**Report on Root2ai Dataset**

Name: Saurabh Narvekar

Approach to solve the problem

The data contains texts so for classifying text in this problem Word Count- Count Vectorizer is used and after that term frequency calculation is done.

**Model Interpretation**

Steps

1. All the libraries required for this model are imported.
2. The data is imported
3. The data is then split into training and testing set. The size of the training set if 0.70 and testing set is 0.30.
4. Word-Count-CountVectorizer: It assigns unique number to each word which is known as Tokenizing. It counts the occurrence of each word. It is done by CountVectorizer.transform(). It stores array and it shape
   1. Term Frequency Inverse Document Frequency :
   2. Frequency: This summarizes how often a given word appears within a document.
5. Inverse Document Frequency:This downscales word that appears a lot across documents.
6. It leans the IDF from the count matrix obtained from CountVectorizer.
7. Then it applied to a classifier model: Here 2 models are used Naive Byes Classier ana Random Forest Classifier.
   1. Naïve Bayes Classifier: SKLEARN has inbuilt Multinomial Naïve Bayes Classifier Package, using this package we can directly train our model with matrix obtained from TDIF Transformer .
   2. Random Forest Classifier : A random forest is a meta estimator that fits a number of decision tree classifiers or an ensemble of Decision Tree Classifier, using this we can directly train our model with matrix obtained from TDIF Transformer.
8. Result:
   1. Training and testing accuracy of the model is obtained
   2. Classification and Confusion Matrix report if obtained

**Train and Test Accuracy Score**

Naïve Bayes Classifier Model:

Training Accuracy: 0.712

Testing Accuracy: 0.712

Random Forest Classifier Model:

Training Accuracy:0.994

Testing Accuracy: 0.994

Limitation of the model:

The model struggles to classify between few classes like Neobanks, Reg Tech Robo Advising , FinTech, Stock Trading.