MACHINE LEARNING-BASED SELECTION OF PHD ADMISSION

Abstract

Machine learning is now becoming a crucial decision-support tool in many academic fields. Both educational institutions and students are considered the intended beneficiaries in the field of education. Student admission is a vital problem in educational institutions. The traditional review method can no longer handle a high volume of doctoral applications. This paper discusses the machine learning algorithms for predicting students' chances of admission to a doctoral program. Students will be able to predict their chances of acceptance of ahead of time. We present a novel dataset called Phd_admission_dataset and examine it to determine the performance of several machine learning methods, such as Logistics Regression and KNN. Experimental results show that the KNN model outperforms the Logistics Regression model.

Keywords: K-nearest Neighbor; Logistic Regression; Machine Learning; Student Admission

Authors

Brajen Kumar Deka
Department of Computer Science
NERIM Group of Institutions
Guwahati, India
brajendeka@gmail.com

Chinmoy Talukdar
Department of Computer Science
NERIM Group of Institutions
Guwahati, India
Chinmoy03@gmail.com

I. INTRODUCTION

Admittance to doctoral programs for students is essential in any educational institution. Each student must be carefully selected. However, the department head is not permitted to predict each applicant's chances of admission. This model will assist in calculating each applicant's probability based on their performance on three exams (written, presentation, and viva). We have historical data from the past applications that has been used to train the Logistic Regression and KNN algorithms. The effectiveness of the classifiers and the outputs should be assessed using a variety of metrics. This training set contains the applicant's three exam scores and the admission decision for each training example. All experiments are carried out in a simulated environment using the Jupyter Notebook platform. The outcome of this procedure will be used to decide whether or not students are qualified to enrol in programs. This model predicts the likelihood of admission to a PhD program. This model will generate more accurate results than the existing systems.

Employers are continuously searching for employees with the best knowledge and experience, as global marketplaces are rapidly changing. Young professionals who want to advance in their careers or specialize in a specific field are constantly looking for advanced degrees that will allow them to expand their knowledge and skills. Therefore, more students have applied for doctoral programs in the last ten years [1] [2] [3]. This reality motivated us to research student grades and the probability of admission to PhD programs to support institutions in evaluating their capability to accept a certain number of PhD candidates each year and allocating the appropriate funds.

II. RELATED WORK

Machine learning can be used to predict the likelihood of admission to a university by analyzing data on student applications and grades. Acharya et al. [4] conducted an excellent study in which they tested four regression algorithms, including linear regression, support vector regression, decision trees, and random forest, to find the best model for graduate admissions.

Chakrabarty et al. [5] compared linear regression and gradient boosting regression in predicting admit chance and discovered that gradient boosting regression generated better results.

The model that analyses the graduate admissions procedure in American colleges or universities was created by Gupta et al. [6] using machine learning methods. The purpose of this study was to guide students in selecting the best educational institution for their application. Five machine learning models, including logistic classifiers, AdaBoost, and SVM (Linear Kernel), were created for this investigation.

An innovative study by Waters and Miikkulainen [7] was presented that improves the effectiveness of examining applications using statistical machine learning and helps in classifying graduate admission applications depending on acceptance level.

Sujay [8] calculated the likelihood of approving graduate applicants as postgraduates using linear regression. However, other models were not studied.

III. GRADUATE ADMISSION PROCESS

We present a high-level description of the PhD admissions system to help readers understand the consequences of the prediction approach. Departments may only accept applications for doctoral programs through an online or offline mechanism. There are numerous forms that students must fill out detailing their academic history, exam scores, areas of research interest, and other pertinent data. Each student who submits an online application has their information saved in a departmental database. After the application period has closed, faculty members evaluate the applications using a confidential web-based system. After reviewing each file, a reviewer sends feedback to the other reviewers and assigns a real-valued score to represent the applicant's level of qualification. The time required for each full review varies depending on the reviewer's style and skill, the application's quality and substance, and the stage of the review process, but a typical full review takes roughly 10-15 minutes. The committee frequently goes through the pool several times before approving or rejecting each candidate based on the ratings and opinions of the examiners who reviewed each applicant's file. Although quality is a crucial factor in this selection, the numbers of new students required by the faculty in each research subject and current research opportunities in the department have a significant impact.

IV. OBJECTIVES OF THE PROPOSED APPROACH

Machine learning has improved the Ph. D admissions process. The main goal of the PhD Admission Selection Committee is to select the best candidates for admission to the program.

1. The study's objective is to create a system that can manage multiple variables while getting around the challenges of physical labour.
2. This system enables the department head to evaluate the outcomes instantly.
3. It will decrease time wastage while simultaneously promoting increased technology use.

V. METHODOLOGY

1. **Logistics regression:** The likelihood of a discrete result given an input variable can be estimated using the logistic regression approach. The most common logistic regression models contain a binary outcome, which is anything with two possible values, such as true or false, yes or no, and so on. Multinomial logistic regression can depict events with more than two distinct likely outcomes. Consider that there are classification challenges in cyber security, such as attack detection, logistic regression is a powerful analytical tool for classifying new data. Logistic regression is an effective supervised machine learning method for binary classification problems. It is a kind of linear regression used to solve classification issues [9]. The main difference between logistic and linear regression is that its range is limited to 0 and 1. Also, unlike linear regression, logistic regression does not require a linear relationship between input and output variables. It is due to the odds ratio having undergone a nonlinear log transformation. Logistic regression predicts the probability of the default class and converts the likelihood into a binary value (0 or 1) for classification using the “sigmoid” function as follows:

$$f(x) = \frac{1}{1+e^{-x}} \quad (1.1)$$

2. **K-Nearest Neighbors:** K-Nearest Neighbor assigns a case to the class with the highest frequency of occurrence among its k neighbors. Distance functions such as Euclidean are used to compute the distance between an instance and its neighbors.

$$D_{Euclidean} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (1.2)$$

K-nearest Neighbor (KNN) is a supervised machine learning technique used to solve classification and regression problems [10]. It operates according to the similarity measurement principle. Therefore, predicting a new value requires considering neighbors. In a regression problem, KNN is used to get the mean of the k labels. For classification tasks, it will return the mode of k labels.

3. **Dataset used:** The dataset used in this study relates to the field of education. The 768-row dataset contains the three independent variables listed below:
   - **Written:** the score obtained in the written exam.
   - **Presentation:** the score obtained in the presentation skill.
   - **Viva:** the score obtained in the viva.

4. **Flow diagram for the proposed approach**

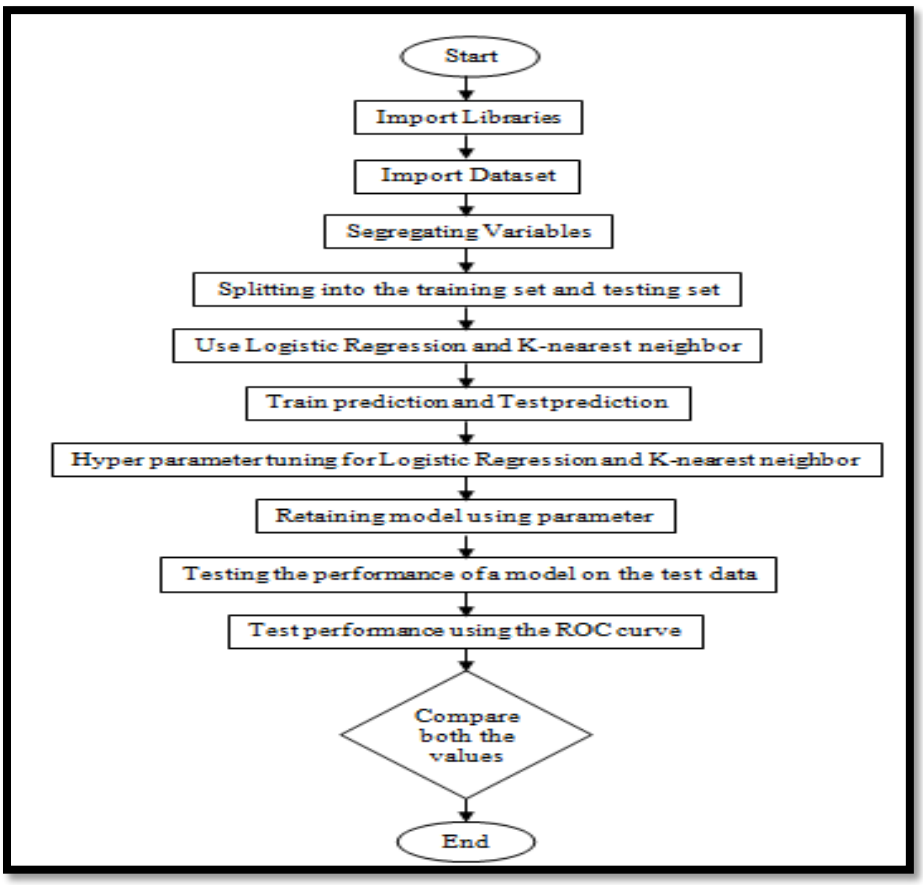

Figure 1: Flow Diagram of the Proposed Approach

One dependent variable, ***admission_chance***, can be predicted and will range from 0 to 1. The dataset was imported and split randomly into two halves using the holdout method. The training with 614 observations using 80% of the dataset is shown in the first section. The second part depicts the testing with 154 cases and 20% of the dataset.

VI. RESULTS AND DISCUSSION

The approach comprises many parts. First, the system receives application files from the departmental database and performs preprocessing to normalize the data. The files then become high-dimensional feature vectors. This feature-encoded historical data is then applied to train a logistic regression classifier. After that, the classifier determines the likelihood of each new applicant being admitted and generates data for the admissions committee to review. After that, the feature-encoded data is used in the same manner to train the KNN classifier. The classifier then determines the chance of each new applicant being accepted and generates data for the admissions committee to review. Finally, depending on this information, decisions are made on which documents should be thoroughly investigated and quickly verified to support the model's predictions.

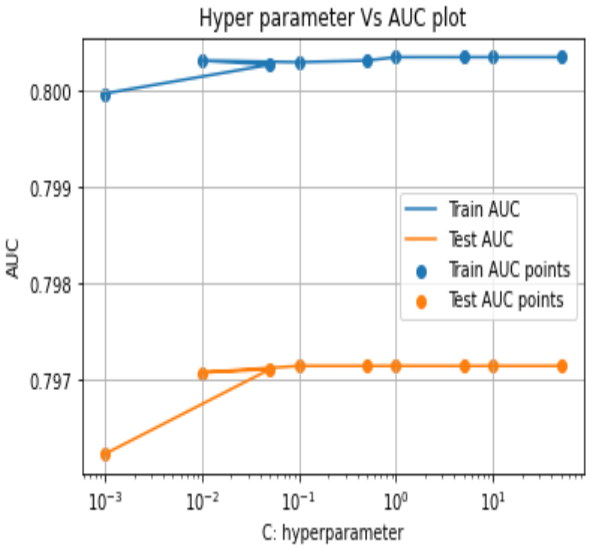

Figure 2: ROC curve for Logistic Regression

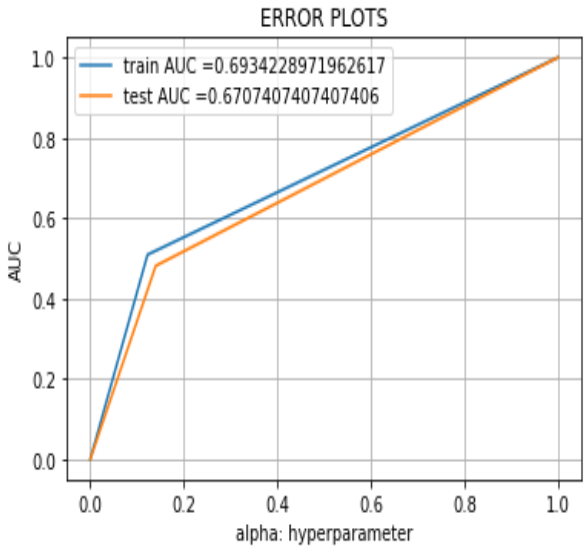

Figure 3: Error Plots for Logistic Regression

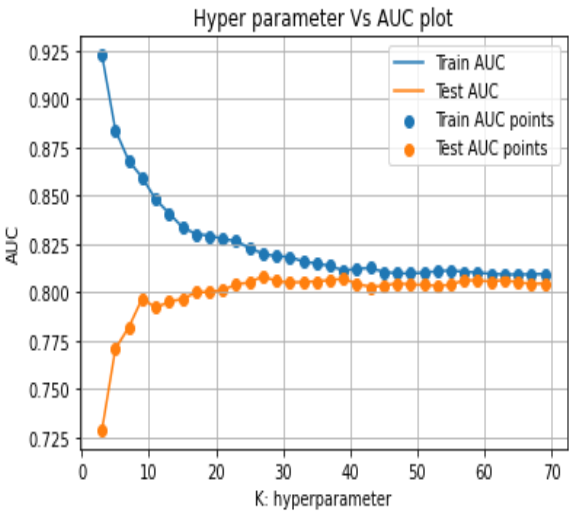

Figure 4: ROC curve for KNN

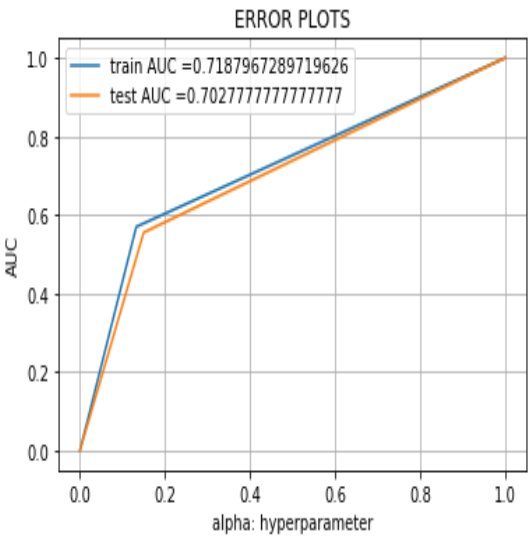

Figure 5: Error Plots for KNN

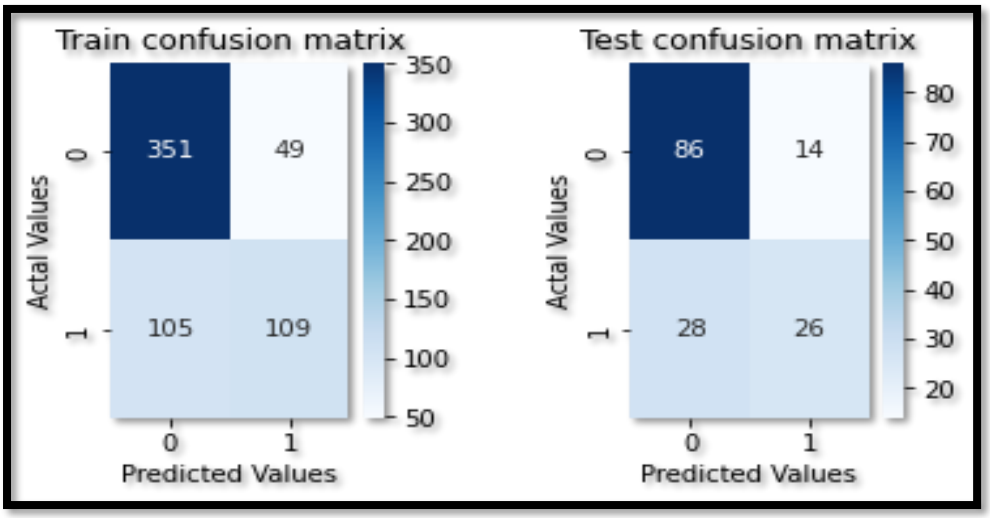

Figure 6: Confusion Matrix for Logistic Regression

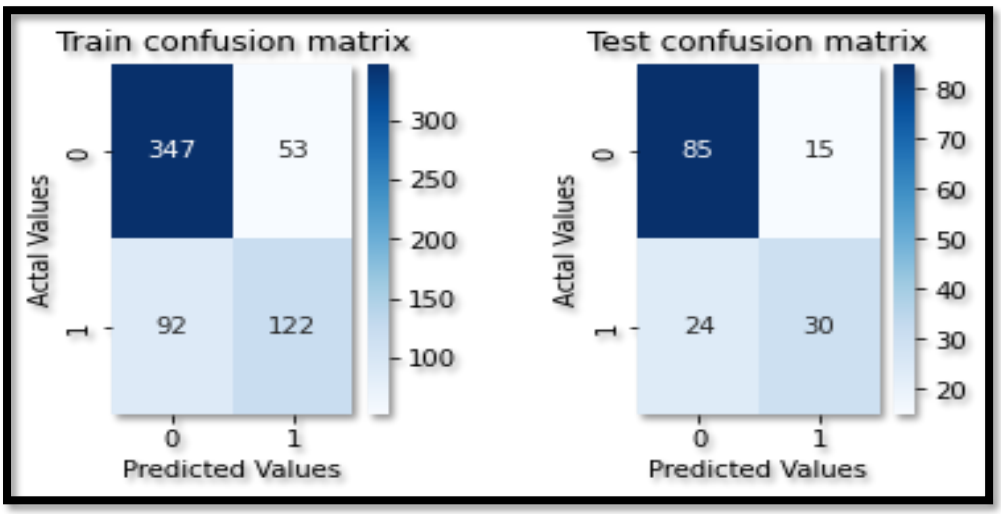

Figure 7: Confusion Matrix for K-Nearest Neighbor

Table 1: Logistic Regression and KNN Model Performance Analysis

Model wise Measure	Logistic Regression		K-Nearest Neighbor	
	Training	Testing	Training	Testing
Precision	0.689873	0.65	0.697143	0.666667
Recall	0.509346	0.481481	0.570093	0.555556
F1-Score	0.586022	0.553191	0.627249	0.606061
Accuracy	0.749186	0.727273	0.763844	0.746753

Table 2: Best Hyperparameter and AUC value

Model	Hyperparameter	AUC
Logistic Regression	0.1	0.6707407407407406
KNN Classifier	27	0.7027777777777777

VII. CONCLUSION

Machine learning algorithms were employed in this paper to predict a student's chances of admission to a doctoral program. We describe two machine learning algorithms: logistic regression and k-nearest neighbors. Experimental results demonstrate that the K-nearest neighbors model outperforms other models. Future studies will allow testing of the additional models on larger datasets to see which model performs best.

REFERENCES

[1] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, “Multi-split Optimized Bagging Ensemble Model Selection for Multi-class
[2] Educational Data Mining,” Applied Intelligence., Vol. 5 https://doi.org/10.1007/s10489-020-01776-3
[3] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, “Systematic ensemble model selection approach for educational data
[4] mining,” Knowledge-Based Systems, Vol. 200, pp. 105992, 2020. https://doi.org/10.1016/j.knosys.2020.105992.
[5] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, “Machine learning towards intelligent systems: applications, challenges, and
[6] opportunities,” Artificial Intelligence Review, Vol. 54, Issue 5, pp.3299-3348, 2021. https://doi.org/10.1007/s10462-020-09948-w
[7] M. S. Acharya, A. Armaan, and A. S. Antony, “A comparison of regression models for prediction of graduate admissions,” Second
[8] International Conference on Computational Intelligence in Data Science (ICCIDS-2019), pp.1-5, 2019. DOI: 10.1109/ICCIDS.2019.8862140
[9] N. Chakrabarty, S. Chowdhury, and S. Rana, “A Statistical Approach to Graduate Admissions,” Chance Prediction, pp. 145-154, 2020.
[10] N. Gupta, A. Sawhney, and D. Roth, “Will I Get in? Modelling the Graduate Admission Process for American Universities,” IEEE Int.
[11] Conf. Data Min. Work. ICDMW, Vol. 0, pp. 631–638, 2016.
[12] A. Waters and R. Miikkulainen, “GRADE: Graduate Admissions,” AI MAGAZINE, pp. 64–75, 2014. DOI: https://doi.org/10.1609/aimag.v35i1.2504

[13] S. Sujay, "Supervised Machine Learning Modelling & Analysis for Graduate Admission Prediction," International Journal of Trend
[14] in Research and Development (IJTRD), Vol. 7, no. 4, pp. 5–7, 2020.
[15] L. Lei, "Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning," International Conference on Robots & Intelligent Systems (ICRIS), pp. 157-160, July 2018. DOI: 10.1109/ICRIS.2018.00049
[16] S. Aljasmi, A. B. Nassif, I. Shahin, and A. Elnagar, "Graduate Admission Prediction Using Machine Learning," International Journal of Computers and Communications, Vol. 14, pp. 79-83, 2020. DOI: 10.46300/91013.2020.14.13

