1. **Logistic Regression**
   1. **Binary Classification**
      1. **Prediction of Cancer from Smoking**

Given the data on a binary response variable telling us whether the cancer is present or not and a single binary independent variable telling whether the person smokes or not we want to predict the possibility of cancer due to smoking. In nutshell we want to know “how more likely is a person to have cancer if he/she smokes rather he/she doesn’t”. We are supposed to do the following:

A study was performed on lung cancer possibility due to smoking habits. Data on presence/absence of two attributes viz. lung cancer and smoking was collected for 25 individuals.

**Practice Question 1:** Consider the dataset **Smoking and Cancer.xlsx** and perform the following objectives.

1. Build a logistic regression model for cancer possibility using smoking as an independent variable.
2. Test for the Significance of independent variable.
3. Construct the Confusion (Classification) Table and report the percentage of correct classification in the given emails. Also calculate specificity and sensitivity of the model.
4. For each person obtain the probability of him/her having cancer and hence the prediction of cancer using the Logistic Classifier you have built.
5. Estimate the odds ratio and interpret it.

**Case 1:** A researcher wants to understand how smoking habits affect the possibility of cancer. For that purpose she randomly selected (**Smoking and Cancer.xlsx**) 25 individuals out of which some are smokers and some are non-smokers, and some have cancer and some do not have cancer. Help her in estimating the probability of having cancer if a person smokes using some appropriate statistical model. What are the estimated probabilities of having cancer for the given 25 individuals?

She is also interested in knowing if smoking is at all significant in causing cancer and also wants to know if the model that you are producing is ‘good’ or not. Also she is specifically interested in knowing what is the extra chance of a person having cancer if he/she smokes over the person who doesn’t, if at all there exists a significant relationship between cancer and smoking?

* + 1. **Predicting Skull Type using some Physical Attributes of Skull**

We are interested in predicting the type of skull of humans as one of two possible types I and II based on some five physical measures available related to the skulls.

**Practice Question 2:** Consider the dataset **Skull Type Prediction.xlsx** and perform the following objectives.

1. Build a logistic regression model for classifying a human skull as Type I/Type II using the given independent variables.
2. Test for the Significance individual independent variables.
3. Test for the overall Logistic Regression using Hosmer and Lemeshow Test (It’s a Chi-Square Test).
4. Construct the Confusion (Classification) Table and report the percentage of correct classification in the given skulls. Also calculate specificity and sensitivity of the model.
5. For each skull obtain the probability of it being Type I or Type II, and hence predict the skull Type using the Logistic Classifier you have built.
6. For a set of five physical measures given for a new skull in the dataset **Skull Type Prediction – Validation Data.xlsx** predict the skull type using the Logistic Classifier you have built.

**Case 2:** A zoologist wants to understand and predict the human skull type based on some five physical measures of a skull. For the purpose of research she collects data (**Skull Type Prediction.xls**x) on skull type and all five measures of the skull for 19 individuals. Help her in predicting the skull type using some appropriate statistical model. What are the predicted skull types for the given 25 individuals?

She is also interested in knowing if some or all of five measures are at all significant in predicting the skull type and also wants to know if the model that you are producing is ‘good’ or not. Based the model you have developed she wants to predict the skull type for five new individuals for whom the five physical measures are provided in the file **Skull Type Prediction - Validation Data.xls***.*

* + 1. **Sentiment Analysis – What makes a US Presidential Candidate Win?**

What we are interested here in knowing that depending upon what and how a politician give speeches, his/her chances of winning the elections are affected. The idea here is similar to the email spam detection. The speech and more explicitly the content of the speech and it is delivery will have the information about the fact that the audience is convinced enough to vote for or against him/her. Sentiment Analysis is a discipline in itself, we are trying to understand the basics of to solve a particular problem. Commonly if politician is polite but passionate enough to serve the people, talks about development, remain optimist in his speech, talks about facts and figures related to government policies to explain his point to the audience is expected to win and vice-versa. But we want to examine what does data say?

The first aspect of the problem is to understand the data itself. I hope there is no confusion that we are going to use past data meaning by past win/loss statistics and the corresponding speeches. Clearly the response variable will indicate the win/loss information.

But what will be my independent variables? The independent variables will be the characteristics of the speech which may affect the win/loss which are commonly the following:

1. Proportion of words in the speech showing *Optimism*
2. Proportion of words in the speech showing *Pessimism*
3. Proportion of words in the speech showing the use of *Past*
4. Proportion of words in the speech showing the use of *Present*
5. Proportion of words in the speech showing the use of *Future*
6. Number of time he/she mentions his/her own party
7. Number of time he/she mentions his/her opposite parties

There are some more independent variables possible for which we need to understand the concept of big five personality traits which represent the personality traits of human which are the following:

1. Openness: *Curious, original, intellectual, creative and open to new ideas*.
2. Conscientiousness: *Organized, systematic, punctual, achievement oriented and dependable*.
3. Extraversion: *Outgoing, talkative, social and enjoys being in social situations*.
4. Agreeableness: *Affable, tolerant, sensitive, trusting, kind and warm*.
5. Neuroticism: *Anxious, irritable, temperamental and moody*.

Other that these big five personality traits the emotional content of the speech may also affect the win/loss. Thus we consider the following more independent variables.

1. Some measure indicating the content of speech showing *Openness*
2. Some measure indicating the content of speech showing *Conscientiousness*
3. Some measure indicating the content of speech showing *Extraversion*
4. Some measure indicating the content of speech showing *Agreeableness*
5. Some measure indicating the content of speech showing *Neuroticism*
6. Some measure indicating the content of speech showing *emotionality*

Once we get this data, task is all with the statistical analyst to make an efficient model with good predictive power.

**Practice Question 3:** Consider the **US Presidential Data.xlsx** and perform the following objectives:

1. Build a logistic regression model for classifying win/loss using the given independent variables.
2. Test for the Significance individual independent variables.
3. Test for the overall Logistic Regression using Hosmer and Lemeshow Test (It’s a Chi-Square Test).
4. Construct the Confusion (Classification) Table and report the percentage of correct classification in the given skulls. Also calculate specificity and sensitivity of the model.
5. For each speech obtain the probability of winning, and hence predict the win/loss status using the Logistic Classifier you have built.

**Case 3:** A US News Channel wants to understand “What makes a US Presidential Candidate Win?” For that purpose it collects the county-wise Win/Loss results for the last general elections (**US Presidential Data.xls**) and the speeches of the candidates. Help the new channel in understanding what all features can be created out of the speech which possibly have relation with the Win/Loss. Using appropriate statistical model the news channel wants to understand what is the prediction of Win/Loss based on different features created from speech for the given candidates?

It is also interested in knowing if some or all of speech features are at all significant in predicting the Win/Loss and also wants to know if the model that you are producing is ‘good’ or not.

* + 1. **Admission Prediction based on GRE Score, GPA and Class Rank**

**Case 4:** An educational services agency wants to understand how chances of admissions at UCLA are affected by the GRE Score, GPA and Class Rank of an applicant. The data on the same is available at <http://www.ats.ucla.edu/stat/data/binary.csv>. Help the agency in understanding the impact of the three features on the chances of admission. Using appropriate statistical model the agency wants to understand what is the admission prediction based on the three features for the given applicants?

Also find out if an applicant with a GRE Score of 720, GPA of 3.9 and Class Rank as 1 will get admission at UCLA?

* 1. **Multiclass Classification**
     1. **Multiclass Classification by Decomposing a Multiclass Problem into several Binary Classification Problems (One-vs-All Classification) – Flower Species Prediction**

**Practice Question 5:** Consider the dataset **Flower Species.xlsx** dataset which has data on Sepal Length, Sepal Width, Petal Length and Petal Width and Species Type for 150 different flowers and perform the following objectives.

1. Decompose the multiclass (3-class) classification problem into three Binary Classification Problems and perform the following for each problem:
   1. Test for the Significance individual independent variables.
   2. Test for the overall Logistic Regression using Hosmer and Lemeshow Test (It’s a Chi-Square Test).
   3. Construct the Confusion (Classification) Table and report the percentage of correct classification for the given emails. Also calculate specificity and sensitivity of the model.
   4. For each flower obtain the predicted probability and hence the predicted class using the Logistic Classifier you have built.
2. Obtain the multiclass predicted flower species for the original problem using the three sub problems.
3. Construct the classification matrix and report the percentage of correct classification.

**Case 5:** A botanist wants to predict the species of a flower based on some four measurements of a flower viz. sepal length, sepal width, petal length and petal width. For the purpose of research she takes the Fisher’s iris data (**Flower Species.xlsx**)which has data on the aforementioned measurements for 150 flowers of three different species viz. setosa, virginica and versicolor. Help her in solving the prediction problem using appropriate statistical model and obtain the predicted flower species for all the given 150 flowers.

She is also interested in knowing if some or all of four measurements are at all significant in predicting the flower species and also wants to know if the model that you are producing is ‘good’ or not.

* + 1. **Multiclass Classification using Multinomial Logistic Regression – Flower Species Prediction**

**Practice Question 6:** Consider the dataset **Flower Species.xlsx** dataset which has data on Sepal Length, Sepal Width, Petal Length and Petal Width and Species Type for 150 different flowers and perform the following objectives.

* 1. Built a multinomial logistic regression model for the given problem.
  2. Obtain the predicted class for each flower using the multinomial logistic regression.
  3. Construct the classification Matrix and report the percentage of correct classification.

**Case 6:** A botanist wants to predict the species of a flower based on some four measurements of a flower viz. sepal length, sepal width, petal length and petal width. For the purpose of research she takes the Fisher’s iris data (**Flower Species.xlsx**)which has data on the aforementioned measurements for 150 flowers of three different species viz. setosa, virginica and versicolor. Help her in solving the prediction problem using appropriate statistical model (different from case 9) and obtain the predicted flower species for all the given 150 flowers.

She is also interested in knowing if some or all of four measurements are at all significant in predicting the flower species and also wants to know if the model that you are producing is ‘good’ or not.

* + 1. **Comparative Study of the two approaches of Multiclass Classification for the Flower Species Prediction Problem**

**Case 7:** The botanist after solving the Flower Species Prediction Problem using two different statistical techniques in case 5 and case 6, now wants to compare the two methodologies for Flower Species Prediction Problem on following grounds:

* 1. Compare the predicted class for each flower based on both the approaches.
  2. Compare the diagonal entries of the classification matrix based on both the approaches.
  3. Compare the percentage of correct classification for both the approaches.