DAV Team Recruitment Assignment Submission Report

Aditya Singh

Roll Number: 22B1844

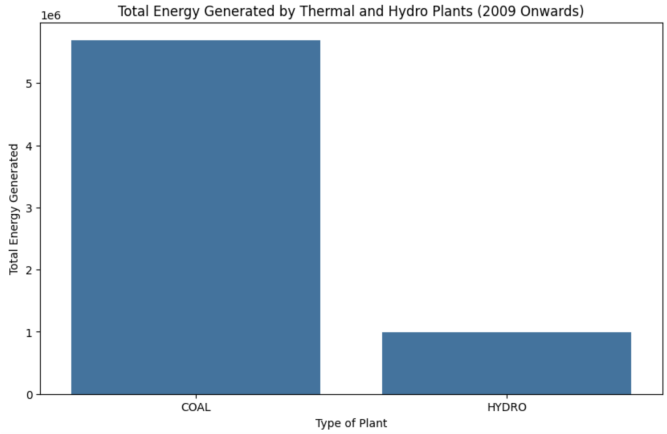
Question 1

We are given the dataset from NDAP on Power Generation in India. We need to analyse the data and ask questions. Following is the analysis:

1. **Find the installed capacities for each state and find the state with the maximum installed capacity.**
   The installed capacities for each state has been calculated in the jupyter notebook and the state with the maximum installed capacity is Maharashtra with 2224914.000 MW.

2. **Find the company with the highest number of power plants.**
   TNGDCL has the highest number of power plants in India and it has 2889 plants.

3. **Compare the actual energy generation from Thermal and Hydro Power Plants.**

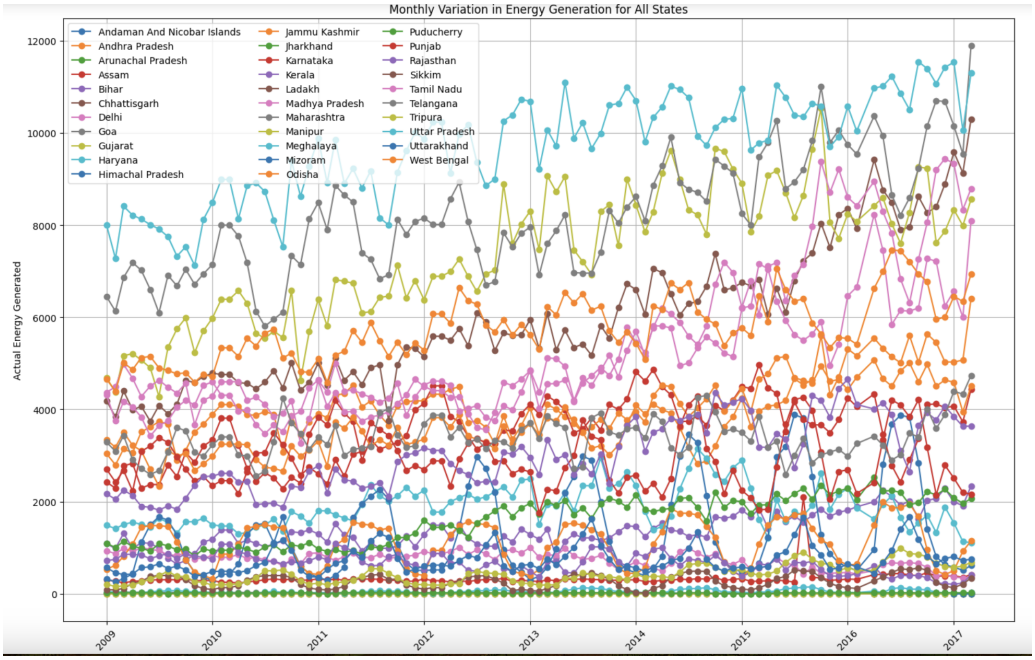


We find that the Thermal Power Plants generate 5692707.13 GWH and the Hydro Power Plants generate 986202.64 GWH. The T-

Statistic score between them is 79.52047508278449

4. **Find the Category of Plant most prevalent in each state**
   The answer for this question has been put found in the Jupyter Notebook.

5. **Find the state-wise monthly variation of energy generation 2009 onwards**
   We draw a plot for this question showing the variation in energy generation across the years.

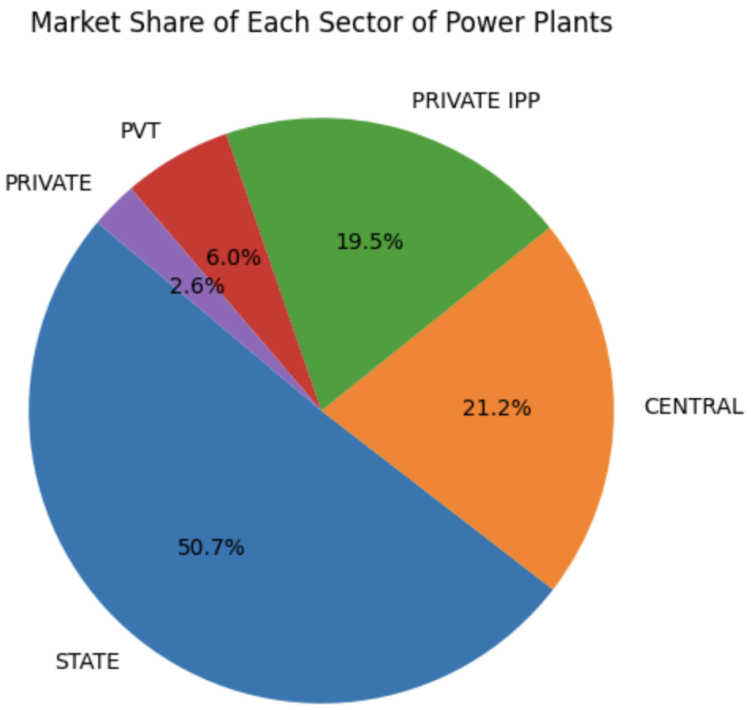


6. **Find the fraction of actual energy generated with respect to the installed capacities in each state. Find the state with the highest ratio.**
   The fractions for all states is present in the jupyter notebook. The state with the highest value of this ratio is Puducherry at 56.16%

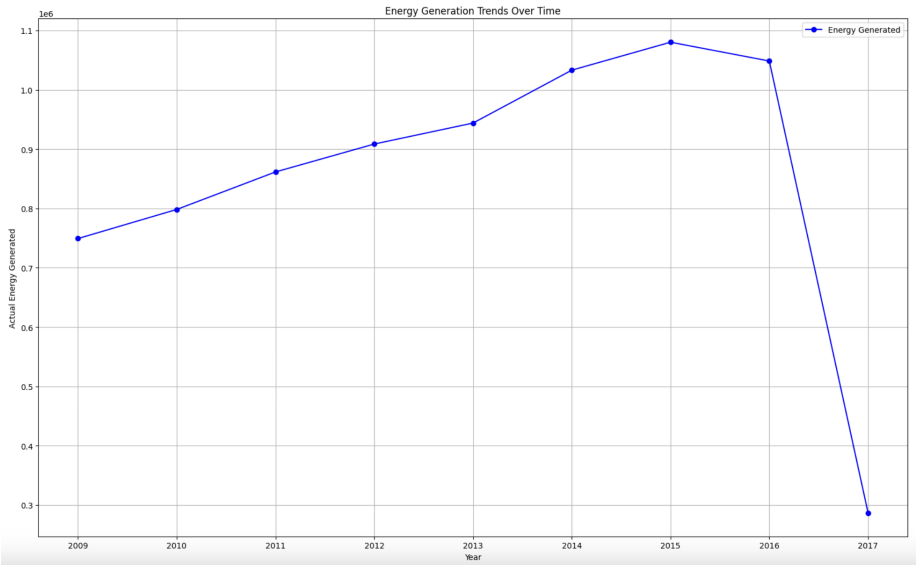
7. **Find the type of fuel used most commonly in each region.** The below image shows the most commonly used fuel in each region of the country.

	Region	Type of fuel used
0	EASTERN	COAL
1	NORTH EASTERN	HYDRO
2	NORTHERN	HYDRO
3	SOUTHERN	HYDRO
4	WESTERN	COAL

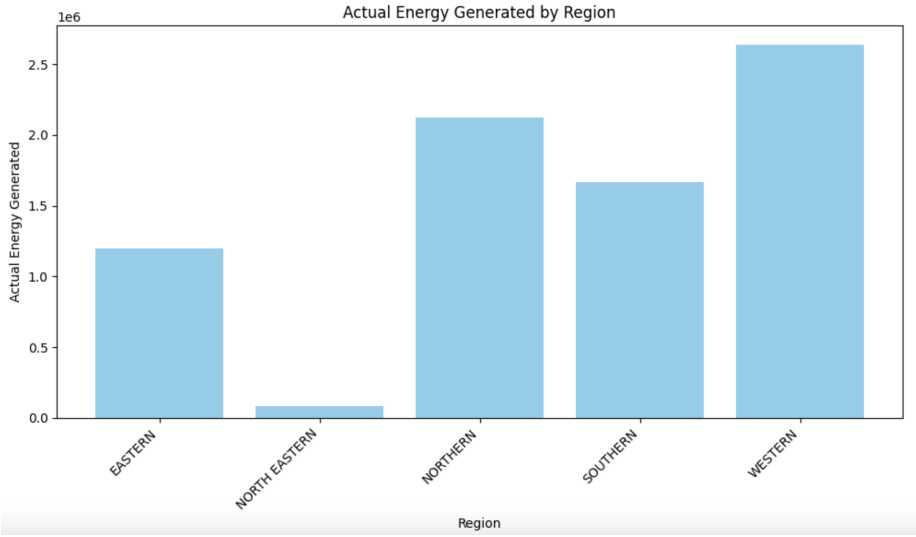
8. **Draw a pie chart showing the Market Share of each sector of power plants**



9. **Draw a plot showing the Actual Energy Generation Trend over the years**



10. **Draw a Bar Chart showing Actual Energy Generation per Region**



Question 2

The given dataset is not sufficient to identify pollution trends as it does not provide information about the emissions of pollutants from the power plants. It also doesn't have the data about the pollution levels in the regions where the power plants are located.

To analyse the pollution trends we need:

- Particulate matter emissions(like CO2, NOx, SOx) data from the plants. This can be obtained from **Central Pollution Control Board (CPCB)**.
- Concentrations of pollutants (like PM2.5, PM10, NO2, SO2, O3, CO) of the regions where the plants are located. This can be obtained from **National Air Quality Monitoring Programme (NAMP)**.
- We may also study the meteorolgical data which affects pollution dispersion. This can be obtained from **India Meteorological Department (IMD)**.
- The locations of the power plants and air quality monitoring stations may also help us in studying the pollution patterns. This can be obtained from **Geographic Information System (GIS)** data from the government.
- We can also use the reports released by the power plants themselves about their pollution management and emissions.

Now we can combine the obtained pollution data with the given dataset and use statistical analysis tools like regression and correlation to identify relationships between power plant emissions and regional pollution levels. Using time series plots we can observe the trends in emissions and pollution levels.

Question 3

In this problem we are given a labelled dataset containing Consumer Information and their classifications into 4 groups - A,B,C,D. Based on this we need to label an unlabelled dataset. This is a Supervised Learning - Multiclass Classification problem.

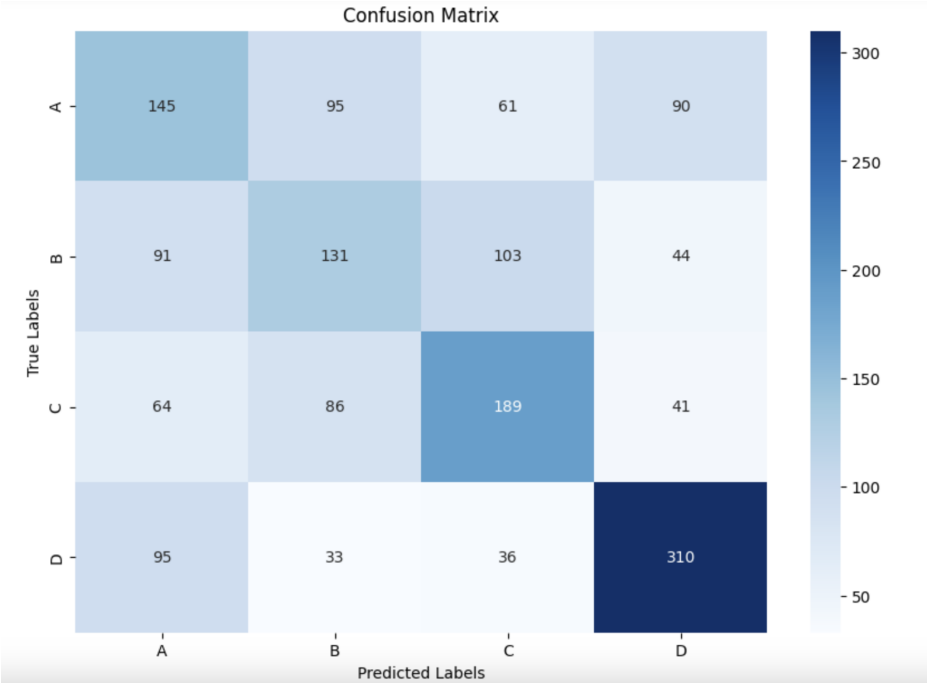
We use pandas library to store the dataset in a manageable way and the sci-kit learn library to apply Data Analysis methods on the dataset. The code has been implemented in the Jupyter Notebook.

Explaining the Code

We split the labelled dataset into Training and Testing sets. The numerical and categorical features need to be dealt separately during preprocessing. We use transformers to normalize the values of the features. To handle the missing values we use the simple imputer which puts the mean value for the numerical features and most frequent value for the categorical features. We use a Random Forest Classifier which works well for Multiclass Classification problems and prevents overfitting.
Now we train the model and test it on the test data. For the given dataset I got an accuracy of 48%.

	precision	recall	f1-score	support
A	0.37	0.37	0.37	391
B	0.38	0.36	0.37	369
C	0.49	0.50	0.49	380
D	0.64	0.65	0.65	474
accuracy			0.48	1614
macro avg	0.47	0.47	0.47	1614
weighted avg	0.48	0.48	0.48	1614

To better visualize the test results we can plot the confusion matrix:



Now we use this trained model on the given unlabelled dataset. The predictions have been given in the drive link in the file named: Predictions_for_Unlabelled_Dataset.csv

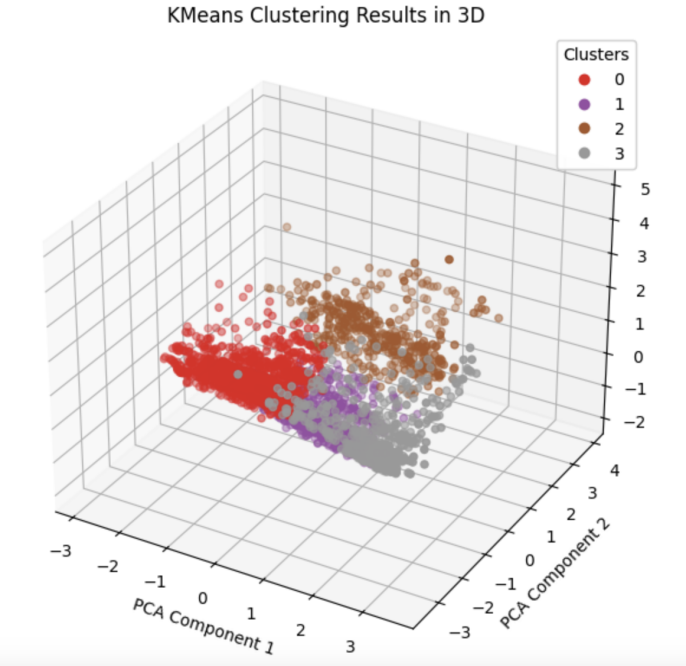
Question 4

If we were not given the labelled consumer dataset, then this would be a problem of Unsupervised Learning and we would have to observe the patterns in the given datapoints to find relations and then divide them into groups. One of the most effective algorithms for this is Clustering, specifically K Means Clustering in which we find the clusters by minimizing the variance within each cluster. The details about the implementation of this algorithm are in the next question.

Question 5

As described in the previous question we use K Means Clustering algorithm for this problem. Preprocessing of data is similar to Q3, for implementing the K-Means algorithm we use the Sci-Kit learn library.

To visualise the results of the clustering algorithm we can draw a 3d plot representing the clusters by compressing the dataset into 3 dimensions using Principal Component Analysis(PCA). The plot obtained is shown below:



Upon relating the results of Q3 with the clusters obtained here I found the relation between the cluster numbers and the Group name to be: 0 is A, 1 is B, 2 is C and 3 is D.

Using a map from cluster numbers to group name we obtain the final predictions for the unlabelled dataset. The predictions have been stored in a file named: Predictions_using_Clustering.csv

Question 6

Prediction of the sales of our products would hugely depend on the weather and seasonal conditions. Hence it is quite difficult to predict the sales and we need to take these factors into account. Some possible causes for the large error in the predictions are:

- The model might not be considering the external factors like weather patterns and change in market demand appropriately.
- The model might not be considering the temporal dependencies and seasonality in the data.
- The model may not be using the most relevant features, or there might be redundant or irrelevant features included.
- The model might not be the best for the problem we are trying to solve.
- The model might be overfitted or underfitted over the data.

To improve the accuracy of the model we can take various steps:

- Considering weather forecasts and historical weather data to incorporate its impact on renewable energy production influences sales a lot.
- Incorporating data on market trends, consumer behavior, and economic indicators that might influence demand. This will help in improving the predictions.
- We can use time series analysis to capture the temporal dependencies. We can use models specifically designed for time series data, such as ARIMA, SARIMA, or LSTM networks. This helps in accurate forecasting in time-dependent data.

- We need to continuously monitor the performance of the model and regularly update and retrain the model to maintain accuracy over time.
- We can use techniques such as Cross Validation which ensure that the model generalizes well to unseen data and we need to identify patterns in the errors and understand where the model is failing.
- We can try various models like Random Forest, Gradient Boosting, Neural Networks, and Support Vector Machines and use the one which gives the best accuracy.

These are a few ways in which we can improve the performance of our model and reduce the error in our predictions.

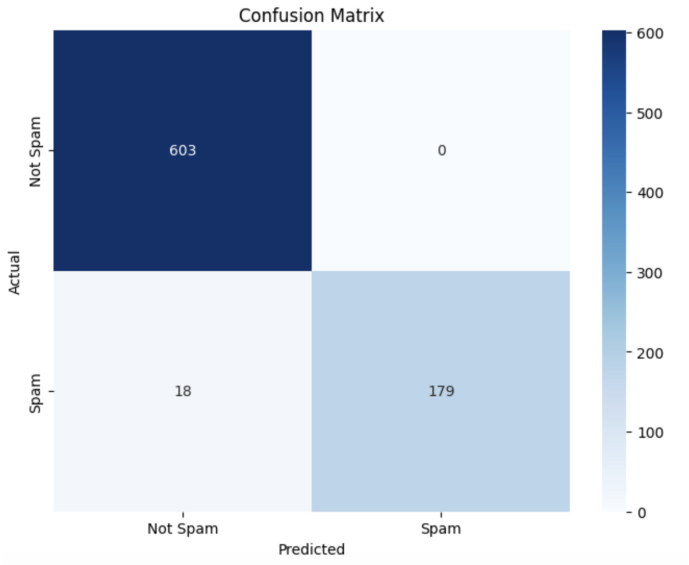
Question 7

This problem involves Spam Detection which is a binary classification problem. We need to use Natural Language Processing tools to deal with the messages in the dataset. Alongwith the sklearn library we use the Natural Language Tooklit(nltk) for stop words and the wordnet, BeautifulSoup for parsing the HTML documents and the re module for matching regular expressions.

We create the function preprocess_text for preprocessing the email messages from the dataset. It involves removing HTML tags, non alphabetical characters, converting to lowercase, tokenization, removing stop words and then applying lemmatization.

After preprocessing the text we now apply vectorisation using the TF-IDF (Term Frequency-Inverse Document Frequency) technique which converts text data into numerical features that can be used by the machine learning model.

Now we train the model and test on the test data. I got an accuracy of 97.75% on the test dataset. The results of the test can be visualised using the Confusion Matrix:



The model is then applied to the unlabelled dataset and the results of the classification have been stored in the file named:
Classified_Emails.csv