THYROID DISEASE DETECTION PREDICTION

Detail Project Report(DPR)

Objective:

- Development of a predictive model for thyroid disease detection. The model will determine whether a patient has hyperthyroid or hypothyroid.

Benefits:

- To Detect early.
- Accurate identification.
- Make proper decision and better treatment.

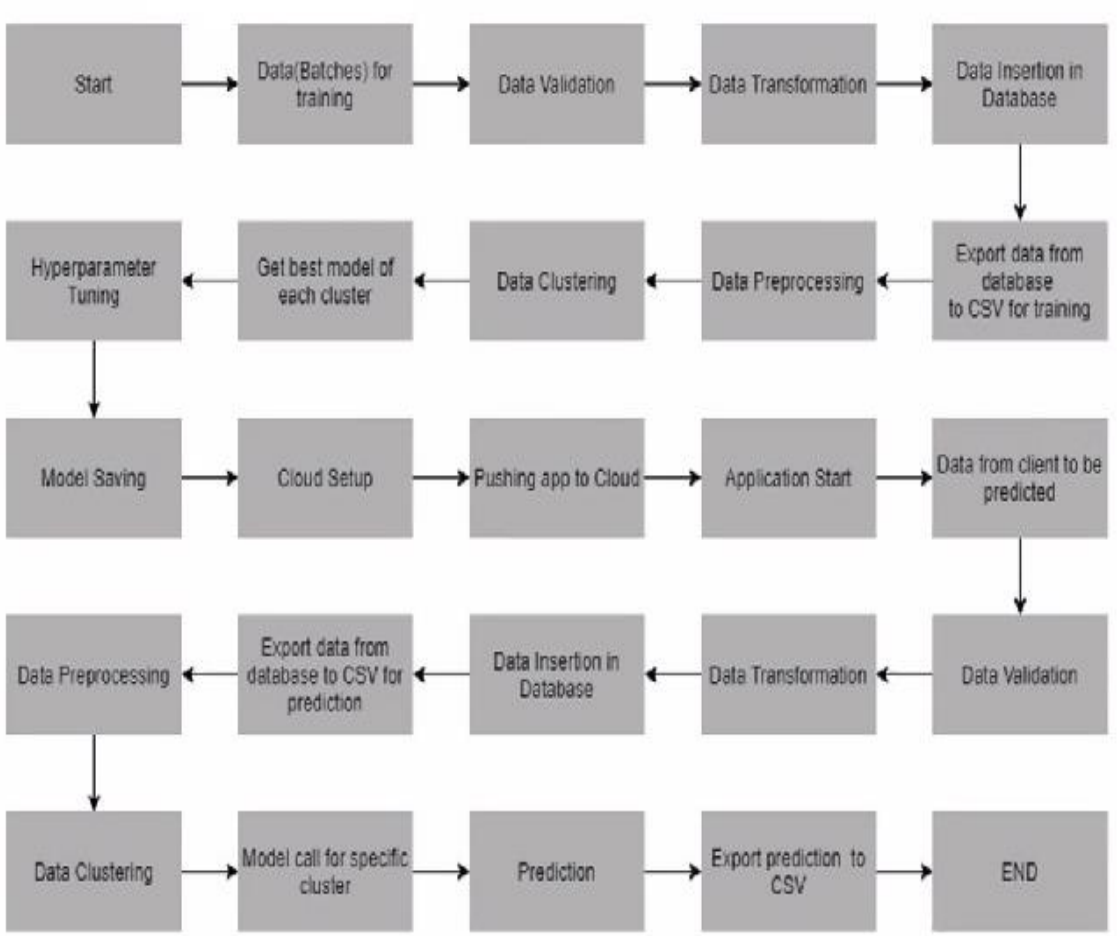
Data Sharing Agreement:

- Sample file name (ex hypothyroid_25092022_120000)
- Length of date stamp (8 digits)
- Length of time stamp (6 digits)
- Number of Columns (For Training 30 and For Prediction 29)
- Column name

[age, sex, on_thyroxine, query_on_thyroxine,on_antithyroid_medication, sick, pregnant, thyroid_surgery,I131_treatment,query_hypothyroid,query_hyperthyroid,lithium,goitre, tumor,hypopituitary,psych,TSH_measured,TSH,T3_measured,T3, TT4_measured, TT4, T4U_measured,T4U,FTI_measured,FTI,TBG_measured,TBG,referral_source, Class]

- Column data type (Numerical and Categorical)

ARCHITECTURE



Data Validation and Data Transformation :

- **Name Validation** - Validation of files name as per the DSA. We have created a regex pattern for validation. After it checks for date format and time format if these requirements are satisfied, we move such files to "Good_Data_Folder" else "Bad_Data_Folder.“
- **Number of Columns** – Validation of number of columns present in the files, and if it doesn't match then the file is moved to "Bad_Data_Folder.”
- **Name of Columns** - The name of the columns is validated and should be the same as given in the schema file. If not, then the file is moved to "Bad_Data_Folder".
- **Data type of columns** - The data type of columns is given in the schema file. It is validated when we insert the files into Database. If the datatype is wrong, then the file is moved to "Bad_Data_Folder“.
- **Null values in columns** - If any of the columns in a file have all the values as NULL or missing, we discard such a file and move it to “Bad_Data_Folder”.

Data Insertion in Database:

- **Table creation** :- Table name “Training“ is created in the database for inserting the files. If the table is already present then new files are inserted in the same table.

- **Insertion of files in the table :-** All the files in the "Good_Data_Folder" are inserted in the above- created table. If any file has invalid data type in any of the columns, the file is not loaded in the table.

Model Training:

- **Data Export from Db :** The accumulated data from db is exported in csv format for model training.

- **Data Preprocessing :** Performing EDA to get insight of data like identifying distribution , outliers ,trend among data etc. Check for null values in the columns. If present impute the null values. Encode the categorical values with numeric values. Perform Standard Scalar to scale down the values.

Clustering:

- KMeans algorithm is used to create clusters in the preprocessed data. The optimum number of clusters is selected by plotting the elbow plot, and for the dynamic selection of the number of clusters, we are using Knee Locator function. The idea behind clustering is to implement different algorithms on the structured data.
- The Kmeans model is trained over preprocessed data, and the model is saved for further use in prediction.

Model Selection:

- After the clusters are created, we find the best model for each cluster. By using 2 algorithms “KNN” and “RandomForest". For each cluster both the hyper tunned algorithms are used. We calculate the AUC scores for both models and select the model with the best score. Similarly, the model is selected for each cluster. All the models for every cluster are saved for use in prediction.

Prediction :

- The testing files are shared in the batches and we perform the same Validation operations ,data transformation and data insertion on them.
- The accumulated data from database is exported in csv format for prediction
- We perform data pre-processing techniques on it.
- KMeans model created during training is loaded and clusters for the preprocessed data is predicted
- Based on the cluster number respective model is loaded and is used to predict the data for that cluster.
- Once the prediction is done for all the clusters. The predictions are saved in csv format and shared.

Q & A

Q1) What's the source of data?

- The data for training is provided by the client in multiple batches and each batch contain multiple files.

Q 2) What was the type of data?

- The data was the combination of numerical and Categorical values.

Q 3) What's the complete flow you followed in this Project?

- Refer slide 4th for better Understanding

Q 4) After the File validation what you do with incompatible file or files which didn't pass the validation?

- Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the **bad data folder**.

Q 5) How logs are managed?

- We are using different logs as per the steps that we follow in validation and modeling like File validation log , Data Insertion ,Model Training log , prediction log etc.

Q 6) What techniques were you using for data pre-processing?

- Removing unwanted attributes.
- Visualizing relation of independent variables with each other and output variables.
- Checking and changing Distribution of continuous values.
- Removing outliers.
- Cleaning data and imputing if null values are present.
- Converting categorical data into numeric values.
- Scaling the data.

Q 7) How Prediction was done?

- The testing files are shared by the client .We Perform the same life cycle till the data is clustered .Then on the basis of cluster number model is loaded and perform prediction. In the end we get the accumulated data of predictions.

Q 8) How training was done or what models were used?

- Before diving the data in training and validation set we performed clustering over fit to divide the data into clusters.
- As per cluster the training and validation data were divided.
- The scaling was performed over training and validation data
- Algorithms like KNN , RandomForest were used based on the recall final model was used for each cluster and we saved that model .

Q 9) What are the different stages of deployment?

- When the model is ready we deploy it in Fire environment .Where SIT and UAT is performed over it.
- Once We get Sign off from Fire we deploy in Earth and UAT is performed over it.
- After getting the sign off from Earth we deploy in production