**Settyl Data Science And Machine Learning Engineer Task**

**DOCUMENTATION:**

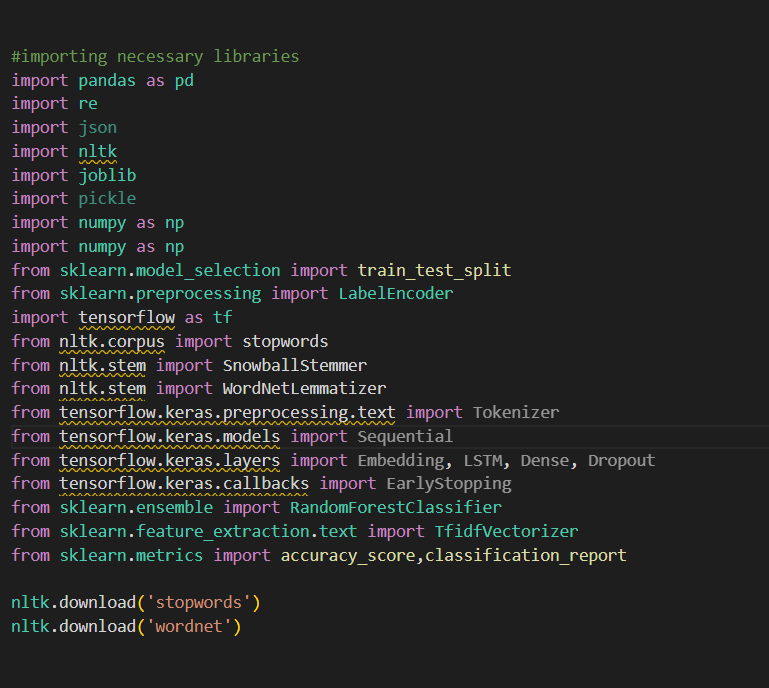
**STEP1:**

1. The Github link provided was referred to download the “Dataset.json” file from the repository
2. Analysed the data present in the dataset and came to a conclusion that the prediction of “internalStatus” can be done by adapting a classification-based Machine Learning algorithms.

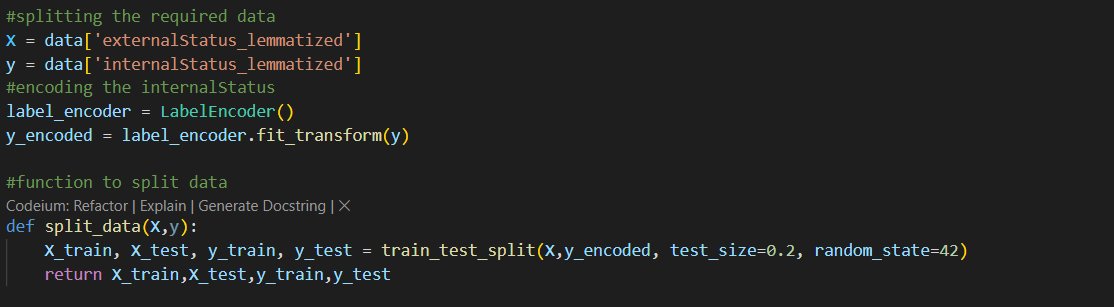
**STEP 2:**

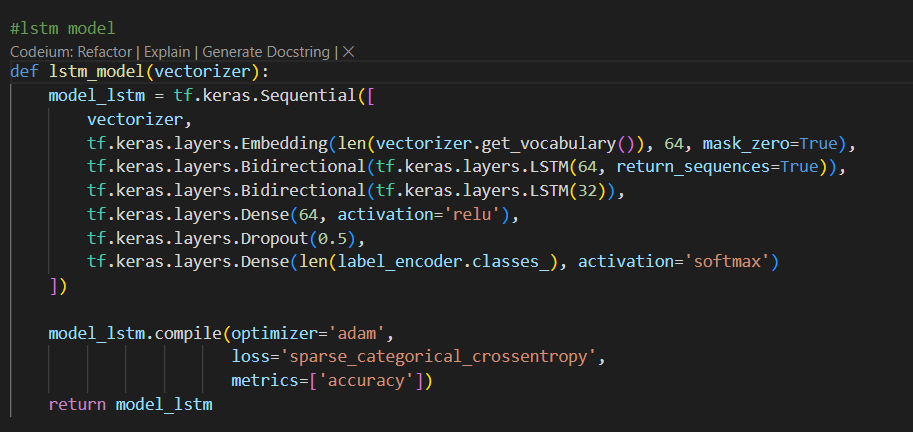
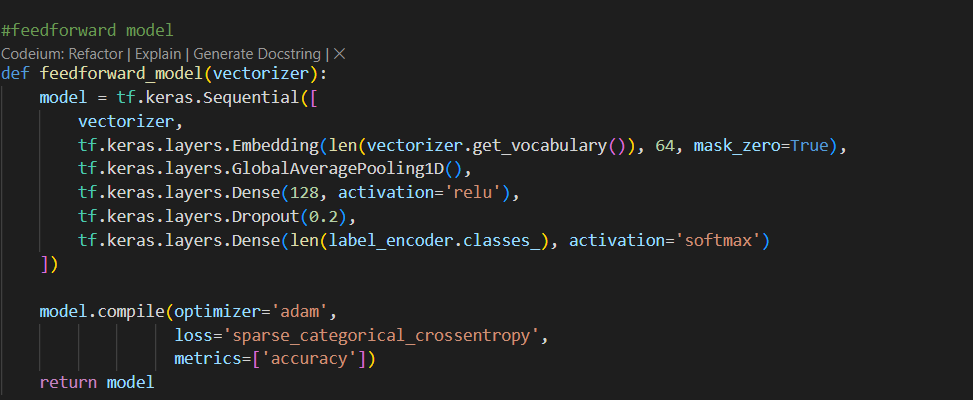
1. After careful consideration I selected 3 models to carry out the prediction task
2. I chose Feedforward, LSTM , Random forest Classifiers for the same.

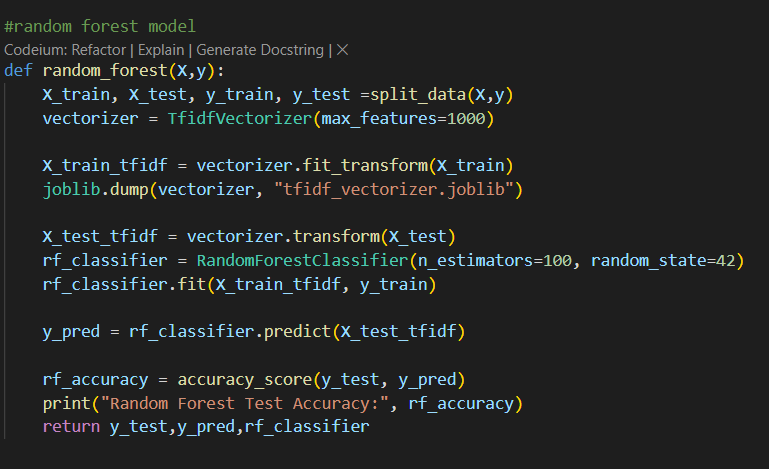
**STEP 3:**

1. I carried out the execution part by first framing out the necessary and highlighted things of my project.
2. Then I imported the necessary modules iteratively into the project file.
3. Imported the dataset into the project using Pandas.
4. Then as the primary step, I performed preprocessing of the “externalStatus” and “internalStatus”.
5. The following preprocessing methods were followed:
   1. Tokenization
   2. Normalizing the input-converted into lower case
   3. Removal of special characters
   4. Removal of stop words
   5. Stemming and lemmatization of input data.

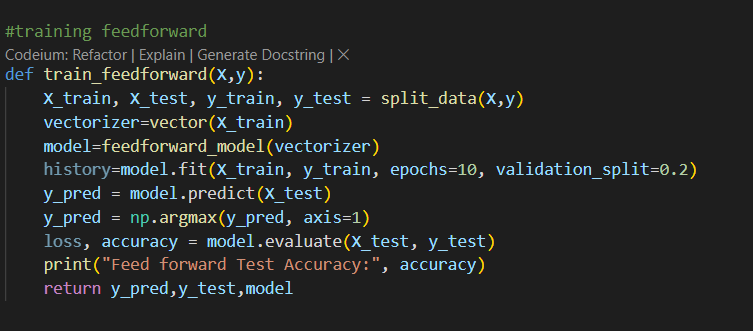
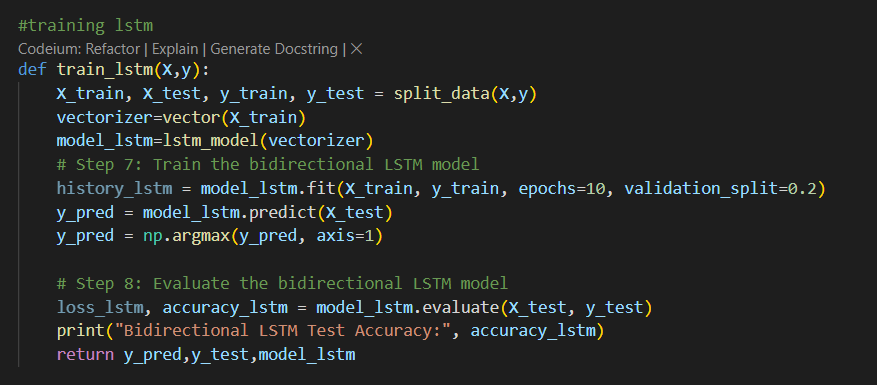


1. Now the pre-processed data is split into test and train dataset and carried out encoding so that it can be used in neural network.
2. Then the three models were created separately followed by training them with the training data that was split.



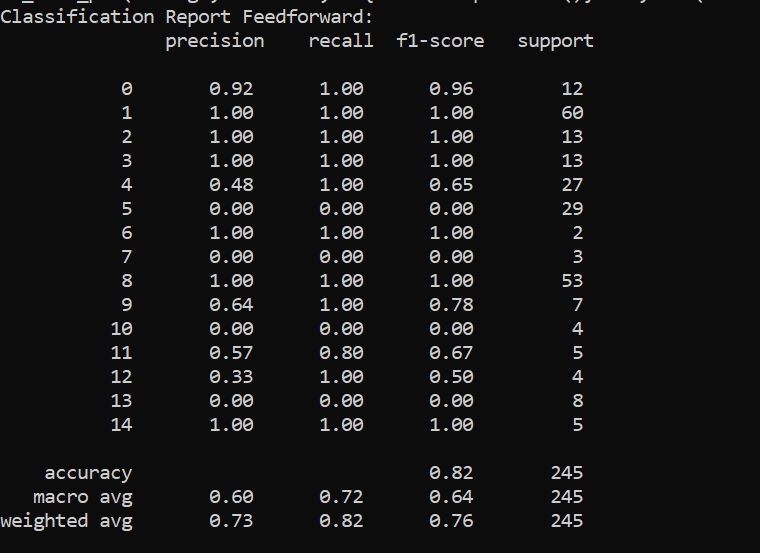
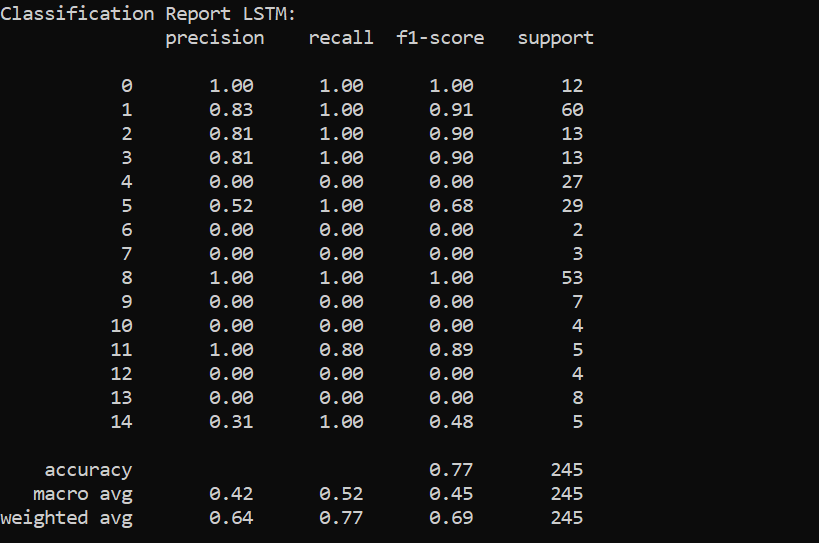


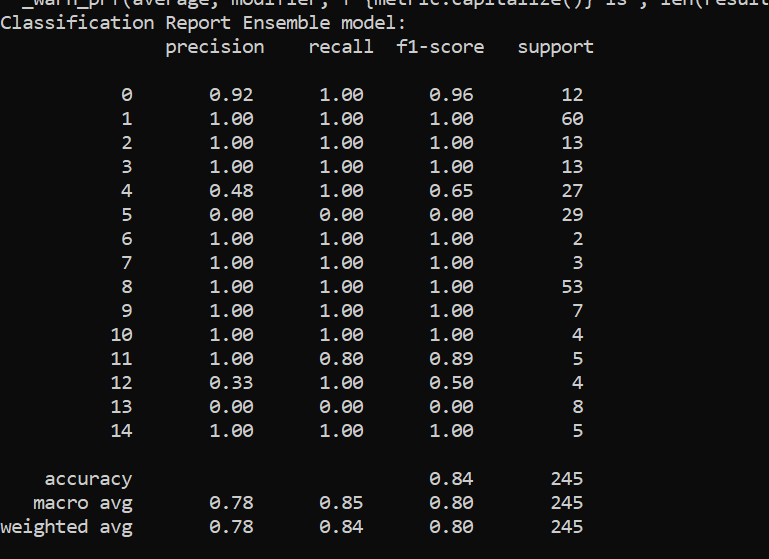
**TRAINING DATASET**:



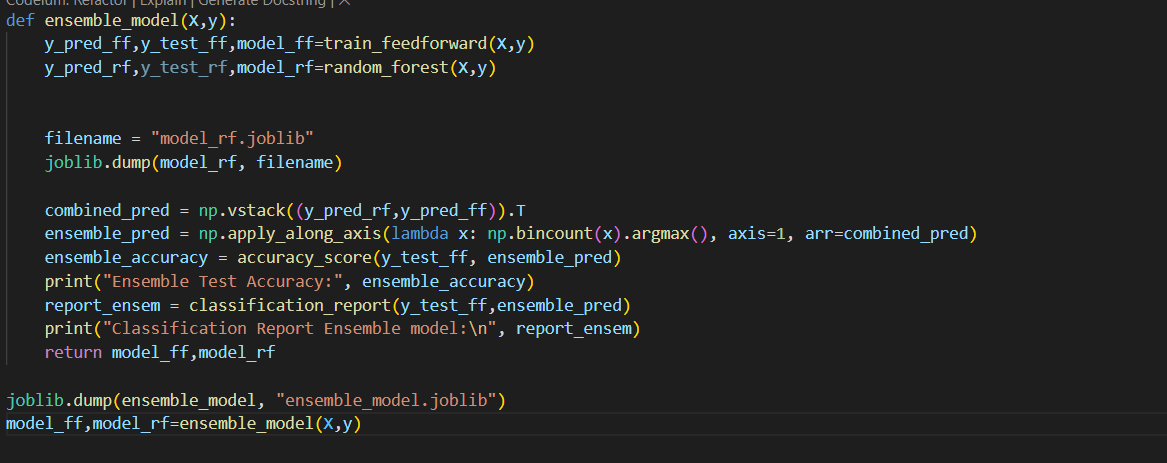
**STEP 4:**

1. The trained models were evaluated individually to compare and conclude which model works better with the data.
2. For that I tried incorporating an Ensemble model combining Random-forest and Feedforward models.
3. These are the classification report of each model:









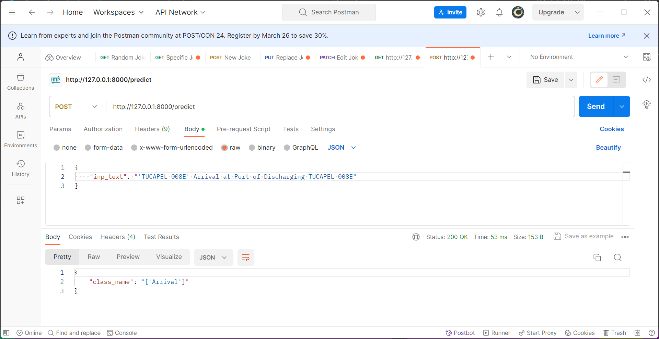
**STEP 5:**

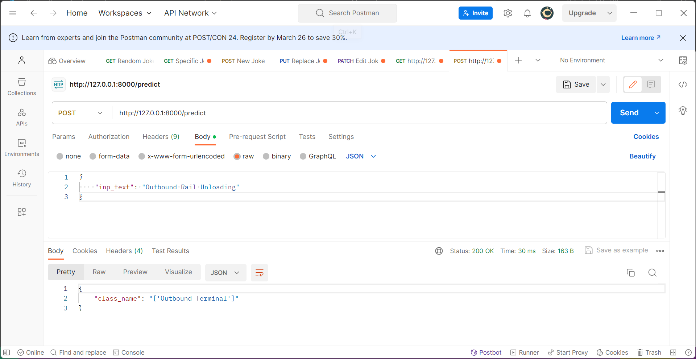
1. After careful consideration of Randomforest for prediction of test values because there isn’t much difference between the ensemble model and random forest and taking into consideration of robustness of Random forest classifier I decided to proceed with rf .
2. The model of each type were saved using “joblib “ ,which were loaded in prediction part.

**STEP 6:**

1. Fast api method is used to handle the api part for this task
2. I imported the necessary modules and packages into the file.
3. Using pydantic models to create input and output class so that request and response data types can be defined properly.
4. Defined the URL handling part of the api.
5. Created a function to receive the input and predict the output.
   1. Within the function , I created another function to preprocess the input text to match the trained data.
6. Loaded the random forest model which was saved and called the model.predict function to pass the preprocessed input text.
7. The resulting class was mapped with the label\_encoded classs of the internalStatus to return the string type of predicted class.

**STEP 7:**

1. After successfully coding the modules and api, I created an python environment and checked for missing dependences and started the uvicorn server
2. Once the server was successfully initated,I tested the api using Postman and the results were positive.



**NOTE:**

* I tried to extend the project by incorporating the frontend part by including an HTML form to receive the input and pass on to the backend server and get back the output .
* A bug which occurred in the code and I am still trying to solve the implementation part
* I request that if time provided I will be able to figure out the missing part and update the team regarding the same.
* I have attached the HTML part along with the other files for reference.

**CONTACT:**

**DHARUNESWAR**

9176766351

dharun12351@gmail.com