1A. we get the statistics of the data frame by using describe function. Hence loaded the data from flu data.csv to the data frame and used describe function to get the statistics count, mean, standard deviation, minimum, 25% percentile, 50% percentile, 75% percentile, and maximum.

Table

Description automatically generated

1B. Yes. The Female column consists of values like 1.0 and 0.0 which are float. These can be converted to categorical or string by using a function called astype() and then replacing the values by yes and no for 1.0 and 0.0 respectively. Similarly, we can also do it for sick and flu variables. The sick value consists of values like 0.0,1.0,2.0. These are also float values. By using astype() we can convert these to string and then replace these values by not, cold-like symptoms, flu-like symptoms for 0.0,1.0,2.0 respectively. The Flu value consists of values like 0.0,1.0. These are also float values. By using astype() we can convert these to string and then replace these values by not, flu-like symptoms for 0.0,1.0 respectively.

Chart, scatter chart

Description automatically generated

1C. The describe function of a data frame gives the statistics count, unique value, top value, and  
frequency of top value for the categorical values.

Table

Description automatically generated

1D. To find the redundant column we can find the correlation matrix using the corr function. Then find the columns with the highest correlation that is values close to 1. After analyzing the correlation matrix, I see there is no correlation value close to 1. The correlation value between flu and sick seems to be 0.78

Graphical user interface, application, table, Excel

Description automatically generated

1E. The correlation value between SociDist and PersnDist is found to be 0.11. Hence they are not correlated which means that people who maintain social distancing does not tend to maintain personal distance

Graphical user interface

Description automatically generated with medium confidence

2A. Implemented logistic\_regression and sigmoid functions by taking Risk as input and Flu as target or output variable which would return weights and the intercept. And implemented predict function to find the predicted values for the given input and find the accuracy of the model.



2B. written code to find the accuracy of the model. Which is used to understand the performance of the model. The accuracy of the model is found to be 0.81.

Graphical user interface, text, application

Description automatically generated

3A. Implemented logistic\_regression function with Flu(y) as target variable and Vaccin, HndWshQual, HndWshFreq, SociDist, NoFaceContact, RespEttiqu, PersnDist, HandSanit, Risk, Inefficacy, KnowlTrans, KnowlMgmt as input features for the model. Calculated the accuracy for the model and found to be 0.81.

3B. Implemented forward\_selection function and found out that the following features 'Risk','HandSanit','HndWshFreq','SociDist','KnowlTrans','Vaccin','RespEttiqu','Inefficacy' as best as they improved the accuracy from the above model to 0.83.

3C. Comparing 3A and 3B models I found the accuracy of 3B model is high i.e., 0.83. Hence I feel the model 3B is good compared to 3A. The 3B features performed better than the 3A feature set.

4A.i. As per the above observation in 3C. I choose the model from 3B i.e., model with selected features as it has more accuracy than model in 3A. After performing regularization for the model selected in 3C I don’t see any difference in the accuracy obtained in 3B and 4A.i. so, I feel the given data does not require any regularization as it is not improving any accuracy.

4A. ii. we can do two types of feature scaling a. standardization and b. Normalization. I choose to do standardization. When I implemented standardization to the model in 4A.i. I observed that the accuracy is decreasing slightly i.e., approximately 0.1. hence, I feel for the given data there is no need of feature scaling with regularization as it is decreasing the accuracy which leads to decrease in performance of the model.

4B.i. Implemented logistic\_regression\_with\_new\_cf function by changing the cost function given in the question. Found that the given cost function is the cost function of linear regression model.

4B. ii. After calculating the accuracy of the model with the new cost function it gave the same result as that of the model in 4A.i. It gave the same result even with the different cost function.