Machine Learning-Based Selection of PhD Admission

Brajen Kumar Deka1,\*

Department of Computer Science,

NERIM Group of Institutions,

Guwahati, India

brajendeka@gmail.com

Chinmoy Talukdar2

Department of Computer Science,

NERIM Group of Institutions,

Guwahati, India

Chinmoy03@gmail.com

ABSTRACT

Machine learning is now becoming a crucial decision-support tool in many academic fields. The intended beneficiaries in the field of education include both educational institutions and students. Admission of students is a vital problem in educational institutions. The enormous volume of applicants for the PhD programme cannot be handled by the traditional review process any more. This paper discusses machine learning algorithms for predicting a student's chances of admission to a doctoral program. Students will be able to predict their chances of acceptance ahead of time. We provide a novel dataset called the Phd\_admission\_dataset and evaluate it to determine the effectiveness of different machine learning methods, including Logistics Regression and KNN. According to experimental findings, the KNN model outperforms the Logistic Regression model.

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

# INTRODUCTION

In an educational institution, student entrance to doctoral programs is crucial. Each student must be carefully selected. However, the Head of the Department cannot evaluate each student's chances of acceptance. This model will assist in calculating each applicant's chances based on their performance on three exams (written, presentation, and viva). We have historical data from past applications that has been used to train the Logistic Regression and KNN algorithms. The effectiveness of the classifiers and their 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 programmes.This model predicts the likelihood of admission to a PhD program. This model will generate more accurate results than the existing systems.

Employers are constantly looking for employees with the most knowledge and experience, as global markets are fast evolving. Young professionals who want to advance in their careers or specialize in a specific industry are constantly looking for higher degrees that may help them develop their knowledge and skills. As a result, more students have applied to pursue PhD degrees in the last ten years [1] [2] [3]. This fact motivated us to research student grades and the probability of admission to PhD programmes to support institutions in determining the capacity to accept the number of PhD candidates each year and allocating the necessary funds.

# RELATED WORK

Several studies and research activities have already been conducted on datasets relevant to graduate admissions using various machine learning methods. 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 predict the likelihood of admission based on the best model with the lowest MSE, which was multi-linear regression.

Furthermore, 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.

Waters and Miikkulainen [7] proposed a remarkable article that aids in evaluating graduate admission applications according to the level of acceptance and enhances the performance of reviewing applications using statistical machine learning.

Sujay [8] calculated the percentage chance of admitting graduate students to master's programs using linear regression. Other models, however, were not examined.

# GRADUATE ADMISSION PROCESS

To better understand the impact of the prediction system, we provide a high-level explanation of the PhD admissions procedure. Departments may only accept applications for doctoral programs through an online or offline mechanism. Students must fill out several documents that include their educational background, test results, research interests, and other information. When a student submits an online application, information about them is kept in a departmental database. After the application period has ended, faculty members assess the applicants using a private 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 based on the reviewer's style and expertise, the quality and substance of the application, and the stage of the review process, but a typical full review takes about 10-15 minutes. The committee often goes through the pool several times before admitting or rejecting each candidate based on the scores and opinions of the reviewers who looked at his or her file. Although quality is the key criterion for this selection, it is heavily influenced by current research opportunities in the department, i.e. how many fresh students the faculty request to be admitted in each research field.

# OBJECTIVES OF THE PROPOSED APPROACH

Machine learning is used to improve the PhD admissions process. The main goal of the PhD Admission Selection Committee is to select the best candidates for admission to the program.

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

# METHODOLOGY

## **Logistic Regression**

Logistic regression is a method for estimating the likelihood of a discrete outcome given an input variable. The most popular logistic regression models include a binary result, which is anything that can take two values like true or false, yes or no, and so on. Multinomial logistic regression can represent events with more than two unique probable outcomes. Logistic regression is a valuable analytical technique for classifying new samples, considering that cyber security involves classification problems such as the identification of attacks. Logistic regression is an excellent supervised machine learning technique for binary classification problems. Logistic regression is a kind of linear regression used to solve classification issues [9]. The main difference between logistic regression and linear regression is that its range is restricted to 0 and 1. Also, unlike linear regression, logistic regression does not require a linear relationship between input and output variables. This 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:

(1.1)

## **K-nearest Neighbors**

K-Nearest Neighbor assigns a case to the class with the highest frequency of occurrence among its k nearest neighbours. Calculating the separation between an instance and its neighbour involves using distance functions like Euclidean.

(1.2)

K-nearest Neighbor (KNN) is a supervised machine learning method for classification and regression problems [10]. It works on the similarity measurement principle. As a result, predicting a new value requires considering neighbours. KNN uses mathematical equations to compute the distance between points in order to find neighbours. KNN is used in a regression problem to determine the mean of the k labels. It will return the mode of k labels for classification tasks.

## **Dataset Used**

The dataset presented in this study is related to the educational domain. Phd\_admission\_dataset is a dataset with 768 rows that includes the following three independent variables:

* ***Written***: the score obtained in the written exam.
* ***Presentation***: the score obtained in the presentation skill.
* ***Viva***: the score obtained in the viva.

One dependent variable, ***admission\_chance***, can be predicted and will range from 0 to 1. After importing the dataset, it was randomly partitioned into two sections using the holdout method. The first section covers 80% of the dataset to demonstrate training with 614 observations. The second section uses 20% of the dataset to show testing with 154 instances.

## **Flow diagram for the proposed approach**

Hyperparameter testing

Splitting into train set, cross-validation set, and test set

Import Libraries

Import Dataset

Segregating variables

Compare both the values

Test performance using the ROC curve

Use KNN

Hyperparameter tuning for KNN

Train prediction and Test prediction

Retraining model using best parameter

Testing the performance of a model on the test data

Use Logistic Regression

Train prediction and test prediction

Hyperparameter tuning for logistic regression

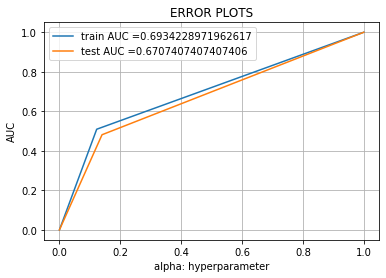
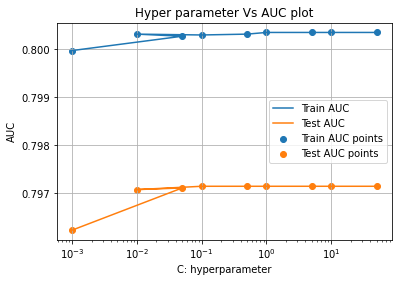
Retraining model using best parameter

Testing the performance of a model on the test data

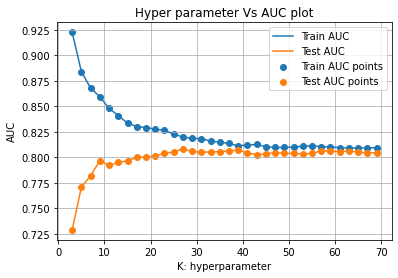
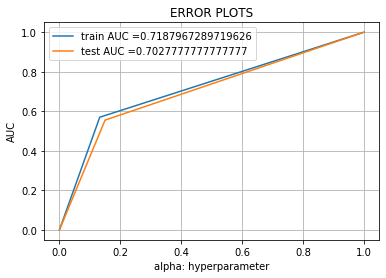
**Figure 1: Flow diagram of the proposed approach**

# RESULTS AND DISCUSSION

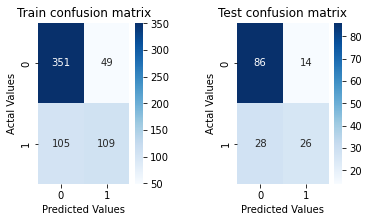
The approach comprises many parts. First, the system reads applicants' files from the departmental database and conducts preprocessing to normalize the data. The files are then encoded as high-dimensional feature vectors. After that, a logistic regression classifier is trained on the feature-encoded historical data. The classifier then calculates the likelihood of each new applicant being accepted and produces data to be sent to the admissions committee. Then the KNN classifier is trained on the feature encoded data in the same way. The classifier then calculates the likelihood of each new applicant being accepted and produces data to be sent to the admissions committee. Finally, based on this data, decisions are made on which files should be thoroughly examined and which may be verified quickly to confirm the model's predictions.

****

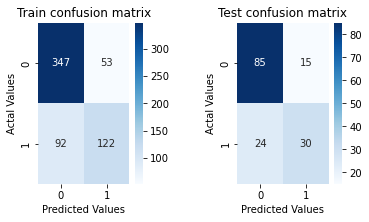
**Figure 2: ROC curve for Logistic Regression Figure 3: Error Plots for Logistic Regression**

****

**Figure 4: ROC curve for KNN Figure 5: Error Plots for KNN**



**Figure 6: Confusion matrix for Logistic Regression in both training and testing**



**Figure 7: Confusion matrix for KNN in both training and testing**

**Table 1: Performance Analysis of Logistic Regression and KNN model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Logistic Regression | | K-Nearest Neighbor | |
| Train | Test | Train | Test |
| 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 |

# CONCLUSION

In this paper, machine learning models were used to predict the opportunity of a student to get admitted to a PhD program. The machine learning models included are logistic regression and k-nearest neighbour. Experimental results show that the KNN model surpasses logistic regression. As for future work, more models can be conducted on more datasets to learn the model that gives the best performance.

##### REFERENCES

1. M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, “Multi-split Optimized Bagging Ensemble Model Selection for Multi-class

Educational Data Mining,” *Applied Intelligence.*, Vol. 50, No. 12, pp. 4506–4528, 2020.

<https://doi.org/10.1007/s10489-020-01776-3>

1. M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, “Systematic ensemble model selection approach for educational data

mining,” *Knowledge-Based Systems*, Vol. 200, pp. 105992, 2020.

<https://doi.org/10.1016/j.knosys.2020.105992>.

1. M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, “Machine learning towards intelligent systems: applications, challenges, and

opportunities,” [*Artificial Intelligence Review*](https://dl.acm.org/toc/artr/2021/54/5), Vol. 54, Issue 5, pp.3299-3348, 2021.

<https://doi.org/10.1007/s10462-020-09948-w>

[4] M. S. Acharya, A. Armaan, and A. S. Antony, “A comparison of regression models for prediction of graduate admissions,” Second

*International Conference on Computational Intelligence in Data Science* (*ICCIDS-2019*), pp.1-5, 2019.

**DOI:**[10.1109/ICCIDS.2019.8862140](https://doi.org/10.1109/ICCIDS.2019.8862140)

[5] N. Chakrabarty, S. Chowdhury, and S. Rana, “A Statistical Approach to Graduate Admissions,” *Chance Prediction*, pp.

145-154, 2020.

[6] N. Gupta, A. Sawhney, and D. Roth, “Will I Get in? Modelling the Graduate Admission Process for American Universities,” *IEEE Int.*

*Conf. Data Min. Work. ICDMW*, Vol. 0, pp. 631–638, 2016.

[7] A. Waters and R. Miikkulainen, “GRADE: Graduate Admissions,” *AI MAGAZINE*, pp. 64–75, 2014.

DOI: <https://doi.org/10.1609/aimag.v35i1.2504>

[8] S. Sujay, “Supervised Machine Learning Modelling & Analysis for Graduate Admission Prediction,” *International Journal of Trend*

*in Research and Development (IJTRD)*, Vol. 7, no. 4, pp. 5–7, 2020.

[9] 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](https://doi.org/10.1109/ICRIS.2018.00049)

[10] 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