can help universities, colleges and students improve their outcomes.

III. RELATED WORK

In recent years, there has been interest in utilising machine learning to predict students' employability. Many research have been carried out to create models that can forecast student's employability based on a variety of variables, including academic performance, talents, and prior work experience.The goal of Moumens et al (2020) work is to refresh the understanding of current trends and applications of various data mining techniques for young people's employability.Their overview claims that data mining techniques are useful for analysing employability issues in various contexts and methodologies using a variety of dimensions and features.[2]. Machine learning methods were utilised in a different study by Li et al. (2020) to forecast the employability of university graduates. The model was developed by the study using a variety of variables, including academic achievement, talents, and work experience. The study's findings demonstrated that the SVM algorithm had the highest accuracy, at 80%.Students are more likely to get job if they have strong academic achievements, few failed classes, volunteer work in organizations, or receive scholarships. In general, jobs are available to students with a mediocre academic record and few failed classes. But, finding a job is more challenging for those who have bad academic records.[9]. Study conducted by Chen et al (2021) creates an employment prediction model and examines the influencing elements of employment based on the CART decision tree algorithm.Then, in order to increase the learning accuracy, they further used the random forest algorithm. The findings of the experiment demonstrate that the decision tree and random forest algorithms can accurately forecast students' job status.[9]. Causat et al (2019), their research served as a basis for machine learning methods that predict students' employability. The results of learning algorithms were significantly impacted by the preparatory phase. Researchers came to the conclusion that Support Vector Machine (SVM) creates a predictive model that has the highest precision and recall measures of.991 or 91.15% and 91%, respectively, and an accuracy of 91.22%.[1]. According to Dubey et al(2020) for predicting students employability Random Forest, K-Nearest Neighbor, and Support Vector Machines outperformed Logistic Regression and Decision Tree for the smaller dataset. The classifiers with the highest accuracy, precision, recall, and F1-score were Logistic Regression, Random Forest, and Support Vector Machine for the bigger bootstrapped datasets. [3]. Bharambe et al (2017) offered a methodology for predicting students' employment status using data mining techniques like categorization. According to classification studies, Random Forest has a 99% accuracy rate when compared to other classification algorithms, including decision trees, KNN, Random forests, Naive Bayes, logistic regression, SVM (Linearsvc), Multi-class Ada Boosted, and Quadratic Discriminant Analysis (QDA). Consequently, the Random forest classification method was employed on pupils employability. Also, students' strengths and specific flaws were analysed so that they can overcome them and land the job they want. [4]

IV. PROPOSED METHODOLOGY

Methodology for "Students Employability Prediction using Supervised Machine Learning Algorithms" is executed as follows:

a) ***Data Collection:*** Collecting relevant data related to past students placements from multiple sources such as placement reports, student resumes, and recruitment data.

b) ***Data Preprocessing***: Cleaning and preprocessing the collected data by removing inconsistencies, missing values, and irrelevant features.

c) ***Feature Selection***: Selecting the most important features using techniques such as correlation analysis and feature importance ranking.

d) ***Data Partitioning***: Splitting the preprocessed data into training and testing sets in a ratio of 70:30 or 80:20, respectively.

e) ***Algorithm Selection:*** Choosing appropriate supervised machine learning algorithms based on the problem statement, available data, and performance metrics.

*In this research, the researchers have used three **supervised learning algorithms** viz. Naïve Bayes, Random Forest and Decision Trees, so-as-to predict the students employability.*

f) ***Model Training***: Training the selected algorithms on the training data and evaluating their performance using various metrics such as accuracy, precision, recall, and F1-score.
g) ***Hyperparameter Tuning:*** Fine-tuning the hyperparameters of the selected algorithms to improve their performance on the testing data.
h) ***Model Selection***: Comparing the performance of different algorithms and selecting the best-performing model based on the chosen evaluation metric.
The proposed methodology is depicted in the Fig.1.

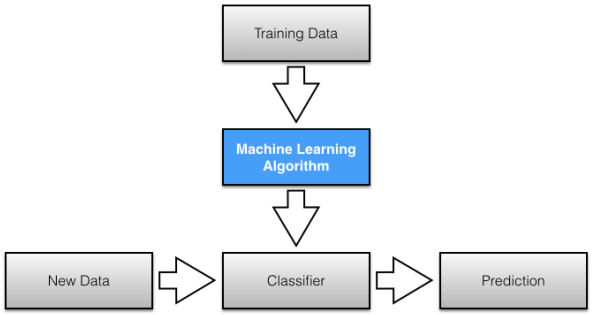

Fig.1 :Proposed Methodologyof Students employability prediction

V. EMPIRICAL WORK

a) Data Collection and Preprocessing

The sample data has been collected from IT students of INDIRA college in Pune. The dataset consists of approximately 3000 instances of students. Preprocessing is required to make the data ready for analysis. There are quite a few steps involved in data preprocessing which includes data cleaning, handling missing values, and attribute selection. In a dataset, there can be few attributes which may be irrelevant and hence hamper the accuracy of the result. Considering the same, the attributes which directly affect the classification and prediction are retained. In some cases, the dataset may contain blank or missing values which needs to be handled and equipped and with an appropriate value. A missing value can be replaced by a default value, or the mean of that column or most simple solution can be to remove the whole row.

b) Algorithm Selection

i. *Naïve Bayes*

Naïve Bayes is a supervised machine learning algorithm based on Bayes theorem which is used for classification. Its a probabilistic classifier which means it predicts based on the probability of a certain event. As it works on the principle of Bayes theorem, the formula for the same is as follows : P(A/B) = (P(B/A).P(A))/P(B). Naive Bayes classifiers assume that value of a particular feature is independent of the value of other feature, given class variable.

ii. *Decision Trees*

Decision trees algorithm is a significant machine learning technique which is frequently used for classification and regression problems. Decision trees have a sequence well defined set of rules which is used to classify patterns. As the name suggests, a tree has a graph like structure which comprises of branch and leaf nodes which makes it easily understandable and applicable. Decision trees are generally preferred for rational decision making and prediction of forthcoming circumstances based on historical data[9]

iii. *RandomForest*

Random forest (RF) algorithm is one of the most popular machine learning technique used for classification and regression problems. Forest, the term refers to multiple trees and more the number of trees the algorithm will be robust enough. Hence more number of trees ensure higher accuracy and problem solving and correct prediction ability. Random forest develops from decision trees. There are multiple decision trees and when a new example/case has to be classified, each decision tree delivers a classification for the provided input data. Random forest gathers the classification result and uses quorums/votes to choose the most preferred prediction as the result.

c) Training and Test Data

Training data is a set of data which is used to train the model and Test data is the set of data which is used to test the model after successful training. After preprocessing the data, the next step will be to split our dataset into two. Training set and a Test set. We will train our machine learning models on our training set, i.e. our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to examine how accurately it will predict. A general rule of the thumb is to assign 80% of the dataset to training set and therefore the remaining 20% to test set.

d) Tool used for Experiment.

The widespread tool WEKA was used for carrying out the analysis and prediction of the dataset. In the supervised learning category, there are various algorithms used. In this research, the algorithms used are Naïve Bayes, Decision Tree, and Random Forest algorithms. The dataset was analyzed for the aforesaid algorithms. For every algorithm, accuracy is the most important parameter which specifies how correctly the algorithm has classified the instances of the dataset. Apart from accuracy, Precision, Recall and F-score is also considered for the comparison of all three algorithms.

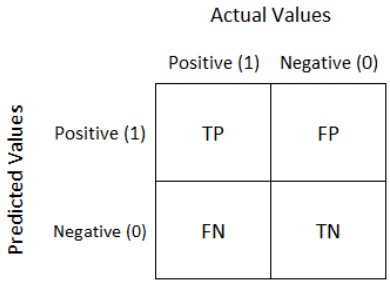

Fig. 2 : Confusion Matrix To Be Used To Capture Empirical Results [11]

A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm. [9]. This matrix consists of True positive (TP): Observation is predicted positive and is positive. False positive (FP): Observation is predicted as positive and negative. True negative (TN): Observation is predicted negative and is negative. False negative (FN): Observation is predicted negative and is actually positive.[9] .

Precision quantifies the number of positive class predictions that belong to the positive class. [9]Recall quantifies the number of positive class predictions made from all positive examples in the dataset. [9]. F-Measure provides a single score that balances both the concerns of precision and recall in one number.[9]

VI. RESULT ANALYSIS

a) Experiment Result

After executing the three mentioned algorithms, the results obtained are placed in Table-1, Table-2 and Table-3 respectively. Based on these tables, algorithms are evaluated using the metrics accuracy, Recall, Precison and F-Score, which is shown in Table-4.

Table-1: Confusion Matrix for Naïve Bayes Classifier

a	b	<-- classified as
1002	727	a = Employable
484	769	b = Less Employable

Table-2:Confusion Matrix for Decision Tree Classifier

a	b	<-- classified as
1729	0	a = Employable
1253	0	b = Less Employable

Table-3:Confusion Matrix for Random Forest

a	**b**	**<-- classified as**
1597	132	a = Employable
131	1122	b = Less Employable

Table-4: Performance analysis of algorithms.

Evaluation Parameters	**Naïve Bayes**	**Decision Tree**	**Random Forest**
Correctly Classified Instances	1771	1729	2719
Incorrectly Classified Instances	1211	1253	263
Accuracy	59.38 %	57.98 %	91.18 %
Recall	0.58	1	0.924
Precision	0.674	0.58	0.924
F Score	0.623	0.734	0.924

As per the results obtained, Random Forest gives best result.

b) Graphical Analysis

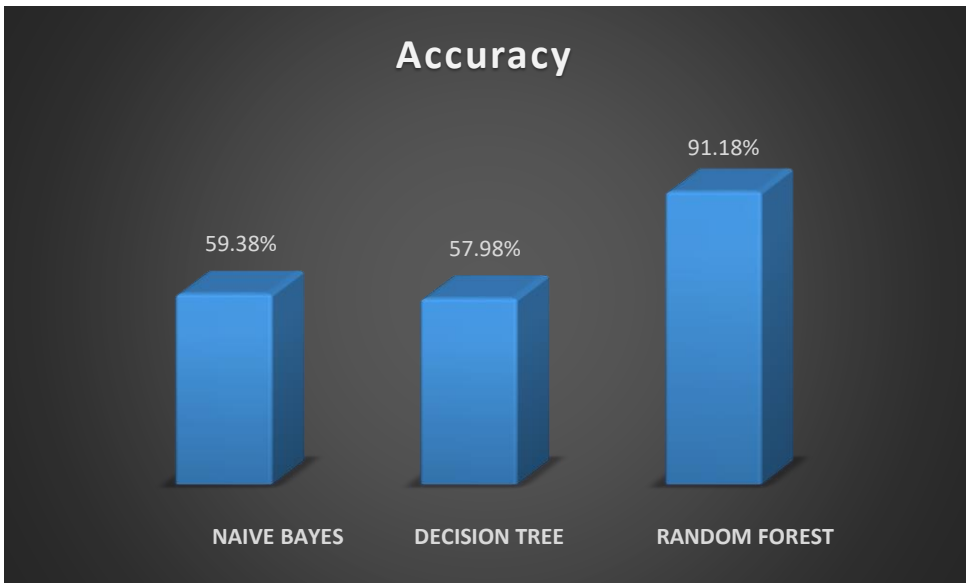

Fig.3 : Performance of Classifiers [Compiled by Researcher]

c) Discussion

Forecasting students' employability is essential for universities,students and colleges to identify students who are likely to be successful in their chosen career paths. Machine learning algorithms have the potential to improve the accuracy of predicting students' employability. In this respect an effort to study and predict the employability using the supervised machine learning classification algorithms Decision Tree, Naïve Bayes, and the Random forest algorithm have been used to authenticate the methodologies. The algorithms are applied to the data set and features are selected to build the model. As seen in Table 4 and Figure 3, the accuracy obtained after analysis for the Decision tree is57.98%, for Naïve Bayes is 59.38% and for the Random Forest is 91.18%. These results recommend that amongst the various supervised machine learning algorithms, the Random Forest classifier has the potential to more correctly classify and predict the employability of students. This study can be definitely of great help to colleges and universities to carry out the placement process smoothly. However, there are several limitations to using these algorithms, such as lack of transparency and bias which needs to be taken care of.

VII. CONCLUSION

Some common supervised machine learning algorithms used in students employabiity prediction include logistic regression, decision trees, random forests, support vector machines, and artificial neural networks. These algorithms can be trained on historical data to learn the relationships between various factors such as academic performance, work experience, and personal characteristics, and the likelihood of getting employable.

The evaluation of the performance of these algorithms is typically done using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. The results of these evaluations can help in selecting the most effective algorithm for a given dataset and application scenario.

In conclusion, supervised machine learning algorithms have the potential to improve the accuracy and efficiency of students employability prediction. However, it is important to carefully choose and evaluate the algorithms, as well as the dataset and feature selection process, to ensure the best possible results.

VIII. FUTURE RESEARCH WORK:

There are several potential areas for future research work on students employability prediction using supervised machine learning algorithms. Some possible directions are:

- ***Feature selection and engineering:*** Although many studies have focused on selecting the most relevant features for students employability prediction, there is still room for improvement in this area. Future research could explore more sophisticated techniques for feature selection and engineering, such as deep learning-based methods or natural language processing techniques to extract features from resumes.
- ***Ensemble learning:*** Ensemble learning is a technique that combines multiple models to improve prediction accuracy. Future research could investigate the use of ensemble learning techniques, such as stacking or boosting methods.
- ***Online and real-time prediction:*** Many prediction models are designed to be used offline, based on historical data. Future research could focus on developing online and real-time prediction models that can be used to predict t outcomes in real-time, based on streaming data from job applicants.
- ***Comparative analysis of algorithms:*** Although several supervised machine learning algorithms have been applied, there is a need for comparative analysis to determine the most effective algorithm for a given dataset and application scenario. Future research could focus on comparative analysis of different algorithms to identify the strengths and weaknesses of each algorithm in terms of prediction accuracy and efficiency.

IX. REFERENCES

1. Casuat, C. D., & Festijo, E. D. (2019, December). Predicting students' employability using machine learning approach. In *2019 IEEE 6th international conference on engineering technologies and applied sciences (ICETAS)* (pp. 1-5). IEEE.
2. Moumen, A., Bouchama, E. H., & EL IDIRISSI, Y. E. B. (2020, December). Data mining techniques for employability: Systematic literature review. In *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)* (pp. 1-5). IEEE.
3. Dubey, A., & Mani, M. (2019, October). Using machine learning to predict high school student employability–A case study. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 604-605). IEEE.
4. Bharambe, Y., Mored, N., Mulchandani, M., Shankarmani, R., & Shinde, S. G. (2017, September). Assessing employability of students using data mining techniques. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 2110-2114). IEEE.
5. Casuat, C. D., Castro, J. C., Evangelista, D. C. P., Merencilla, N. E., & Atal, C. P. (2020, November). StEPS: A Development of Students' Employability Prediction System using Logistic Regression Model Based on Principal Component Analysis. In *2020 IEEE 10th international conference on system engineering and technology (ICSET)* (pp. 17-21). IEEE.
6. C.J. Yue, J. Xia, W.Q. Qiu, "An empirical study on graduates' employment: Based on 2019 national survey," Journal of East ChinaNormal University(Educational Sciences), no. 4, pp. 1–17, 2020
7. T. Mishra, "Students' Performance and Employability Prediction through Data Mining: A Survey", 2017
8. Mishra, T. Students' Performance and Employability Prediction through Data Mining: A Survey, International Journal of Applied Engineering Research 11, no. 4 (2016): 2275-228

9. He, S., Li, X., & Chen, J. (2021, May). Application of data mining in predicting college graduates employment. In *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)* (pp. 65-69). IEEE.