1. **Use the dataset in** [**this**](https://umkc.box.com/s/de9bdscv3ys9ff1sjogbfzppzy1o477a) **link. Predict the temperature using the weather details specified in the columns**

**Description of the dataset:**

**This dataset has 12 columns and we want to predict “Temperature (C)” using other independent variables.**

**a) Apply some Exploratory Data Analysis to draw some insight from the data**

**a) Visualize the data and draw the model line**

**b) Evaluate the model and try to interpret the performance that you get**

**Solution:**

1. In this task initially loaded all required libraries and read dataset. Checked for any null values in the datset as we found nulls in ‘Precip Type’ column replaces null value with rain value and loaded dependent and independent values into X and y. Then splitted complete dataset into training and testing data and performed analysis on the dataset. Used linear regression, Polynomial regression and Randomforest regression to check for error rates. Randomforest yields much less error rate so applied that model. Below are screenshots of working code along with output images.

![A screenshot of a cell phone

Description automatically generated]()

![A screenshot of a social media post

Description automatically generated]()

![A screenshot of a social media post

Description automatically generated]()

![A screenshot of a social media post

Description automatically generated]()

![A screenshot of a cell phone

Description automatically generated]()

![A screenshot of a social media post

Description automatically generated]()

c)

Below is the accuracy score for both test and training data as it yields better accuracy score by applying randomforest.

![A screenshot of a cell phone

Description automatically generated]()

1. Use the dataset in [this](https://umkc.box.com/s/mlmyznn0667tbgui5urnzbos1i1c48qf) link and apply classification on that.

**Description of the dataset:**

The dataset is Spam Dataset and has two columns

**1) class**

**2) text**

You need to initially clean the text data and apply the techniques we have learned for transforming the text data into numeric format (TFIDF, Count\_Vectorizer, …)

Evaluate the model and try to interpret the result

**Solution:**

Importing basic required libraries from nltk and sklearn and we can see our data has ham which means not spam and spam data . Reading data set initially.

Importing ‘re’ for use of regular expression and importing nltk for use of stopwords, stemming etc. Importing porter stemmer which helps us to find base root form of word and initializing the variable of porter stemmer and creating empty list where after all data processing we can append to this. Running a for loop through all messages and removing all extra letters or symbols except a-z and A-Z and replacing with space after removal and at last lowering them to small letters and splitting each word so it would be helpful for stemming. Checking for each word whether it is present in stop words if it doesn’t exist then it performs stemming on the word and joining them to sentence and appending to list.

![A screenshot of a cell phone

Description automatically generated]()

Next creating bag of words or document matrix using CountVectorizer and object is called with max\_features to select frequent words and performing fit\_transform on this and converting it to array. As Class column is two categorial data using dummies converting to numerical data upon existence with value 1 and non-existence with value 0. Using tain\_test\_split splitting data to training data and testing data then using NaïveBayes classification technique as it yields better results for nltk problems and fitting with train data and predicting using test data. At printing confusion matrix and accuracy scores.

![A screenshot of a cell phone

Description automatically generated]()

Below is screenshot of code and output samples with accuracy score of model and confusion matrix.

![A screenshot of a cell phone

Description automatically generated]()