**Problem Statement**

To build a regression model to predict the calories burnt based on the given indicators in the training data.

**Architecture**



**Data Description**

Data Description: This dataset predicts the visibility distance based on the different indicators as below:

1. Id: The customer ID
2. ActivityDate: The date for which the activity is getting tracked.
3. TotalSteps: Total Steps taken on that day.
4. TotalDistance: Total distance covered.
5. TrackerDistance: Distance as per the tracker
6. LoggedActivitiesDistance: Logged
7. VeryActiveDistance: The distance for which the user was the most active.
8. ModeratelyActiveDistance: The distance for which the user was moderately active.
9. LightActiveDistance: The distance for which the user was the least active.
10. SedentaryActiveDistance: The distance for which the user was almost inactive.
11. VeryActiveMinutes: The number of minutes for the most activity.
12. FairlyActiveMinutes: The number of minutes for moderately activity.
13. LightlyActiveMinutes: The number of minutes for the least activity
14. SedentaryMinutes: The number of minutes for almost no activity
15. Calories(Target): The calories burnt.

Apart from training files, we also require a "schema" file from the client, which contains all the relevant information about the training files such as:

Name of the files, Length of Date value in FileName, Length of Time value in FileName, Number of Columns, Name of the Columns, and their datatype.

**Data Validation**

In this step, we perform different sets of validation on the given set of training files.

1. Name Validation- We validate the name of the files based on the given name in the schema file. We have created a regex pattern as per the name given in the schema file to use for validation. After validating the pattern in the name, we check for the length of date in the file name as well as the length of time in the file name. If all the values are as per requirement, we move such files to "Good\_Data" container else we move such files to "Bad\_Data" container.
2. Number of Columns - We validate the number of columns present in the files, and if it doesn't match with the value given in the schema file, then the file is moved to "Bad\_Data" container.
3. 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" container.
4. The datatype of columns - The datatype of columns is given in the schema file. This is validated when we insert the files into Database. If the datatype is wrong, then the file is moved to "Bad\_Data" container.
5. 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" container.

**Data Insertion in Database**

1) Database Creation and connection - Create a database and collection with the given name passed. If the database is already created, open the connection to the database.

2) Insertion of files in the table - All the files in the "Good\_Data" container are inserted in the above-created collection. If any file has invalid data type in any of the columns, the file is not loaded in the table and is moved to "Bad\_Data\_Folder".

**Model Training**

1) Data Export from Db - The data in a stored database is exported as a CSV file to be used for model training.

2) Data Preprocessing

a) Drop columns not useful for training the model. Such columns were selected while doing the EDA.

b) Replace the invalid values with numpy “nan” so we can use imputer on such values.

d) Check for null values in the columns. If present, impute the null values

e) Scale the training and test data separately

3) 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

To train data in different clusters. The Kmeans model is trained over preprocessed data and the model is saved for further use in prediction.

4) Model Selection - After clusters are created, we find the best model for each cluster. We are using Multiple algorithms, "KNN", "SVR", "Voting\_Regressor", "Bagging\_Regressor", "Stacking\_Regreesor","Linear\_Regression",and "RandomForest\_Regressor". For each cluster, all the algorithms are passed with the best parameters derived from GridSearch. We calculate the Rsquared scores for all 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 Data Description**

Client will send the data in multiple set of files in batches at a given location. Data will contain climate indicators in 10 columns.

Apart from prediction files, we also require a "schema" file from client which contains all the relevant information about the training files such as:

Name of the files, Length of Date value in FileName, Length of Time value in FileName, Number of Columns, Name of the Columns and their datatype.

**Data Validation**

In this step, we perform different sets of validation on the given set of training files.

1) Name Validation- We validate the name of the files on the basis of given Name in the schema file. We have created a regex pattern as per the name given in schema file, to use for validation. After validating the pattern in the name, we check for length of date in the file name as well as length of time in the file name. If all the values are as per requirement, we move such files to "Good\_Data\_Folder" else we move such files to "Bad\_Data\_Folder".

2) Number of Columns - We validate the number of columns present in the files, if it doesn't match with the value given in the schema file then the file is moved to "Bad\_Data\_Folder".

3) Name of Columns - The name of the columns is validated and should be same as given in the schema file. If not, then the file is moved to "Bad\_Data\_Folder".

4) Datatype of columns - The datatype of columns is given in the schema file. This is validated when we insert the files into Database. If dataype is wrong then the file is moved to "Bad\_Data\_Folder".

5) Null values in columns - If any of the columns in a file has all the values as NULL or missing, we discard such file and move it to "Bad\_Data\_Folder".

**Data Insertion in Database**

1) Database Creation and connection - Create database with the given name passed. If the database is already created, open the connection to the database.

2) Table creation in the database - Table with name - "Good\_Data", is created in the database for inserting the files in the "Good\_Data\_Folder" on the basis of given column names and datatype in the schema file. If table is already present then new table is not created, and new files are inserted the already present table as we want training to be done on new as well old training files.

3) 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 and is moved to "Bad\_Data\_Folder".

**Prediction**

1) Data Export from Db - The data in the stored database is exported as a CSV file to be used for prediction.

2) Data Preprocessing

a) Drop columns not useful for training the model. Such columns were selected while doing the EDA.

b) Replace the invalid values with numpy “nan” so we can use imputer on such values.

c) Check for null values in the columns. If present, impute the null values.

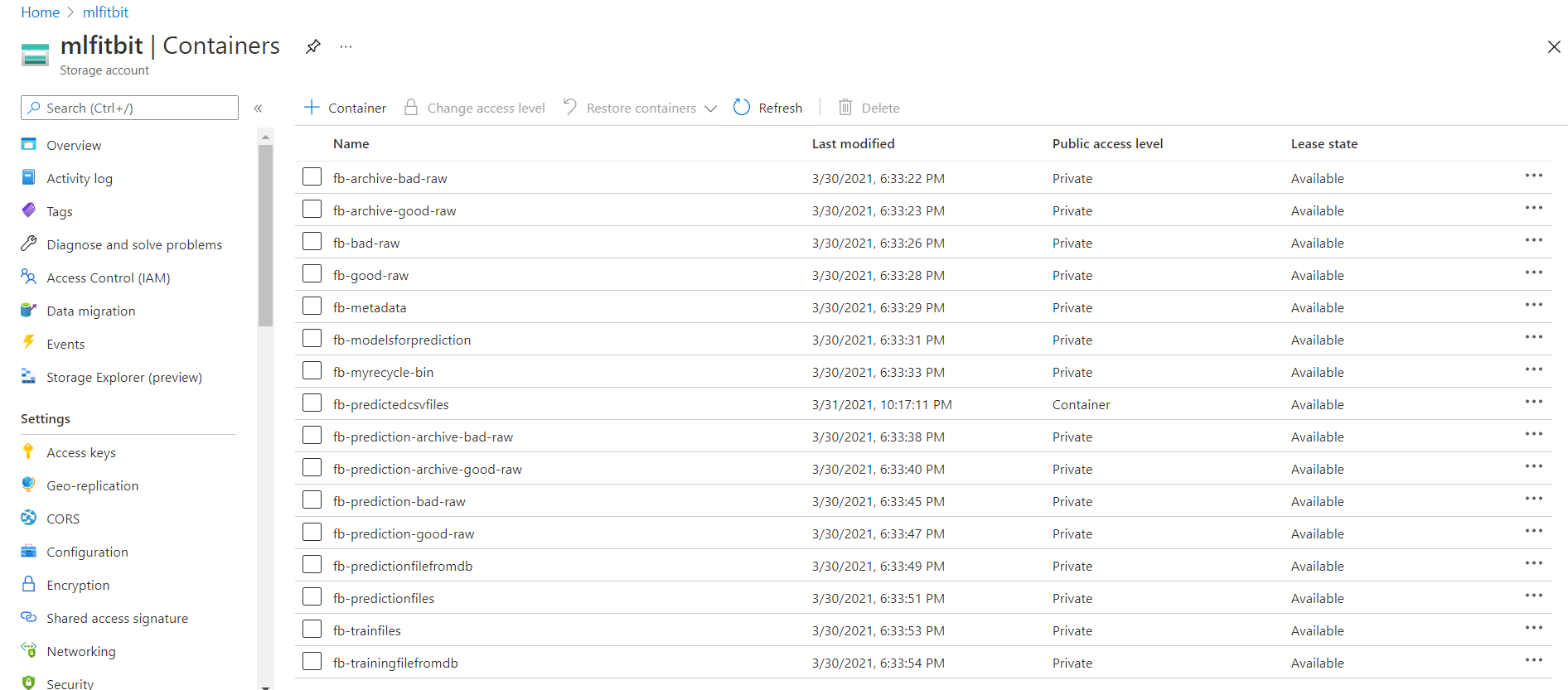
d) Scale the training data.

3) Clustering - KMeans model created during training is loaded, and clusters for the preprocessed prediction data is predicted.

4) Prediction - Based on the cluster number, the respective model is loaded and is used to predict the data for that cluster.

5) Once the prediction is made for all the clusters, the predictions along with the original names before label encoder are saved in a CSV file at a given location and the location is returned to the client.

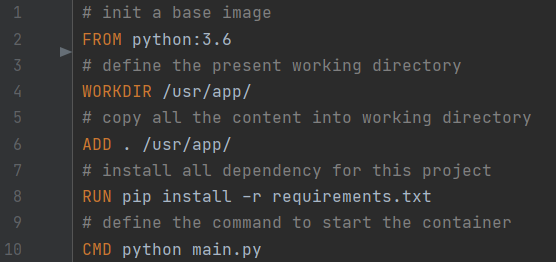
**Now let’s see the fitbit project azure container structure.**



**Deployment**

**requirements.txt** file consists of all the packages that you need to deploy the app in the cloud.

Before deployment we contenarized our project using docker. We need to create DockerFile to bulid docker image. Let’s see what command we need to specify in our DockerFile.



**To create docker image frist we navigate our project directory in command prompt and then we need to run following commands:**

<REPOSITORY> we need to provide name of docker image.

* docker image build -t <REPOSITORY> .
* docker images
* docker ps
* docker run -p 5000:5000 -d <REPOSITORY>
* docker login dockerfitbit.azurecr.io
* docker push dockerfitbit.azurecr.io/mlfitbit:latest

To create a web app on Azure, follow these steps:

* Login on [https://portal.azure.com](https://portal.azure.com/).
* Click on Create a Resource.
* Search for Web App and click on Create.
* Link your ACR image that you pushed in to your app.

**The app is now up and running on Azure Web Services.**