

Insurance Premium Predictor

Objective:

The goal of this project is to give people an estimate of how much they need based on

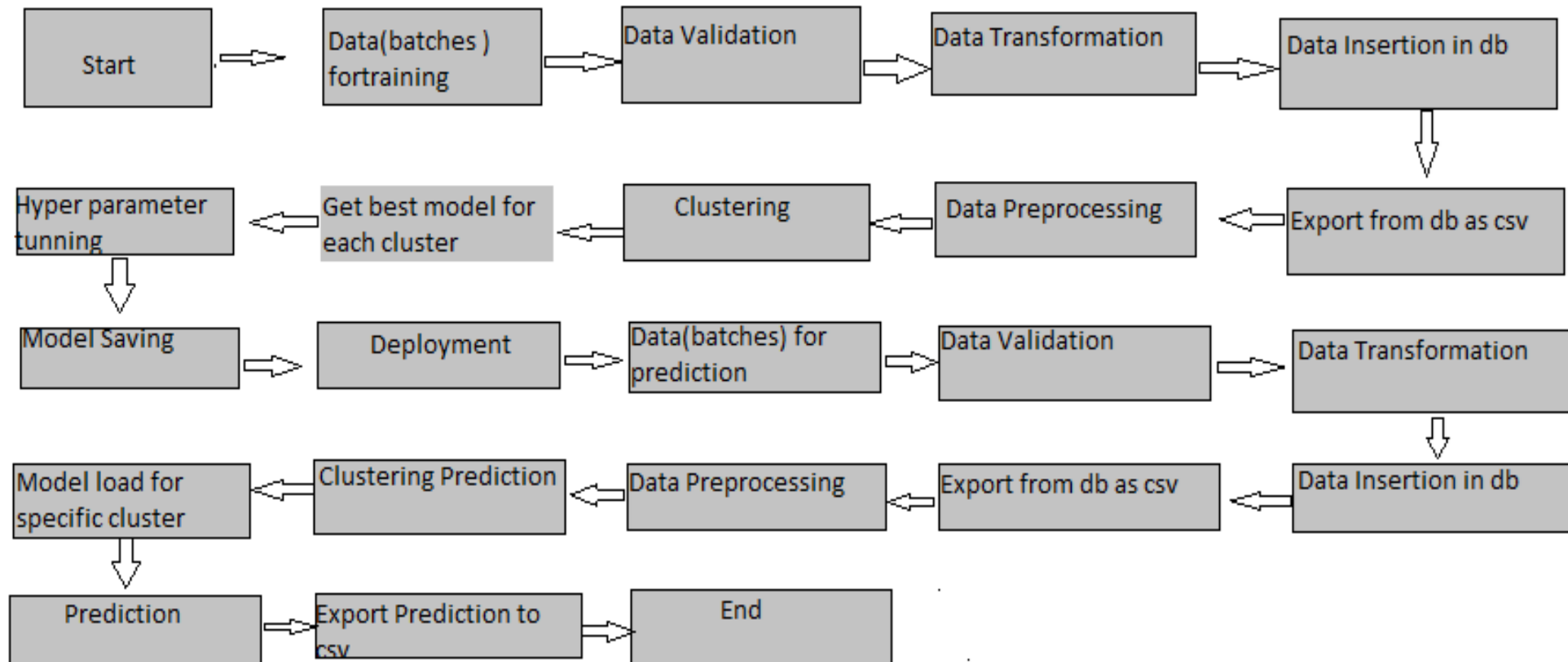
their individual health situation. After that, customers can work with any health

insurance carrier and its plans and perks while keeping the projected cost from our

study in mind. This can assist a person in concentrating on the health side of an

insurance policy rather than the ineffective part.

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 "insurance_predictor" 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 Selection –
 - We found the best model for the dataset. We checked four algorithms out of which Gradient Boosting was selected. We will calculate the r^2 scores and mean absolute error for models and select the model with the best score.
- 
- A series of white lines of varying lengths and orientations are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

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.
- After pre-processing the data, respective model is loaded and is used to predict the data.
- Once the prediction is done, 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 Linear Regression, SVM, Random Forest and Gradient Boosting were used based on the recall final model was used for each cluster and we saved that model .

