Impact of type of food consumed on spread of COVID-19

Architectural Decisions Document

## Data Source

### Technology Choice

The dataset is an ".csv" file obtained from the "https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset”. It contains the information about percentage of energy obtained from different type of food item in a country and number of people affected by COVID-19 in that country.

### Justification

The dataset provided is very comprehensive and clear and it is provided in .csv format which makes it easy to import and process.

## Data Repository

### Technology Choice

The IBM Cloud Storage is used for storage of initial database as well as intermediate data generated during processing.

### Justification

The IBM Cloud storage gives free data storage up to 5 GB and it is very easy to import in the python notebooks using boto3 library.

## Initial Data Exploration

### Technology Choice

The Python notebooks in Watson Studio with pandas and matplotlib are used for initial data exploration.

### Justification

The most important part during initial data exploration is to understand the distribution of different features in the data. This can be very easily done using pandas and matplotlib libraries in Python. The features whose distribution indicates lack of usable knowledge can be removed at this stage.

## Extract, transform, load (ETL)

### Technology Choice

The tools used in this stage are similar to the once in last stage.

### Justification

In this stage we decide the feature to be used as output. We also reduce the dimensionality of the input data using different techniques like correlation analysis with which we can remove the features which are very less related to the output. We also clean the data for any type differences or not available values.

## Feature creation

### Technology Choice

Here we can use Python notebook in IBM Watson studio with pandas, numpy libraries.

### Justification

This step includes converting features depending upon the type of input needed for the model. Ex: converting categorical data to one-hot encoding.

Our dataset has all the data in one format and hence this kind of conversion is not needed.

We used this stage to randomly select training and test data with 85% training data using numpy.

## Model definition

### Technology Choice

We use total 4 models for this problem evaluation:

1. SVR model using all the features.
2. SVR model with feature reduction with PCA.
3. Deep learning model with 5 dense hidden layers.
4. Similar Deep learning model with Dropout and regularization.

We used IBM Watson studio Notebook with skit-learn and keras libraries.

### Justification

In the problem of percentage estimation, the SVR is a good algorithm providing good accuracy. The skit-learn library provides easy implementation of SVR and also the PCA for feature reduction. The keras is a great open source high level deep learning framework on top of tensorflow engine which provides easy and fast implementations for deep learning models.

## Model training

### Technology Choice

The training of all 4 models is done in IBM Watson studio notebook using 8GB RAM environment.

### Justification

The Watson studio provides the use of these environments for free and as the dataset size and model sizes are small there is no need to use spark sessions or GPU’s for training as the case with most of Deep learning models.

## Model evaluation

### Technology Choice

The evaluation of the trained models is done based on the “mean-squared-error”.

### Justification

The output is in the form of percentage of population and hence the mean-squared-error gives a good estimate of accuracy of the trained model on the test set of data.

## Model deployment

### Technology Choice

The deployment of the model best in the evaluation is done in an interactive Python notebook in which user have to input the percentage of total energy obtained from different food types and gets the estimate of the percentage of confirmed patients of corona virus in that country.

### Justification

The Python Notebook is an easy and light way of giving an output to user based on the input and can be shared easily using its URL.

The Link to GIST of notebooks of implemented project is as follows:

https://gist.github.com/prateekshamprasadpawar/2fb8a68b1a9c751b4641d46a1c82466b