

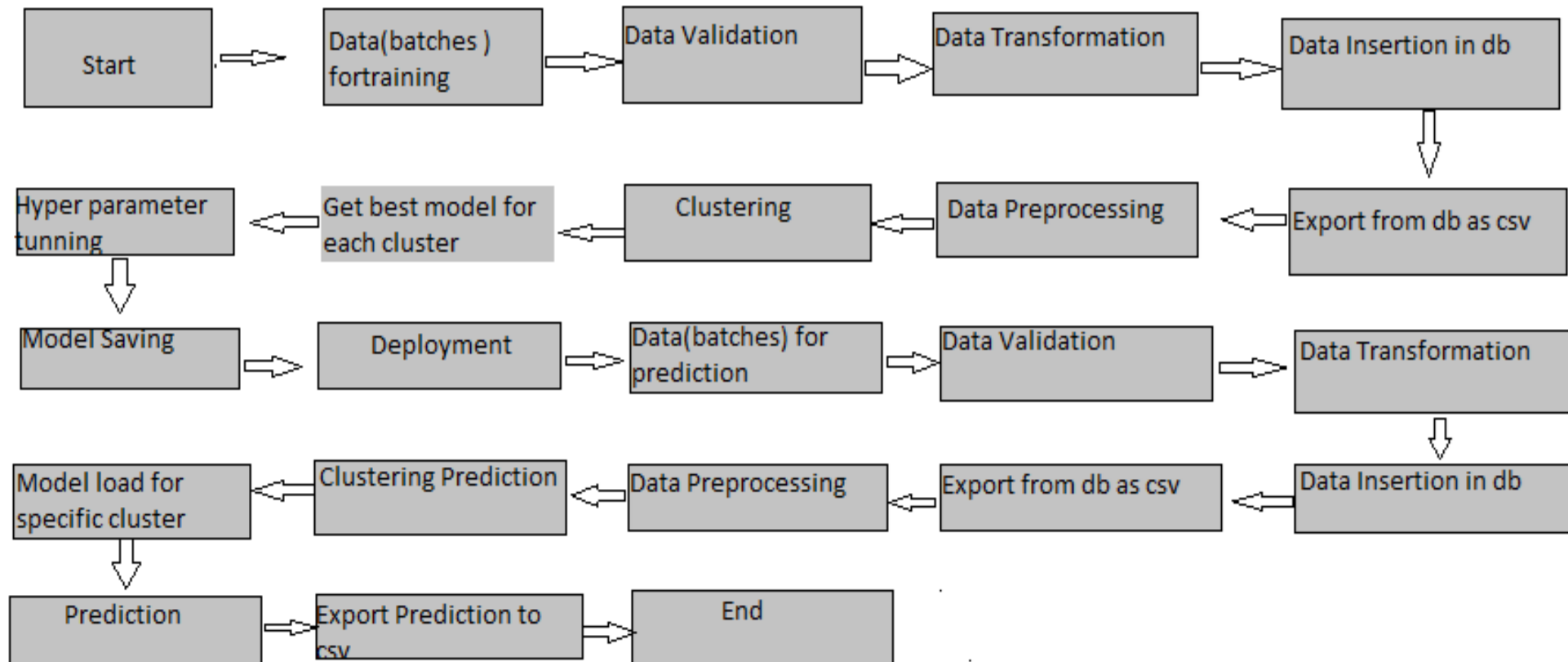
HR Analytics - Absenteeism Data Analysis Report

Several thin, white, parallel lines of varying lengths and slight curves are positioned on the right side of the slide, extending from the top right towards the bottom left.

Objective:

- ▶ **To create a report or use analytics concept to provide a smooth monitoring of workforce for the HR department.**
- **Potential analyses could be to see if there is an association between BMI and absence, as well as season, work load, distance from work and the other factors in the data**
- ▶ **Benefits:**
 - This data analysis can help you find predictors of absence. Gives better insight of customers base.
 - Helps in easy flow for managing resources.
 - Required to use analytics concept to provide a smooth monitoring of workforce for the HR department.

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 “Absent_Hours” 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

➤ 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 KneeLocator 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 “SVM” and "XGBoost". For each cluster both the hyper tuned 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 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 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 SVM , 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.

