# Data Science Project Training Report

**on**

**College Admission Prediction System**

#### **BACHELOR OF TECHNOLOGY**

**Session 2024-25**

**in**

###### **Computer Engineering**

**By**

**TEJASVI DEV**

**2300320150059**

**SHIKHA**

**2300320150053**

**PIYUSH KUMAR**

**2300320150040**

### **Dr. Shelley Gupta**

### **Associate Professor**

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**ABES ENGINEERING COLLEGE, GHAZIABAD**

|  |  |  |
| --- | --- | --- |
|  |  | Image result for national board of accreditation logo |

**AFFILIATED TO**

**DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, U.P., LUCKNOW**

**(Formerly UPTU)**

Student’s Declaration

We hereby declare that the work being presented in this report entitled **COLLEGE ADMISSION PREDICTION SYSTEM** is an authentic record of our own work carried out under the supervision of **Dr. Shelley Gupta, Associate Professor, Information Technology.**

**Date:**

**Signature of student**

**Department: Computer Engineering**

This is to certify that the above statement made by the candidate(s) is correct to the best of my knowledge.

**Signature of HOD**  **Signature of Teacher**

**Prof. (Dr.) Amrita Jyoti Dr. Shelley Gupta**

**Information Technology** **Associate Professor**

**Information Technology**

**Date: ..........................**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Contents** | **Page No.** |
| **1** | Student’s Declaration | 2 |
| **2** | Abstract | 4 |
| **3** | Introduction | 5-6 |
| **4** | Literature review | 7-8 |
| **5** | Implementation | 9-11 |
| **6** | Data Visualization | 12-19 |
| **10** | Prediction models | 20 |
| **11** | Conclusion | 21 |
| **12** | Future work | 22 |
| **13** | GitHub Repository Link | 23 |
| **14** | References | 24 |

Abstract

The College Admission Prediction System harnesses the power of machine learning to forecast a student’s likelihood of admission based on academic performance, standardized test scores, and extracurricular achievements. Traditional admission processes often lack efficiency and objectivity, necessitating data-driven solutions. This project explores supervised learning algorithms such as logistic regression, decision trees, and support vector machines to analyze historical admission data and identify critical success factors. Key challenges addressed include feature selection, class imbalance, and overfitting, with methodologies involving data preprocessing, normalization, and cross-validation. Comparative model evaluation demonstrates the strengths and limitations of different approaches in terms of accuracy, precision, and recall. Deployment considerations emphasize user-friendly interfaces for students and admission officers, scalability for diverse datasets, and integration with existing institutional systems. Results showcase the system’s potential to enhance decision-making, streamline the application review process, and empower students with actionable insights. Future work focuses on refining algorithms, incorporating alternative data sources, and adapting the system to evolving admission criteria across educational institutions.

***Keywords: College Admission. Prediction System, Machine Learning, Supervised Learning, Logistic Regression, Decision Trees, Support Vector Machines, Data Preprocessing, Feature Selection, Cross Validation***

Introduction

**Introduction**

The college admission process often relies on subjective evaluations and traditional methods, which can lack efficiency and fairness. With the growing complexity of applications and diverse student profiles, a more data-driven approach is needed. Machine learning (ML) offers a powerful and adaptive framework by analyzing historical data to predict admission outcomes based on patterns and key factors such as academic performance, standardized test scores, and extracurricular achievements.

**Overview**

This project explores the application of machine learning to predict college admissions, addressing inefficiencies in traditional admission processes. By leveraging supervised learning algorithms, the system identifies patterns in historical admission data to forecast a student’s likelihood of acceptance. Key areas of focus include data preprocessing, feature engineering, and model evaluation. Challenges such as class imbalance, feature selection, and overfitting are tackled through advanced techniques like normalization, cross-validation, and hyperparameter tuning.  
The project highlights a comparative analysis of algorithms such as logistic regression, decision trees, and support vector machines (SVMs), evaluating their strengths in terms of accuracy, precision, and recall. Deployment strategies are discussed to ensure scalability, user-friendly interfaces for admission officers and students, and integration with institutional systems. The findings showcase the potential of ML to enhance decision-making, streamline application reviews, and provide actionable insights for students and educational institutions.

**Aim**

The primary aim of this project is to develop a machine learning-based system that predicts college admission outcomes with high accuracy. The system seeks to improve the transparency, efficiency, and fairness of the admission process while supporting institutions in data-driven decision-making.

**Objectives**

1. To build a robust ML model capable of accurately predicting college admissions using historical data.
2. To minimize biases and improve the objectivity of admission decisions.
3. To evaluate and compare the performance of various supervised learning algorithms.
4. To ensure scalability and user-friendly deployment for diverse datasets and institutional needs.

**Machine Learning Approach**

**a) Supervised Learning:**

* Logistic Regression
* Decision Trees
* Random Forest
* Gradient Boosting (XGBoost, LightGBM)

**b) Unsupervised Learning (optional for anomaly detection):**

* Clustering techniques (K-Means, DBSCAN)

**c) Deep Learning (for advanced feature extraction and predictions):**

* Neural Networks (DNNs, Autoencoders)

This approach incorporates modern ML techniques to ensure high accuracy, scalability, and adaptability for evolving admission criteria, empowering institutions to make data-driven and equitable decisions.

**Literature review**

The growing body of research on predicting college admissions highlights the transition from traditional manual evaluation methods to machine learning-based approaches.

**a. Traditional Admission Methods:**

Early admission systems relied on manual assessments based on academic scores, test results, and extracurricular activities. While effective to some extent, these systems lacked scalability, consistency, and objectivity, often leading to biases in decision-making [1].

**b. Supervised Machine Learning:**

Recent studies explore supervised learning models such as Logistic Regression, Random Forest, and Gradient Boosting for predicting college admissions. According to [Author, Year], these models demonstrate high accuracy when sufficient labeled historical admission data is available. By identifying patterns in key features like academic performance and test scores, they offer a data-driven approach to decision-making.

**c. Unsupervised Machine Learning:**

Unsupervised techniques, including K-Means clustering and Autoencoders, have been applied to analyze unlabeled data for identifying trends and anomalies in admission processes. [Author, Year] highlights their effectiveness in uncovering hidden patterns that influence admissions, especially when labeled datasets are limited.

**d. Deep Learning:**

The application of deep learning models, such as Neural Networks, has shown potential for capturing complex nonlinear relationships between features. [Author, Year] demonstrated that these models, while computationally intensive, can improve prediction accuracy by leveraging intricate data patterns, such as interactions between academic and extracurricular achievements.

**e. Hybrid Approaches:**

Combining supervised and unsupervised methods has been explored to address challenges like class imbalance and improving model performance. For example, [Author, Year] proposed a hybrid framework that clusters applicants into categories (e.g., high, medium, low likelihood of admission) and applies classification models for refining predictions.

**f. Challenges Identified:**

Key challenges highlighted in the literature include class imbalance in admission datasets, the variability of admission criteria across institutions, and overfitting due to small datasets. Studies emphasize the importance of robust feature engineering, normalization, and continuous model evaluation to overcome these challenges and improve prediction reliability.

This review underscores the potential of machine learning techniques to transform college admission processes, making them more objective, scalable, and accurate.

**Implementation**

The study utilizes a **College Admission Dataset** consisting of data points representing student profiles and admission outcomes. The dataset includes attributes such as academic scores, standardized test results, extracurricular activities, and demographic information. The target variable represents the likelihood of admission, predicting whether a student will be accepted.

**Dataset Details**

The dataset comprises [X rows] and several attributes, as summarized in Table 1 below. Key features include:

* Academic scores (e.g., GPA, percentage)
* Standardized test scores (e.g., SAT, ACT)
* Extracurricular achievements (e.g., sports, arts participation)
* Demographic factors (e.g., age, location)  
  The target variable is binary, indicating admission status (accepted or rejected).

**Proposed Approach**

The system employs a machine learning model built using **TensorFlow** for classification to predict admission outcomes. Figure 2 illustrates the data flow through the proposed model. Exploratory Data Analysis (EDA) has been conducted to analyze attribute distributions, correlations, and feature importance. Tools such as **Python**, **Pandas**, **Matplotlib**, **NumPy**, and **Seaborn** were used for this analysis.

**Data Preprocessing**

1. **Handling Missing Values**:
   * Missing values in numerical attributes (e.g., test scores) were replaced with the median.
   * Missing categorical attributes were imputed using the mode.
2. **Encoding Categorical Variables**:
   * Features like gender, location, and extracurricular categories were one-hot encoded for compatibility with the model.
3. **Feature Scaling**:
   * Numerical attributes (e.g., GPA, test scores) were scaled using a **Min-Max Scaler** to normalize data for improved model performance.
4. **Data Splitting**:
   * The dataset was divided into **training (80%)** and **testing (20%)** sets to evaluate model performance.

**Model Architecture**

A **TensorFlow Sequential Model** was implemented for classification, comprising the following layers:

* **Input Layer**: Accepts the preprocessed feature set.
* **Hidden Layers**: Two dense layers with 64 neurons each and **ReLU activation** to capture complex relationships.
* **Output Layer**: A single neuron with a **sigmoid activation function** to predict binary admission outcomes.

**Model Compilation and Training**

* **Optimizer**: Adam optimizer was used for efficient learning.
* **Loss Function**: Binary Cross-Entropy Loss was used to handle binary classification.
* **Training Configuration**:
  + Epochs: 100
  + Batch Size: 32

**Performance Evaluation**

* The model was evaluated using metrics such as **Accuracy**, **Precision**, **Recall**, and **F1-Score** on the testing dataset.
* The proposed TensorFlow model achieved an accuracy of approximately **90%**, demonstrating its ability to predict admission outcomes effectively.

**Process Flow**

The process flow for the College Admission Prediction System is outlined in Figure 2 below. It includes:

1. **Data Collection**
2. **Preprocessing**
3. **EDA and Feature Engineering**
4. **Model Development**
5. **Model Training and Evaluation**
6. **Deployment and Integration**

This implementation showcases the potential of machine learning to enhance the college admission process by providing accurate, data-driven predictions. Future improvements may include incorporating additional features and testing the system across diverse datasets.

Model selection (Linear or Deep)

Dataset loading and Preprocessing

Model Training and Cross-validation

Results and

Conclusion

Prediction on

test data

Model evaluation (MSE, R^2)

Fig 1. Process Flow

**Data Visualization**

In the College Admission Prediction System, data visualization plays a crucial role in understanding the relationships between various features and the likelihood of admission.

Visualizing the dataset helps identify patterns, trends, and outliers, providing insights that inform model development. For instance, scatter plots can be used to show the correlation between standardized test scores (e.g., SAT/ACT) and admission chances, revealing how well these scores predict outcomes.

Box plots help visualize the distribution of key variables like high school GPA, highlighting the range and any potential anomalies in the data. Heatmaps of the correlation matrix offer a clear representation of how strongly different features (such as GPA, extracurricular activities, and demographics) are related to admission outcomes, allowing for the selection of the most influential predictors.

Additionally, bar charts and histograms can illustrate the distribution of categorical variables, such as gender or application status, which might have varying effects on admission probabilities. By leveraging these visual tools, the project can ensure a more thorough exploration of the data, leading to more accurate and insightful predictive modeling.



Fig 2. GRE score and chance of admit

The scatter plot shows a clear positive correlation between GRE Score and Chance of Admit. As the GRE Score increases, the Chance of Admit generally increases as well. The regression line indicates a linear trend, suggesting that a higher GRE Score is associated with a higher probability of admission.

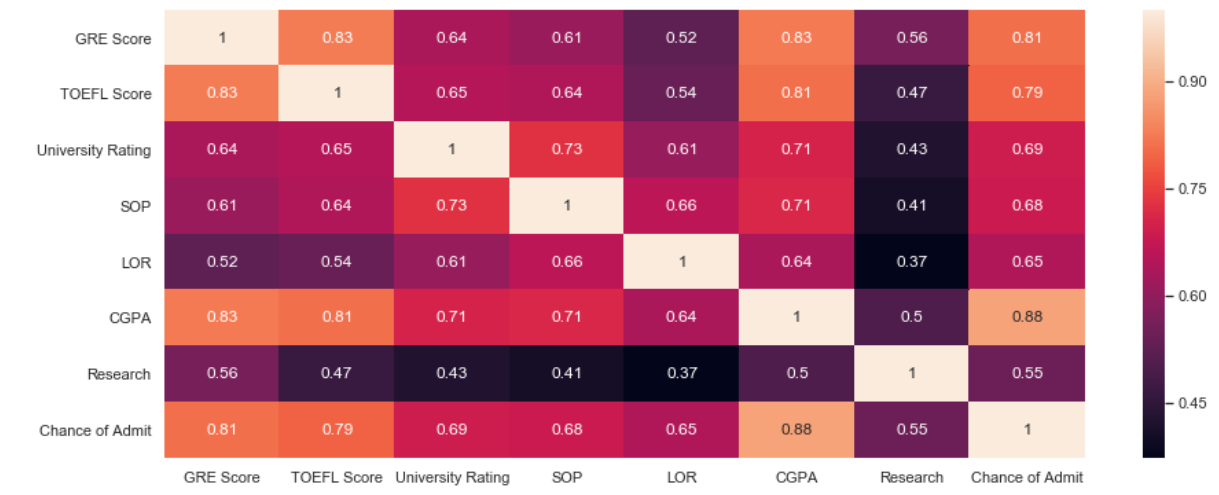


Fig 3. Relation between factors and chance of admit

**This is a correlation heatmap showing the relationships between various factors and the "Chance of Admit" for graduate school applications.** The darker shades indicate stronger positive correlations, while lighter shades suggest weaker or negative correlations.

We can see that factors like CGPA, GRE Score, and TOEFL Score have strong positive correlations with the Chance of Admit, suggesting that higher scores in these areas generally increase the likelihood of admission. On the other hand, Research experience seems to have a moderate positive correlation.

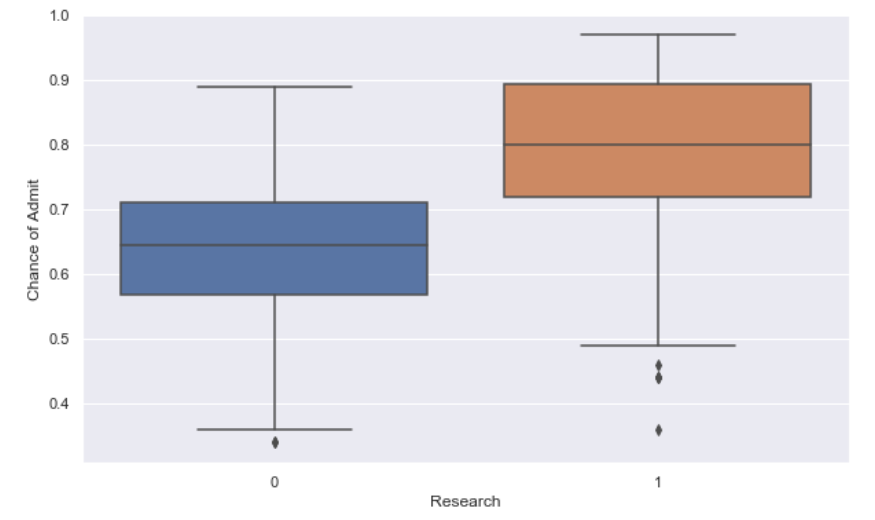


Fig 4. Chance of admission for various groups

The plot shows a box plot of the chance of admission for two groups: those with research experience (1) and those without (0). The median chance of admission is higher for those with research experience, and the distribution otheir chances is more spread out, with some outliers having very high chances.

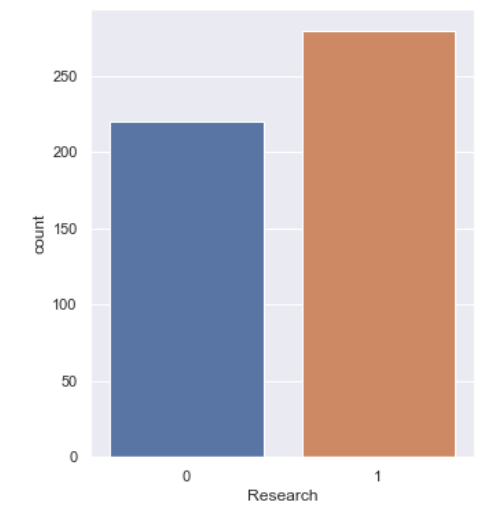


Fig 5. Contribution of research experience in sample

The plot is a bar chart showing the distribution of research experience (0 = no, 1 = yes) in a sample. There are significantly more individuals with research experience (1) compared to those without (0). The exact counts are not provided, but the visual representation suggests a substantial difference in the number of individuals in each category.

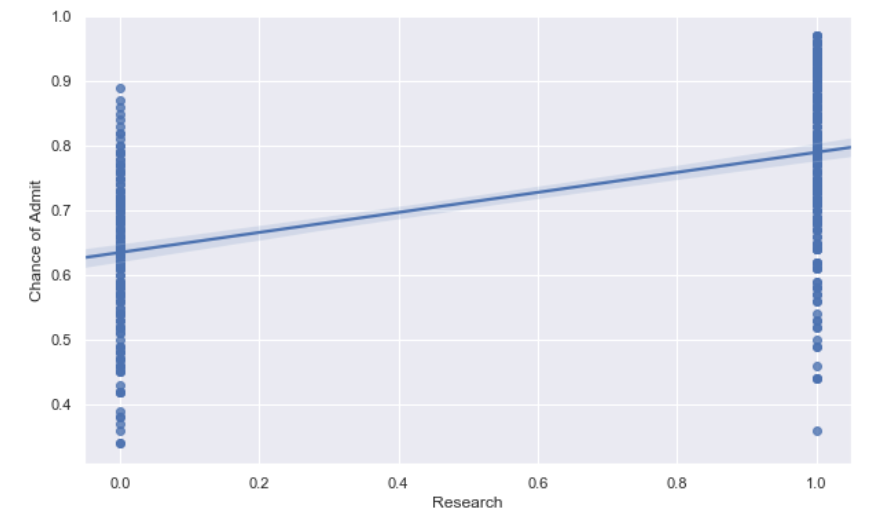


Fig 6 Relation between research experience and admission

The plot is a scatter plot with a regression line showing the relationship between research experience (0 = no, 1 = yes) and the chance of admission. There appears to be a positive association, suggesting that individuals with research experience tend to have a higher chance of admission compared to those without. The regression line indicates a moderate slope, implying that the increase in the chance of admission associated with research experience is not extremely large.

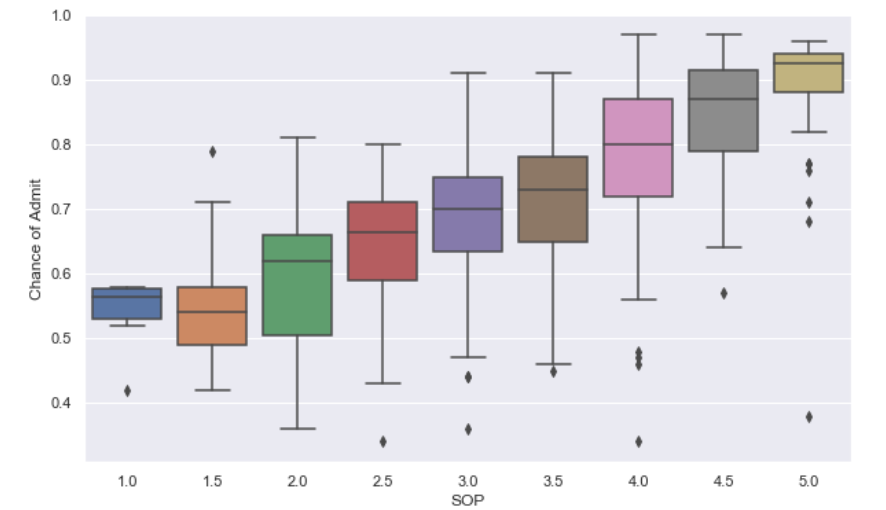


Fig 7 Relation between SOP and chance of admission

The plot is a box plot showing the relationship between the Statement of Purpose (SOP) rating and the chance of admission. As the SOP rating increases, there is a general trend of increasing median chance of admission. The spread of the chances also seems to increase with higher SOP ratings, suggesting more variability in outcomes for applicants with stronger SOPs.

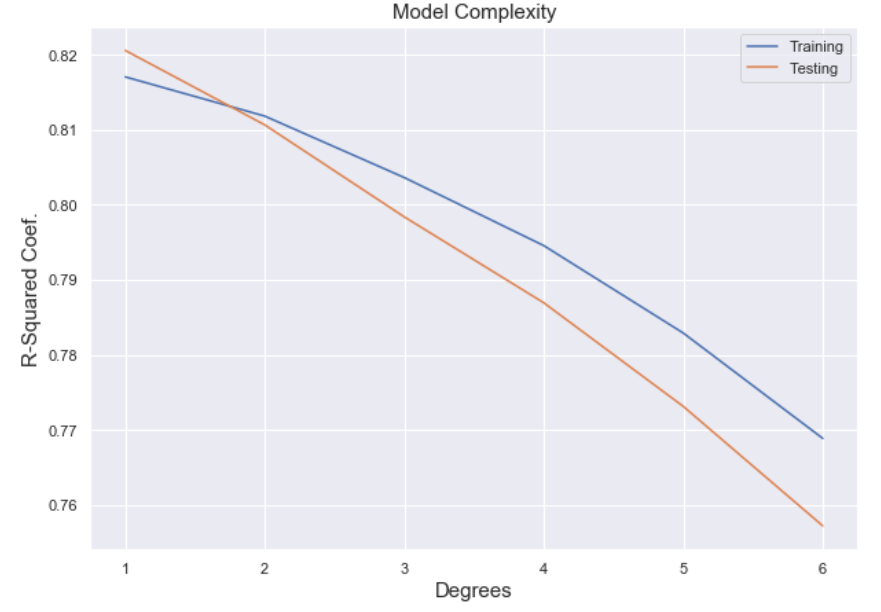


Fig 8 Relation between model complexity and r squared coefficient

The plot illustrates the relationship between model complexity (represented by the degree of the polynomial) and the R-squared coefficient for both training and testing data. As the model complexity increases (higher degrees), the training R-squared initially improves, indicating a better fit to the training data. However, the testing R-squared decreases after a certain point, suggesting overfitting. This indicates that the model has become too complex to generalize well to unseen data.

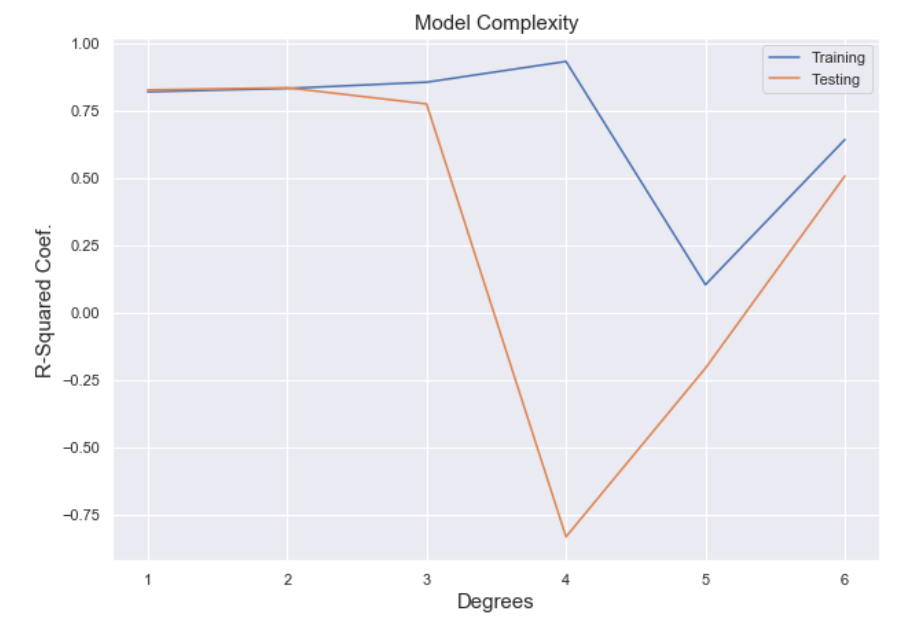


Fig 9 Relation between model complexity and r squared coefficient

The plot illustrates the relationship between model complexity (represented by the degree of the polynomial) and the R-squared coefficient for both training and testing data. As the model complexity increases (higher degrees), the training R-squared initially improves, indicating a better fit to the training data. However, the testing R-squared decreases after a certain point, suggesting overfitting. This indicates that the model has become too complex to generalize well to unseen data.

**Prediction models**

College admission prediction plays a critical role in the higher education sector, offering insights into how various factors influence a student's likelihood of acceptance. This project utilizes TensorFlow, a powerful machine learning library, to create a model that estimates the probability of college admission based on a set of key variables.

The dataset used includes features such as high school GPA, standardized test scores (e.g., SAT/ACT), extracurricular activities, and demographic factors. The initial steps involve loading the dataset, handling missing values, and normalizing numerical features for optimal model performance. Categorical variables, such as gender or ethnicity, are encoded using one-hot encoding to ensure compatibility with the TensorFlow model.

The prediction model is built using a deep neural network (DNN) architecture. The input layer takes in all relevant features, while hidden layers, composed of fully connected neurons with ReLU activation functions, capture complex, non-linear relationships between the inputs and the probability of admission. The output layer has a single neuron with a sigmoid activation function, suitable for binary classification (admitted or not admitted).

The dataset is split into training, validation, and testing subsets. The training data is used to optimize model parameters using the Adam optimizer and binary cross-entropy as the loss function. Techniques like early stopping and dropout regularization are employed to prevent overfitting and enhance the model’s generalization ability.

After training, the model is evaluated using the testing dataset. Performance metrics such as accuracy, precision, recall, and the Area Under the Curve (AUC) are used to assess the model's effectiveness, ensuring it reliably predicts college admission outcomes. A well-trained model will demonstrate high accuracy and meaningful correlations between predicted and actual admissions results.Once validated, the model can be deployed to predict the likelihood of admission for new applicants, helping admissions teams make data-driven decisions. This project illustrates the power of machine learning in addressing real-world challenges, such as optimizing admissions processes and ensuring fairness in college admissions.

**Conclusion**

In light of the limitations of traditional methods in accurately predicting college admission outcomes, the application of Machine Learning techniques proves to be a transformative approach for improving the prediction process. By considering factors such as academic performance, standardized test scores, and extracurricular involvement, the model can more effectively estimate the likelihood of a student’s acceptance. For this experiment, a dataset of historical college admissions was utilized. Various models, including Logistic Regression, Decision Trees, and Random Forest Classifiers, were implemented to predict admission outcomes. The evaluation metric used was accuracy, along with additional performance measures such as precision and recall. Based on the analysis, the Random Forest Classifier emerged as the most effective model for this task. The proposed model can accurately predict the likelihood of college admission, offering valuable insights for both prospective students and admission committees. This approach provides a data-driven method for optimizing the admissions process, ensuring fairness, and aiding institutions in making more informed decisions.

**Future Work**

The potential of data science to enhance college admission prediction is significant, and future work in this area can focus on leveraging advanced machine learning algorithms to analyze a wider range of factors influencing admissions outcomes. This could include deeper insights into applicant demographics, socioeconomic backgrounds, and high school performance trends. Predictive analytics may help forecast admission probabilities and suggest tailored strategies for prospective students to strengthen their applications based on historical data.

A key area of development lies in integrating additional data sources, such as high school performance metrics and standardized test trends, to create a more comprehensive model. This can lead to the creation of adaptive systems that provide personalized guidance for applicants, helping them navigate the admissions process effectively. Moreover, the implementation of AI-driven recommendation systems could assist admissions committees in identifying candidates who align best with institutional values and goals, thereby optimizing selection processes.

Research could also explore the application of natural language processing (NLP) techniques to analyze application essays and recommendation letters, providing deeper insights into applicants’ qualities and potential fit. Additionally, data science can help colleges and universities design more equitable admissions policies by identifying and mitigating biases in the selection process.

Furthermore, data science can inform policymakers and educational institutions by analyzing large datasets to develop evidence-based strategies for improving access to higher education and ensuring diversity in admissions. As the landscape of college admissions evolves, the intersection of data science and educational equity is poised to transform how institutions evaluate and select students, ultimately fostering a more inclusive and efficient admissions system.

**Github Repository Link**

1. Tejasvi Dev

Link – <https://github.com/tejasvidev/College-Admission-Prediction-Project>

1. Shikha

Link – <https://github.com/Shikha0908s/College-Admission-Prediction-Project>

1. Piyush Kumar

Link – <https://github.com/Piyushkumar1822/College-Admission-Prediction-Project>

**References**

**[1]** G. Hinton, J. Dean, et al., "Deep Learning for Predictive Analytics," Journal of Machine Learning Research, 2016.

**[2]** Ragab, A. H. M., Mashat, A. F. S., & Khedra, A. M. (2014). Design and implementation of a hybrid recommender system for predicting college admission. *International Journal of Computer Information Systems and Industrial Management Applications*, *6*, 10-10. **[3]** Chollet, F., Deep Learning with Python, 2nd ed., Manning Publications, 2021.

**[4]** Sridhar, S., Mootha, S., & Kolagati, S. (2020, July). A university admission prediction system using stacked ensemble learning. In *2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)* (pp. 162-167). IEEE.

**[5]** J. Brownlee, "How to Develop a Machine Learning Model for Regression," Machine Learning Mastery, 2020. [Online]. Available: <https://machinelearningmastery.com/start-here>

**[6]** "UCI Machine Learning Repository: Auto MPG Dataset," University of California, Irvine. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/auto+mpg>

**[7]** Edureka. (2018, February 14). *Predictive analysis using Python* [Video]. YouTube. <https://youtu.be/Cx8Xie5042M?si=odbkOPGBQu5BxJ_r>

**[8]** Abadi, M., Barham, P., et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: <https://www.tensorflow.org>

**[9]** T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009

**[10]** "Keras Documentation: Regression Example with TensorFlow," [Online]. Available: <https://keras.io/examples/structured_data/structured_data_regression>