EDS - 6344 ARTIFICIAL INTELLIGENCE FOR ENGINEERS - FINAL REPORT
GROUP 9 – STUDENT'S PERCEPTION OF AI IN EDUCATION

Contents

GROUP 9

Charan Devapatla– 2215421
Hruthika Jinna - 2151639
Leela Prasad Boddu - 2214707
Nikhitha Reddy Kadapakonda – 2162999
Veera Sampath Reddy Alavalapati - 2199279

EDS - 6344 ARTIFICIAL INTELLIGENCE FOR ENGINEERS - FINAL REPORT
GROUP 9 – STUDENT'S PERCEPTION OF AI IN EDUCATION

---

1. INTRODUCTION:

The dataset contains student comments, thoughts, and responses to questions about various applications of AI in educational settings. The data is gathered from surveys, interviews, and students across different education levels such as primary, secondary, and higher education.

This Dataset is taken from Kaggle, and the survey consists of total 16 questions covering topics such as

- Students' level of AI knowledge, Sources of information, Impact of AI on society
- Advantages and disadvantages of using AI in education.
- Demographic information such as gender, year of study, major, and academic performance.

Dataset link: https://www.kaggle.com/datasets/gianinamariapetrascu/survey-on-students-perceptions-of-ai-in-education?resource=download

ATTIBUTES:

Input features:

- Student ID
- AI_knowledge – 1 means not informed, 10 means extremely informed.
- AI_sources – Sources used to learn AI.
- Internet, Books & paper, social media – 1- Students interact, 0 – No Interaction
- Discussion – Discussion with family/friends is one of the option chosen (1- Yes, 0-No)
- AI_dehumanization – 1 – strongly disagree, 2- partially disagree, 3- Neutral, 4- partially agree, 5- fully agree.
- Job replacement – Students opinions of AI on future employment opportunities. - 1 – strongly disagree, 2- partially disagree, 3- Neutral, 4- partially agree, 5- fully agree.
- Domains - Education, Medicine, Agriculture, Constructions, Marketing, Arts, Administration
- GPA, passed_exams, Major, year of study, Gender etc

Target variable:

Utility Grade

- Perception of low usefulness (Ratings between 0 to 5)
- Perception of high usefulness (Ratings between 6 to 10)

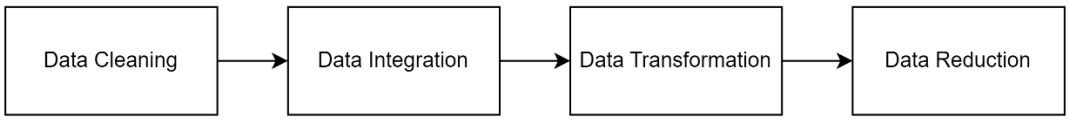
2. PROBLEM STATEMENT AND OBJECTIVE

Problem statement: The issue is that there is a lack of knowledge on how students see artificial intelligence (AI) in the classroom. Understanding how students view the role of AI in education, their attitudes toward its application, and the possible effects it may have on their learning experiences is important. In order to successfully incorporate AI technology into educational settings and resolve any concerns or limitations that may prevent its successful acceptance, educators, policymakers, and AI developers must first understand these attitudes.

Objective: The main objective is to gather and analyze data that provides insights into various aspects, including Attitude towards AI, Acceptance and Adoption of AI and etc. To develop a machine learning model that can accurately predict the utility grade of ai in education from the perspective of students based on their responses to survey questions.

3. DATA PREPROCESSING:

Data needs to be cleaned through some preprocessing techniques to convert raw data into cleaned dataset for further analysis like model performances.



Forms of Data Preprocessing

Data cleaning is a preprocessing approach to clean the raw data by smoothing the noisy data, replacing missing values with averages, and eliminating any outliers. The actions we took in the project's initial phase are listed below.

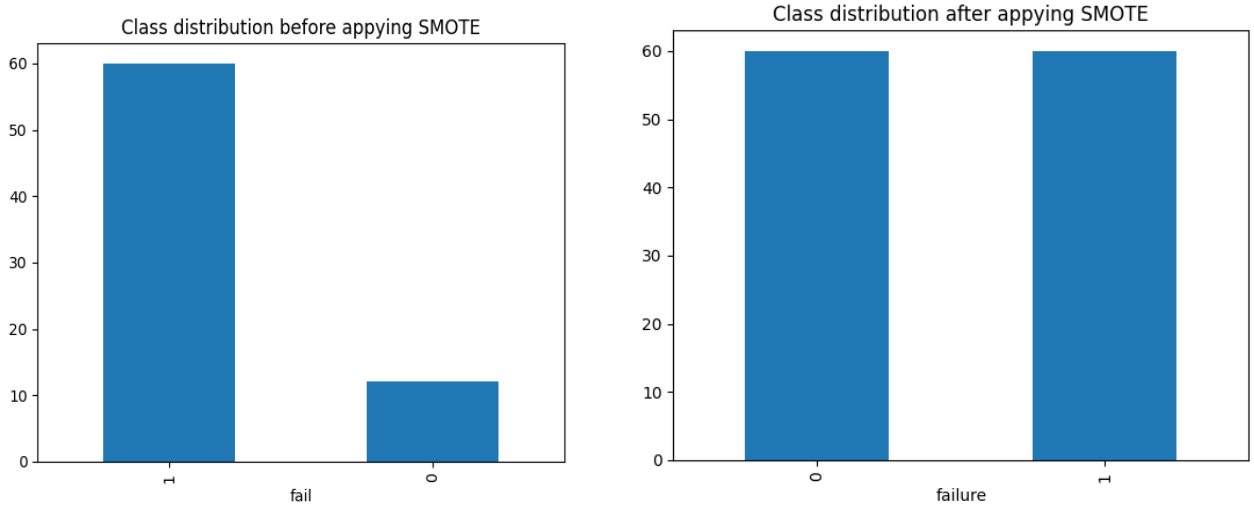
Data Cleaning:

- Checked for the Null values in the dataset, but the data set doesn't contain any Null values.
- Checking if there are any missing, duplicate values present in the dataset to replace it with average values. It doesn't contain any.
- Label encoded the columns using function "multilabelbinarizer", one hot encoding using "get dummies" and dropped the unnecessary columns.
- Used the statistical concepts of interquartile range to identify the outliers in the dataset, but no outliers identified.
- Applied **SMOTE** technique to handle imbalance data points in the target variable.
- Split the Training and Test data into 80 and 20 percent.
- Correlation – Selected the best 10 features from the data using correlation.
- Perform scaling on data using a standard scaler. (StandardScaler is used to resize the distribution of values so that the mean of the observed values is 0 and the standard deviation is 1)

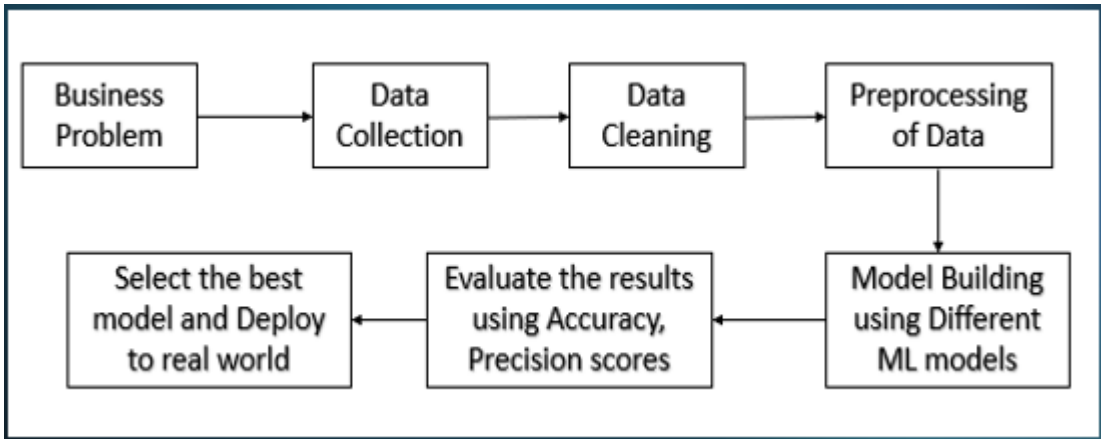
```
Q6#1.Education              0.545814
Q1.AI_knowledge_2           0.379371
Q3#3.Problem_solving_2      0.352369
Q3#3.Problem_solving_5      0.349193
Q9.Advantage_learning_1     0.303660
Q4#3.Economic_growth_5      0.295027
Q5.Feelings_2               0.291249
Q16.GPA                     0.276962
Q3#2.Job_replacement_4      0.271233
Name: Q7.Utility_grade, dtype: float64
```

```
#Scaling the data

# create a StandardScaler object
scaler = StandardScaler()

# fit the scaler to the dataset
scaler.fit(X_train)

# transform the dataset using the fitted scaler
X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

Target class distribution before applying the SMOTE and after that is as follows:



Methodology: Here the workflow started by Identifying the Business problem and then collected the data points. The next step after collected the data is Cleaning & preprocessing of the data, in this stage data points were evaluated for Missing/Null values, removing the outliers if we find any. And then identified the important features using different techniques one in that is Correlation, And Model building is carried out after feature selection. Evaluating the results from all the different models, comparing and choosing the best model is the final stage in the workflow.



4. VISUALIZATION & EXPLORATORY DATA ANALYSIS

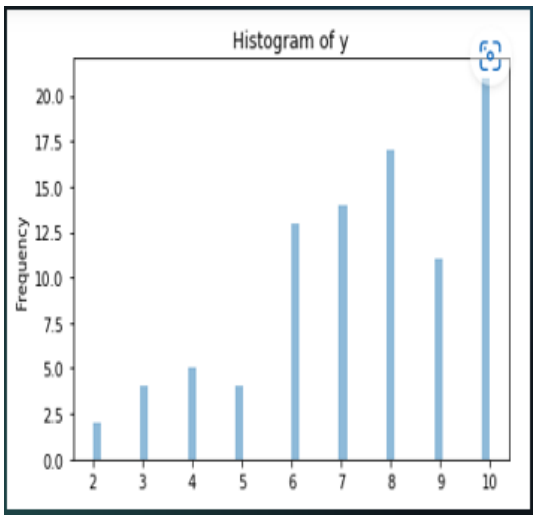
Data visualization allows us to view how the data is organized and what sort of correlations the features of the data have. It is the fastest technique to determine whether the characteristics meet the output.

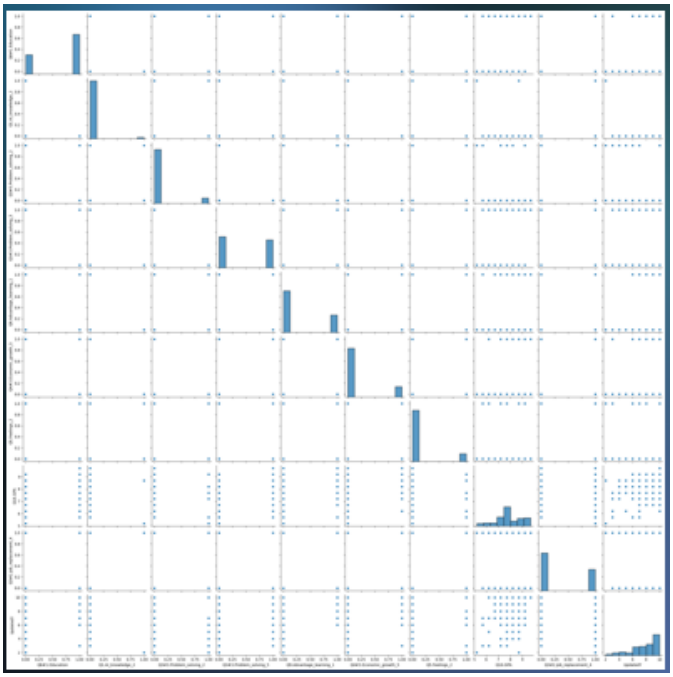
Histogram plot: It provides us with a count of the number of observations for each grade in the target variable.

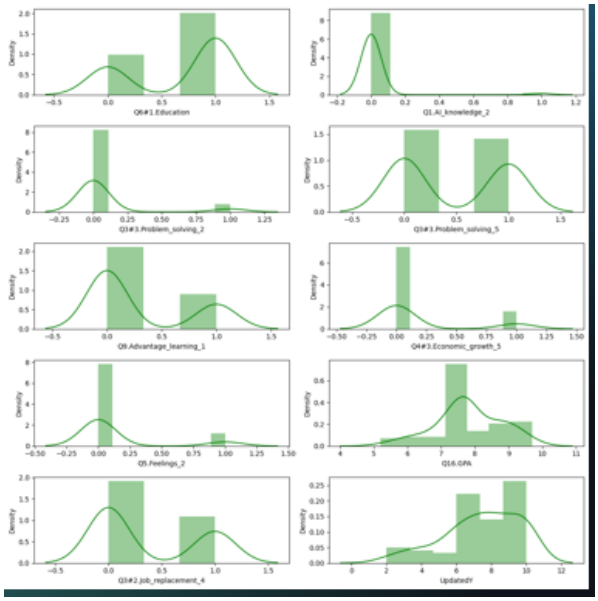
Pair plot: It allows to find the distribution of single variable and relationships between two variables.

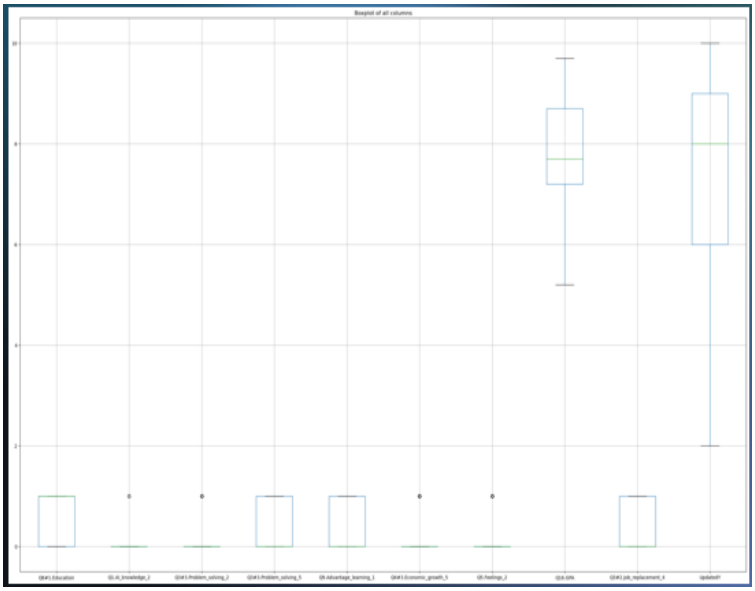
Distribution plot: Used to assess the distribution of each feature with another.

Box plot: Used to identify the outliers by using Interquartile range.









5. RESULTS OF All DIFFERENT MODELS

After preprocessing the dataset, we used the following models to predict the target variable:

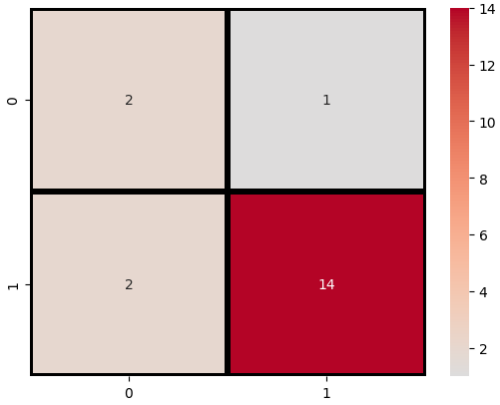
a. Logistic Regression: The main aim is to predict the probability of a given instance belonging to that class or not and it transforms the continuous value output of linear regression to categorical value output using sigmoid function. Since the result of a logistic regression is constantly dependent on the total of the inputs and parameters, it is considered as a generalized linear model. Alternatively, the output cannot be

dependent on the product of the parameters. Logistic regression is a classification model. After training the data using best features got from the correlation the **Accuracy score is 0.84**, Then perform the GridSearch CV technique (with Hyper parameters - {'penalty': ['l1', 'l2'],'C': [0.1, 1]} to get best params. The **Accuracy score** got from **GridSearch CV is 0.79** with best params - {'C': 0.1, 'penalty': 'l2'}

The confusion matrix, classification reports are as follows:

LogReg: Accuracy=0.842, Precision=0.865, Recall=0.842, F1-Score=0.851

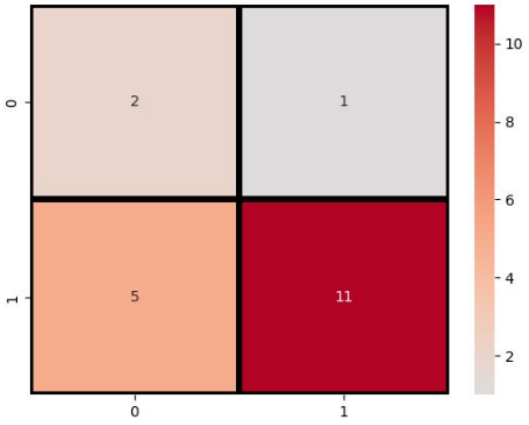
	precision	recall	f1-score	support
0	0.50	0.67	0.57	3
1	0.93	0.88	0.90	16
accuracy			0.84	19
macro avg	0.72	0.77	0.74	19
weighted avg	0.86	0.84	0.85	19



b. Decision Tree: Decision tree is a type of supervised machine learning algorithm and works well with both qualitative and quantitative features. It is a flowchart like tree structure, uses set of rules to make decisions similar to how humans make decisions. Here source set divides based on the attribute value test and the process repeats until subset at a node all has the same values. Here each internal node – attribute, Branch - outcome from the test and leaf node- class label. It doesn't require feature scaling. **Classification Tree** classifies the data into **categories**, whereas Regression Tree predicts numerical value. Entropy or Gini Index are used to build the decision tree.
The **accuracy scores** using GridSearch CV (Hyper params - {'max_depth': [1,2,5],'min_samples_leaf': [2,4,5]} is **0.74** with best parameters {'max_depth': 2, 'min_samples_leaf': 2} whereas the accuracy score using highest correlated features is **0.68**
Below figures represents confusion matrix and classification report by using Decision Tree Model:

DTree: Accuracy=0.684, Precision=0.817, Recall=0.684, F1-Score=0.725

	precision	recall	f1-score	support
0	0.29	0.67	0.40	3
1	0.92	0.69	0.79	16
accuracy			0.68	19
macro avg	0.60	0.68	0.59	19
weighted avg	0.82	0.68	0.72	19



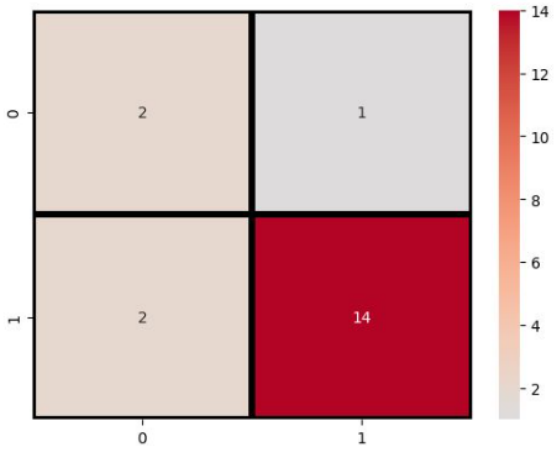
c. Random Forest: Random Forest is an ensemble model made up of many decision trees and it is one of the most powerful machine learning algorithms used for both classification and Regression. It takes the multiple subsets of data from the original dataset and built Decision tree for each set and uses average or aggregate of those to makes predictions. This technique is called **Bagging (Bootstrap + Aggregation).**

The **Accuracy score is 0.84** using the correlated features and it is **0.79** using Best params ({'max_depth': 2, 'min_samples_split': 2, 'n_estimators': 150}) from GridSearch CV

The confusion matrix, classification reports are as follows:

RF: Accuracy=0.842, Precision=0.865, Recall=0.842, F1-Score=0.851

	precision	recall	f1-score	support
0	0.50	0.67	0.57	3
1	0.93	0.88	0.90	16
accuracy			0.84	19
macro avg	0.72	0.77	0.74	19
weighted avg	0.86	0.84	0.85	19

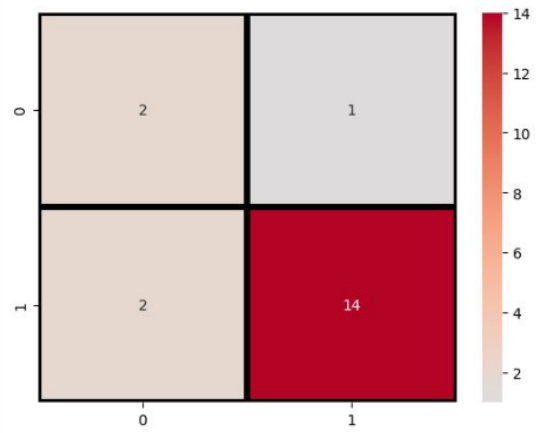


d. SVM with linear kernel: It is mainly used for Classification of any kind of data complex, medium and small. Example face detection, image classification and other. By using SVM, classes can be separated in the best possible approach using maximum margin concept. When the data can be split using a single line, a linear kernel is used. It is one of the most common kernels to be used, when the dataset has a large number of features. SVM can also be used for nonlinear classifications using Nonlinear Kernel if the data is not linearly separable.

The accuracy scores using best correlated features from correlation and best params ({'C': 0.1, 'gamma': 1, 'kernel': 'linear'}) from GridSearch (Hyper params: {'kernel': ['linear', 'rbf'],'C': [0.001,0.01,0.1, 1],'gamma': [1, 10]}) are as 0.84 and 0.84 respectively.

SVM-Linear: Accuracy=0.842, Precision=0.865, Recall=0.842, F1-Score=0.851

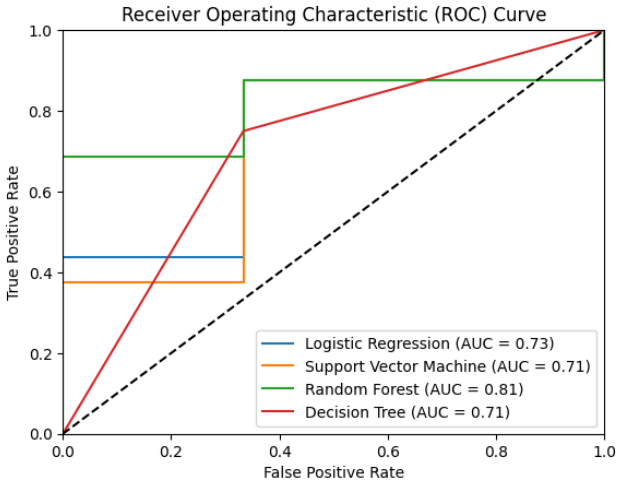
	precision	recall	f1-score	support
0	0.50	0.67	0.57	3
1	0.93	0.88	0.90	16
accuracy			0.84	19
macro avg	0.72	0.77	0.74	19
weighted avg	0.86	0.84	0.85	19



6. RESULTS AND COMPARISON

To compare the performance of different models plotted ROC (Receiver Operating Characteristics) curve and observed that (Area under the curve) AUC is highest for the model Random Forest. By this we have concluded that Random Forest is the best model to predict the Utility Grade target variable. And the least accuracy score is for the model Decision Tree, which is 0.68.

The Accuracy for this model is 0.84 and the classification report is as follows:

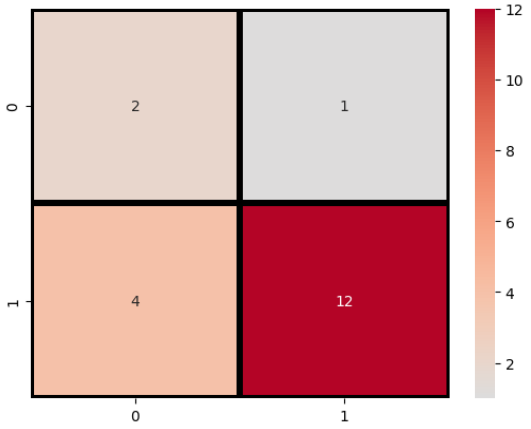


	precision	recall	f1-score	support
0	0.50	0.67	0.57	3
1	0.93	0.88	0.90	16
accuracy			0.84	19
macro avg	0.72	0.77	0.74	19
weighted avg	0.86	0.84	0.85	19

After feature importance, evaluated the Accuracy scores by considering top 1 feature i.e, Education as input feature and utility grade as target variable by using Random Forest classifier. The Accuracy score is 0.74. Similarly evaluated the results by considering Education and GPA as input feature and utility grade as target variable using Random Forest classifier. The Accuracy score is 0.74

The results of the confusion matrix and classification report is as follows:

	precision	recall	f1-score	support
0	0.33	0.67	0.44	3
1	0.92	0.75	0.83	16
accuracy			0.74	19
macro avg	0.63	0.71	0.64	19
weighted avg	0.83	0.74	0.77	19

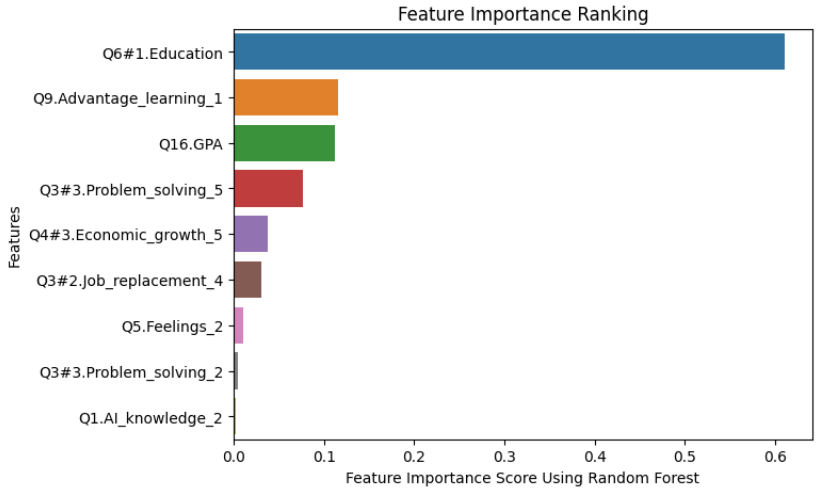


7. CONCLUSION & CHALLENGES

Among all the selected 10 features from correlation, **Education** is the most important feature. GPA and Advantage learning are second and third priorities respectively. And AI Knowledge is the least or minimum important feature out of all, and we also concluded that without this feature we can achieve good accuracy results in predicting the target variable.

Considering the above results using top1 feature and top2 features as input feature, the accuracy score is 0.74 and by considering all top 10 features from correlation, the accuracy score is 0.84. By this we can conclude that using all correlated features gives best results when compared to using individual or top 2 features.

Below is the feature importance graph:



Challenges: One of the main challenges in predicting the target variable is **Data Insufficiency**. As data points are fewer, making predictions will become one of the toughest parts. The other challenges are Privacy concerns, while collecting the data from the students needs to focus on privacy, And Students may not provide accurate responses due to other reasons. Similarly, the other challenges are Data Collection, and it is time dependent i.e, students' opinions of AI in the educational system may change over time.

8. ACKNOWLEDGEMENT

First and foremost, we wish to convey our heartfelt thanks to our supervisor Dr. Rasiah Loganantha raj, for their important advice and assistance during this project. Their knowledge and insight were essential in determining the course of our study, and their constructive remarks allowed us to improve our techniques and solutions. We would also like to thank Teaching assistance Samaikhya for constant support and valuable insights. Not but not least we would also like to thank our members of the team, without individual support it won't be possible to get good results. And finally, we hope that this paper will provide a good understanding of evaluating or predicting the target variable Utility Grade, based on this we can predict that AI is useful in education or not as per students' perception.

9. REFERENCES

1. https://www.kaggle.com/datasets/gianinamariapetrascu/survey-on-students-perceptions-of-ai-in-education?resource=download
2. Kevin_P._Murphy_Machine_Learning_A_Probabilistic_Perspective_AI_page_icog-labs.pdf (ucsd.edu)
3. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition [Book] (oreilly.com)
4. https://files.eric.ed.gov/fulltext/EJ1297477.pdf
5. https://files.eric.ed.gov/fulltext/ED590669.pdf
6. https://www.pewresearch.org/internet/2022/03/17/how-americans-think-about-artificial-intelligence/
7. https://journals.sagepub.com/doi/10.1177/21582440221100463