Vishwakarma Institute of Information Technology, Pune

**(An Autonomous Institute Maharashtra)**

**A Report**

**on**

**A blue and red logo

Description automatically generated**

**CSE (AI) Department**

**Vishwakarma Institute of Information Technology**

**Academic Year: 2023-24**

**“Assignment 7: Classification using Decision Trees”**

**Submitted by**

|  |  |  |
| --- | --- | --- |
| **Sr No** | **Name** | **Email** |
| **1** | **Prajwal Dilip Gaikwad** | [prajwal.22211012@viit.ac.in](mailto:prajwal.22211012@viit.ac.in) |

**Under Guidance of**

Dr. Anuradha Yenkikar

# Problem Statement: -

# Every year many students give the GRE exam to get admission in foreign Universities. The data set contains GRE Scores (out of 340), TOEFL Scores (out of 120), University Rating (out of 5), Statement of Purpose strength (out of 5), Letter of Recommendation strength (out of 5), Undergraduate GPA (out of 10), Research Experience (0=no, 1=yes), Admitted (0=no, 1=yes). Admitted is the target variable.

# Data Set: https://www.kaggle.com/mohansacharya/graduate-admissions

# The counselor of the firm is supposed to check whether the student will get an admission or not based on his/her GRE score and Academic Score. So, to help the counselor to take appropriate decisions, build a machine learning model classifier using a Decision tree to predict whether a student will get admission or not.

# a) Apply Data pre-processing (Label Encoding, Data Transformation....) techniques if necessary. b) Perform data-preparation (Train-Test Split) c) Apply Machine Learning Algorithm d) Evaluate Model.

# Packages / Libraries used: -

* **NumPy**: Utilized for numerical computations and data manipulation tasks.
* **Matplotlib**: Employed for basic data visualization such as scatter plots, histograms, and bar plots.
* **Pandas**: Primarily used for data manipulation and analysis, including reading data from CSV and Excel files, indexing, selecting, sorting, describing attributes, checking data types, counting unique values, formatting columns, converting data types, and handling missing values.
* **Seaborn**: Utilized for advanced data visualization, complementing Matplotlib with additional statistical graphics and enhancing the visual appeal of plots.

# Theory: -

The task involves building a machine learning model to predict graduate admission based on GRE scores and academic performance. The dataset contains various attributes such as GRE scores, TOEFL scores, university ratings, statement of purpose strength, letter of recommendation strength, undergraduate GPA, research experience, and admission status. To aid the counselor in making admission decisions, the objective is to employ a Decision Tree classifier. Initial steps include data preprocessing techniques such as label encoding and data transformation, if necessary. Following this, the data is prepared through train-test splitting to ensure model evaluation on unseen data. The Decision Tree algorithm is then applied to the prepared dataset to create a predictive model. Finally, the model's performance is evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) score. Visualization techniques, including histograms, bar plots, and pair plots, are employed to gain insights into the data distribution and relationships between features. The decision tree's structure is visualized to understand how it makes predictions based on input features. This comprehensive approach aims to provide the counselor with a reliable tool for making informed decisions regarding student admissions, ultimately enhancing the efficiency and effectiveness of the admission process.

## Decision Tree Classification: -

* Decision Tree is a **Supervised learning technique**that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where**internal nodes represent the features of a dataset, branches represent the decision rules,** and **each leaf node represents the outcome.**
* In a Decision tree, there are two nodes, which are the **Decision Node** and**Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
* The decisions or the test are performed based on features of the given dataset.
* ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***
* It is called a decision tree because, like a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

A diagram of a decision tree

Description automatically generated

## Methodology: -

1. Data Loading and Exploration:
   * Load the dataset containing information about GRE scores, TOEFL scores, university ratings, statement of purpose strength, letter of recommendation strength, undergraduate GPA, research experience, and admission status.
   * Perform exploratory data analysis (EDA) to gain insights into the dataset's structure, distribution, and relationships between variables.
2. Data Preprocessing:
   * Handle missing values and ensure data consistency by checking for any anomalies in the dataset.
   * Encode categorical variables using label encoding if necessary and transform the data into a suitable format for modelling.
3. Feature Selection:
   * Select relevant features (in this case, GRE scores and possibly other academic performance metrics) that are likely to have a significant impact on the admission decision.
4. Data Splitting:
   * Split the dataset into training and testing sets using the train\_test\_split function to ensure model evaluation on unseen data.
5. Model Building:
   * Choose an appropriate machine learning algorithm for classification tasks, such as Decision Tree Classifier.
   * Train the model using the training data to learn patterns and relationships between features and the target variable (admission status).
6. Model Evaluation:
   * Evaluate the performance of the trained model using various metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and classification report.
   * Cross-validation techniques can also be applied to ensure the model's robustness and generalization ability.
7. Visualization:
   * Visualize the decision tree structure to understand how the model makes predictions based on input features.
   * Plot histograms, bar plots, and pair plots to visualize the distribution of features and explore relationships between variables.
8. Conclusion and Recommendations:
   * Summarize the findings from the model evaluation and visualization, discussing the model's performance and insights gained from the analysis.
   * Provide recommendations or insights to the counselor based on the model's predictions and evaluation results, helping them make informed decisions regarding student admissions.

## Application: -

1. Admissions Decision Support:
   * The model serves as a valuable tool for university admissions offices and counselors, assisting them in making informed decisions about admitting students based on their academic profiles. By analyzing GRE scores and other relevant factors, the model can provide insights into a student's likelihood of admission.
2. Resource Allocation:
   * Universities can use the model to optimize resource allocation by targeting outreach efforts and scholarships towards students who are more likely to be admitted. This ensures efficient allocation of resources and maximizes the university's impact on student success.
3. Student Counselling:
   * Guidance counsellors can utilize the model to provide personalized advice and guidance to students regarding their chances of admission to specific universities or programs. This helps students make informed decisions about their academic and career paths.
4. Educational Policy Development:
   * Education policymakers can leverage the model's insights to develop effective educational policies aimed at improving access to higher education and promoting diversity and inclusion. By understanding the factors influencing admissions decisions, policymakers can design interventions to address disparities and improve equity in higher education.
5. Research and Analysis:
   * Researchers in the field of education can use the model to conduct analyses and studies on admissions trends, the impact of admissions criteria on student success, and other related topics. The model provides a quantitative framework for studying admissions processes and their outcomes.

# Diagram: -



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

## Conclusion: -

In conclusion, the development of a machine learning model for predicting graduate admissions based on GRE scores and academic performance offers significant benefits and opportunities across various domains. By leveraging predictive analytics and advanced algorithms such as Decision Trees, the model provides valuable insights into students' likelihood of admission, aiding university admissions offices, counselors, policymakers, and educators in making informed decisions. Through data preprocessing, feature selection, model training, evaluation, and visualization, the model offers a structured approach to analyzing admissions data and understanding the factors influencing admissions decisions. The model's applications extend beyond academia, with potential uses in resource allocation, counseling, policy development, research, and personalized learning. Overall, the model represents a powerful tool for enhancing efficiency, equity, and effectiveness in the admissions process, ultimately contributing to improved outcomes and opportunities for students and institutions alike.