**Python for data science Level 2**

Note: please do at least 3 of the 5 case studies provided below

1. **Title: Data exploration of NBA players and predicting their salary**

**Background Information:** It is fascinating to understand the different backgrounds of the NBA super stars. There is also a curiosity to find out about the team, height and salary of an NBA player. We have a very detailed dataset on NBA super star players from 1950 available with Kaggle (given in the link below). We shall use it for our analysis

**Purpose:** The purpose of the exercise is to do a detailed analysis of the super star NBA players. This detailed analysis will focus on what kind of players do well, if taller players do well and whether the team they play has a role in their salary

**Problem description:** We have many input variables such as team name, position of the player, his weight, height and salary. Use pandas.groupby, pandas.crosstab and pandas.pivottable functions to perform exploratory data analysis and visualize the results using and box plots and bar charts using seaborn and matplotlib. Build a classification model using Naïve Bayes or SVM to predict the salary of an NBA player

**Dataset:** nba players dataset

**Link:** <https://raw.githubusercontent.com/sivabalanb/Data-Analysis-with-Pandas-and-Python/master/nba.csv>

**Technique / Approach:** Data analysis using pandas and Classification

**Algorithm:** Naïve Bayes and Random Forest

2. **Title**: **Analysis of immigration data from Iceland to Canada from 1980 to 2013**

**Background Information:** Every year millions of people immigrate to Canada from countries across the world. The data related to such migrants is available in the link given below. There is enormous benefit in understanding the range of the immigration taking place in to a country. Then we can plan for the resources ahead of time.

**Purpose:** The purpose of the exercise is to do an analysis of the number of people who move to Canada from different countries and visualize the results. Specifically we are choosing the people who migrate from Iceland in to Canada and compare with other countries’ data.

**Problem description:** Compare the number of Icelandic immigrants (country = 'Iceland') to Canada from year 1980 to 2013 with the migration data from other countries. Use pandas.groupby, pandas.crosstab and pandas.pivottable functions to perform exploratory data analysis and visualize the results using area plots and box plots using seaborn and matplotlib

**Dataset:** Canadian immigration dataset

**Link:** <https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/DV0101EN/labs/Data_Files/Canada.xlsx>

**Technique / Approach:** Data analysis using pandas

**Tools:** Numpy, Pandas, Matplotlib and Seaborn

3. **Title: Fake news detection using Pandas and Statsmodels**

**Background Information:** In all times of human existence real news has to be separated from fake news or rumour mongering. The need for such classification has never been more acute than in the recent times. Data Science models can be favourably used to do this task.

**Purpose:** The purpose of this exercise is to segregate fake news from real news by using a suitable data science model such as Naïve Bayes or ensemble method

**Problem description:** We use both fake news and real news as training inputs to our TFIDF (Term Frequency and Inverse Document Frequency) features for creating a classifier. Then a regular test set is given as input and the classifier to detect if it is fake news or real news

**Dataset:** fake.csv

**Link:** <https://www.kaggle.com/mrisdal/fake-news>

(click on download in the top right on page)

**Technique / Approach:** Classification using Pandas and Statsmodels

**Algorithm:** Naïve Bayes / Ensemble

4. **Title: Customer segmentation on the retail customer dataset**

**Background Information:** There are thousands of people who visit malls in the US in the peak hour time from 7pm to 9pm. It is a standard practice to group the customers in the retail industry or the mobile segment or auto insurance industry. There are very interesting patterns that come out of such grouping done on retail or insurance customers. If there are many features in the dataset we can do dimensionality reduction such as PCA to reduce the features involved.

**Purpose:** The purpose of the exercise is to segment the retail customers who visit the malls in the united states. We wish to group the customers so that we can have targeted marketing strategies, promotions and discount offers and coupons and hence improve the customer experience across different groups.

**Problem description:** We have the mall customers dataset, which has age, gender, annual income and a spending score (from 1 to 100) as columns for each customer. Develop a model and group them to find interesting buying patterns in the customer segments. Before doing modelling perform dimensionality reduction such as PCA on the dataset so that we spend less time and CPU resources during execution

**Dataset:** mall-customers-dataset.csv

**Link:** <https://www.kaggle.com/shwetabh123/mall-customers>

**Technique / Approach:** We will use clustering approach

**Algorithm:** KMeans clustering with PCA

5. **Title: Building a model to predict Boston housing prices**

**Background Information:** Making a decision to own an apartment or an independent house in any country is a major decision. It is becomes more complicated because of the number of parameters we have to understand before we take a decision. There has always been a need to build a system which predicts the price of the house for us given all the inputs. Regression models in Data Science can be favourably exploited for this purpose.

**Purpose:** The purpose of this exercise is to give a prediction for a house in Boston. We use the Boston housing price prediction dataset given in the link below to build and train the model and do the prediction. This model can also be used for other such housing data in different cities of United States

**Problem description:** If we want to predict the price of the house which is the output parameter, then there are various input parameters such as square feet of the house, number of bed rooms, crime rate in the area, accessibility to national highways, full value tax rate in the area and many more.   
  
Develop a model to predict the housing price based on the many independent variables provided as input. Perform cross validation by choosing the training / testing split value as 0.7 and find the accuracy of the model. Use multivariate polynomial linear regression

**Dataset:** boston\_housing\_price\_prediction.csv

**Data description link :** <https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html>

**Data file link :** <https://www.kaggle.com/vikrishnan/boston-house-prices>

**Technique / Approach:** Regression will be used

**Algorithm:** The algorithm used is Multi variate polynomial regression