Concrete Strength Detection

Objective:

Development of a predictive model for predicting the Cement Strength based on the material used. The model will determine the strength in of the concrete block in MPa.

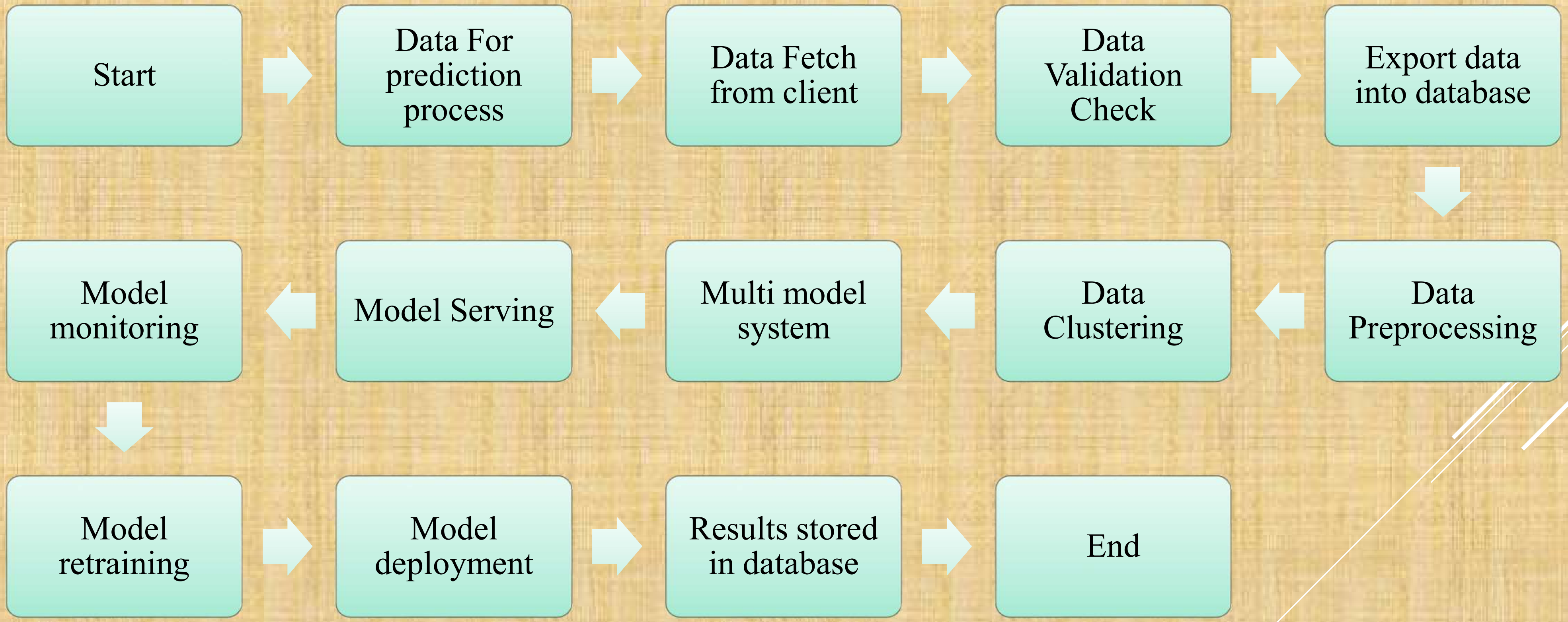
Benefits:

- Instant detection based on the raw materials.
- Saves clients time.
- Helps in easy flow for managing resources.

Data Sharing Agreement :

- Sample file name (ex Concrete_020922)
- Length of date stamp(6 digits)
- Number of Columns: 8
- Column names: Cement, BlastFurnaceSlag, FlyAsh, Water, Superplasticizer, CoarseAggregate, FineAggregate, Age
- Column data type:

Architecture



Data Validation and Data Transformation :

- Name Validation - Validation of files name as per the DSA. A checking condition has been developed to check the file name according to the DSA. After it checks for date 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 MongoDB:

- A new database named “Project” will be created that will contain all the tables related to the prediction.
- Inside the database a table named “Concrete” is created that will store the training data.
- Prediction data set is loaded into a table named “predictConcrete”.
- Final predictions are kept in the table “Results”.

Model Training:

- Data Export from Db :

  The accumulated data from db is exported in csv format for model training

- Data Preprocessing
  - Check for null values in the columns. If present impute the null 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 Silhouette Score. 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 “Random Forest” and "XGBoost". For each cluster both the hyper tunned algorithms are used. We calculate the r2 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 db 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 values.

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

Refer slide 5th 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.
- Scaling the data

Q 7) 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 Random Forest , XGBoost were used based on the recall final model was used for each cluster and we saved that model .

Q 8) 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 9) What are the different stages of deployment?
  - When the model is ready we deploy it in local environment .