STAT 515 FINAL PROJECT

Breast Cancer Prediction

Group 8: Sanika Suhas Dalvi, Snehita Moturu

Why this dataset?

Healthcare is a domain that has a lot of areas that can be studied for analysis. We came across this dataset for Breast Cancer, where we were able to see a good set of predictors and a response variable that could predict the type of cancer (i.e. Benign or Malignant). We decided to make use of this dataset with a vision to help the community in predicting the type of cancer.There is still no cure for malignant cancer. Helping people to understand the type of cancer they have once detected, helps them to take next crucial steps in the process. For instance, if the cancer is benign, next steps would include surgical operation and removal of tumor. However, if the cancer is malignant, it would help the patient to get started with their medication.

Data Set Description:

The breast cancer dataset consists of patient information that can be used to predict whether a person has benign or malignant breast cancer. This data was acquired from the UCI Machine learning repository website where the data has been collected from the Wisconsin Hospitals. The dataset has 699 rows in all, each row is the record of one patient. It consists of 11 attributes in total. Out of which, first 10 attributes are going to be our predictors and the last one “Class” is our class that we are trying to build a model for and predict the data. All the data present in the dataset is numerical data.

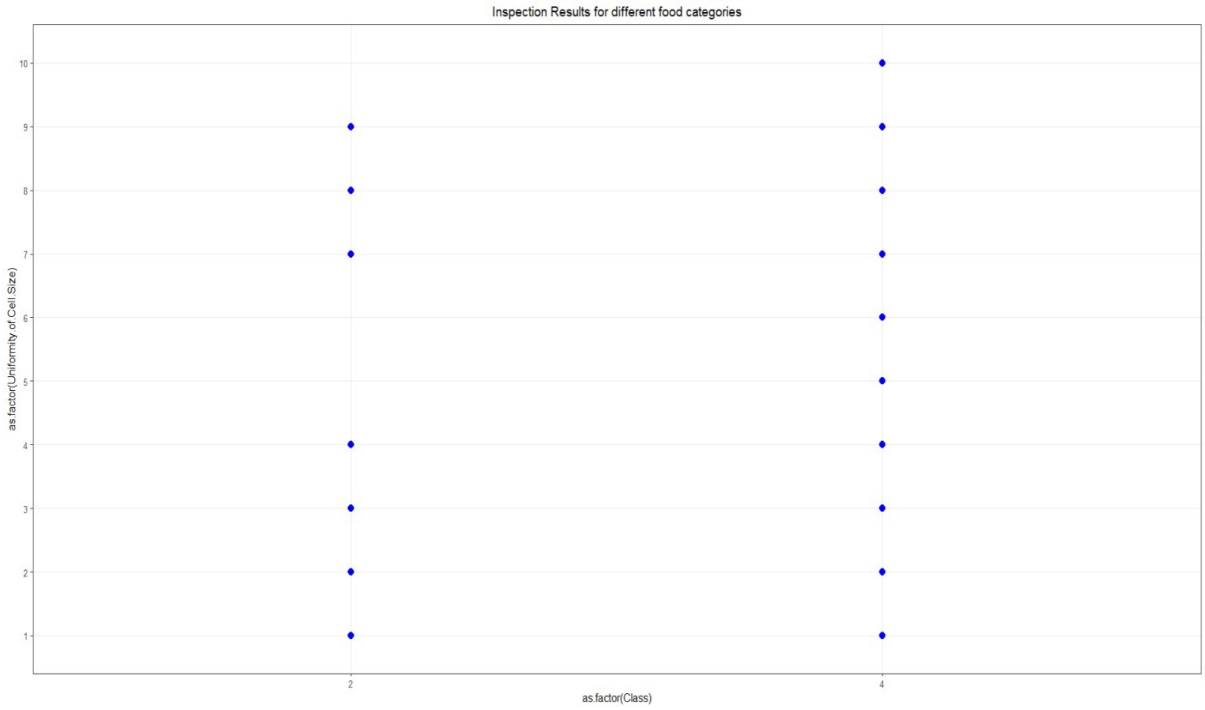
A	B	C	D	E	F	G	H	I	J	K
Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2

Methodology:

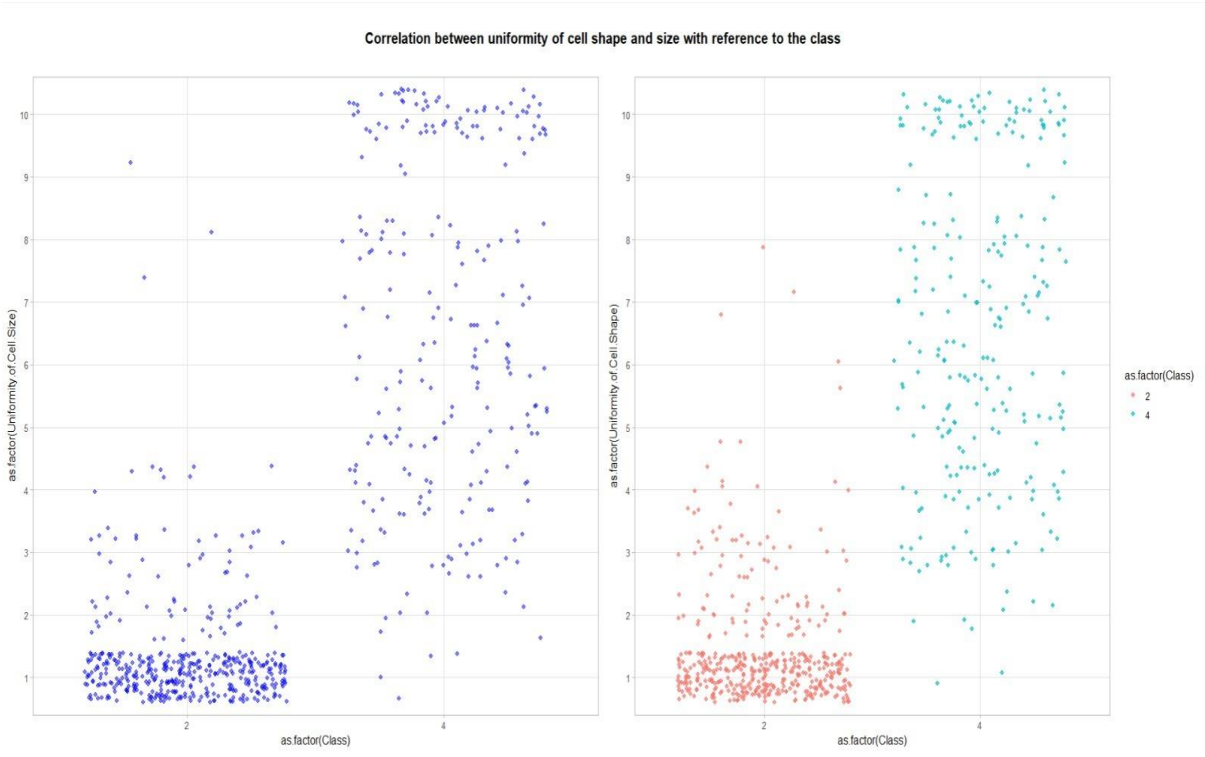
The methodology acquired for building models and visualizations involves several steps. First, we will install and load the required libraries, including mlbench and randomForest. Next step is to pre-process the dataset to make sure we are working on clean data. We will load the breast cancer dataset into a variable and remove records with special character “?” in the "Bare Nuclei" column. We also check if there is any column with a different datatype that needs to be converted fit. Once the data is cleaned, next step will be to divide the dataset into training and testing sets, and fit logistic regression and random forest models to the training set. We will use these models to make predictions on the testing set and calculate their accuracy. Finally, we will visualize the correlations between different columns in the dataset to understand the relationships between different attributes. These steps will enable us to build predictive models for breast cancer, evaluate their performance, and understand the importance of different attributes in the dataset.

Data Visualizations:

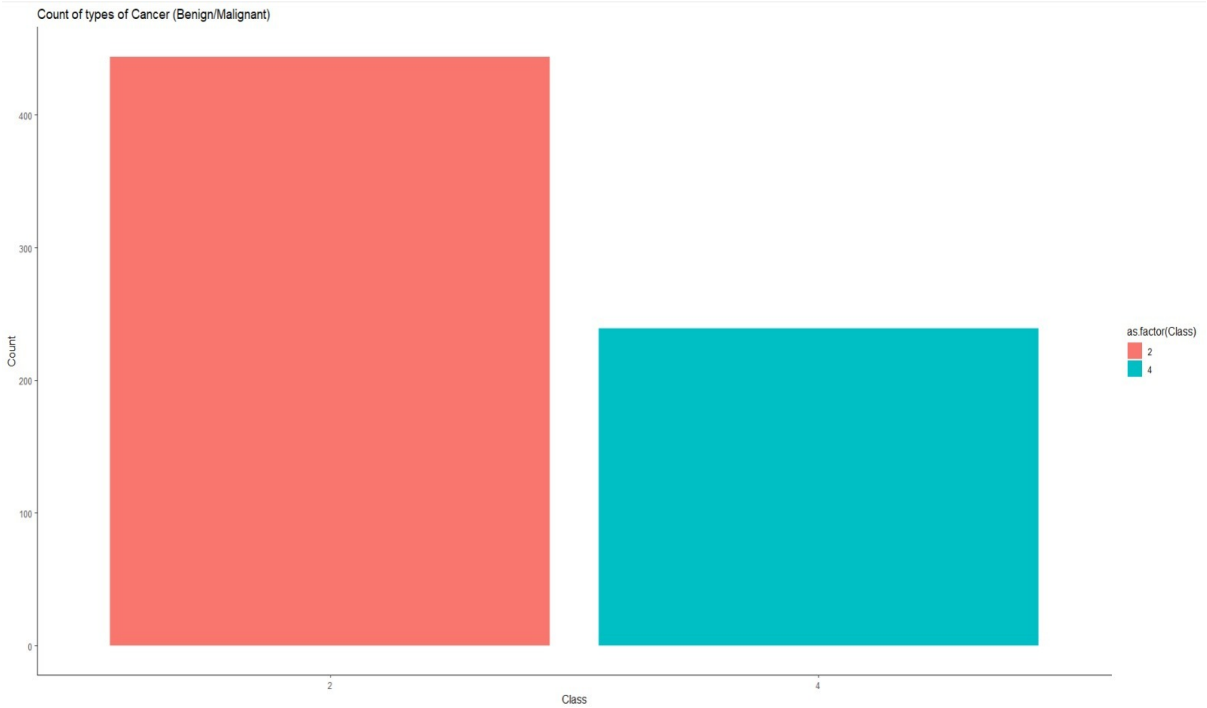
As part of the data visualization process, we generated a correlation graph using ggpairs function from the GGally library as shown below. After studying the graph, we can see that the correlation between class and uniformity of cell shape and size is higher, we can say that the attributes of Uniformity of cell size and Uniformity of cell shape are most relatable with respect to the class and hence we will be building our visualizations using these three correlated variables.



The first visualization that we are going to plot is a scatterplot between variables “Class” and “Uniformity of cell size”. As we can see, for cell sizes – 5,6 and 10, we are not able to see a benign type of cancer. However, for cell sizes- 1 through 10, we are able to see malignant cancer. This means, uniformity of cell size has some correlation with the class and can help for predictions.



The next visualization is between class and uniformity of cell shape and uniformity of cell size, where we are visualization both the graphs beside each other to be able to differentiate any similarities or dissimilarities between them. As we can see, for both the graphs, for benign type of cancer, the cell shape and cell size are clustered below 5. Whereas, for malignant type of cancer, the cell shape and size are scattered all over. From this, we can conclude that, when the cell shape and size is below 5, the probability of cancer being benign and malignant is equal. Whereas, when the cell size and shape are above 5, the probability of cancer being malignant is higher.



Finally, we are trying to understand which type of cancer occurred the most with respect to our dataset. For this reason, we built a bar plot, where we are able to see that the benign cancers were the ones that occurred more in number than the malignant ones. From this we can conclude that, with respect to our dataset, since benign cancer is the one that occurred the most, there are higher chances of going through a surgical operation and removing the tumor, thereby saving lives and giving people a second chance at life.

Model Building:

We built two models for doing predictions using Logistic regression and Random forest. The goal was to see which one performed better in terms of being able to predict the class of Breast cancer with reasonable accuracy. Before building models, we divided the dataset into training and testing data. We allotted 80% of the data for the training set and 20% of the data for the testing set. Also, we have removed 'Sample.code.number' attribute from the set of predictors, as it is not contributing towards making predictions.

1. Logistic Regression Model:

For this regression model, we are going to use generalized linear model (glm) to fit our model. We are using "Class' as our response variable and all the other attributes except "Sample Code number" as our set of predictors. Here, we are using our training data to train our model and we are setting family as "binomial", as we want to fit a logistic regression model. After building the model, we take summary of it by using summary() function.

```
glm.fit = glm(as.factor(Class)~Clump.Thickness+Uniformity.of.Cell.Size+Uniformity.of.Cell.Shape+Marginal.Adhesion+Single.Epithelial.Cell.Size+Bare.Nuclei+Bland.Chromatin+
            Normal.Nucleoli+Mitoses,data=train,family=binomial)
glm.fit
summary(glm.fit)
```

```
> summary(glm.fit)

Call:
glm(formula = as.factor(Class) ~ Clump.Thickness + Uniformity.of.Cell.Size +
    Uniformity.of.Cell.Shape + Marginal.Adhesion + Single.Epithelial.Cell.Size +
    Bare.Nuclei + Bland.Chromatin + Normal.Nucleoli + Mitoses,
    family = binomial, data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.23593  -0.06385  -0.03245   0.00779   2.50567

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -11.9071     1.9475  -6.114 9.71e-10 ***
Clump.Thickness            0.5370     0.2062   2.605  0.00919 **
Uniformity.of.Cell.Size   -0.2217     0.2466  -0.899  0.36877
Uniformity.of.Cell.Shape   0.5201     0.2880   1.806  0.07094 .
Marginal.Adhesion          0.4442     0.1648   2.696  0.00702 **
Single.Epithelial.Cell.Size  0.1241   0.2060   0.602  0.54696
Bare.Nuclei10              4.0654     1.2594   3.228  0.00125 **
Bland.Chromatin            0.5727     0.2436   2.351  0.01875 *
Normal.Nucleoli            0.3330     0.1647   2.021  0.04323 *
Mitoses                    0.9032     0.3423   2.638  0.00833 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 716.293  on 545  degrees of freedom
Residual deviance:  59.703  on 528  degrees of freedom
AIC: 95.703

Number of Fisher Scoring iterations: 18

>
```

As we can see from the summary, we are able to see the minimum, maximum, median, 1st quartile and 2nd quartile, along with coefficients, p-value, etc. The AIC value that we got is 95.7. We even tried different set of predictors. For all of them, the AIC value was greater than 95. So, this is the best set of predictors in comparison to the rest.

```
glm.pred=predict(glm.fit,testData,type="response")
glm.pred[1:10]
```

```
> glm.pred[1:10]
          10           12           13           14           23           25           27           31           39
0.0013472761 0.0005749408 0.3158318055 0.0025272071 0.0009832535 0.0005958141 0.0006960415 0.0008686215 0.9976427586
          52
0.4237078125
```

```
# Let us now predict and classify classes based on whether the cancer is benign(2) or malignant(4)
# For glm.pred value >0.5, breast cancer is malignant(4) and benign(2) is less than 0.5
predicted.classes <- ifelse(glm.pred > 0.5, 4,2)
head(predicted.classes)
```

```
> predicted.classes <- ifelse(glm.pred > 0.5, 4,2)
> head(predicted.classes)
10 12 13 14 23 25
 2  2  2  2  2  2
>
```

After fitting the model, we are finding out the probabilities that the model predicts. After which, we are setting threshold for making predictions. For instance, for probability > 0.5, the cancer should be malignant and for probability < 0.5, the cancer should be benign. After setting these thresholds, we are doing predictions using predict() function on our testing set and we are getting values. If we cross check these predictions with our dataset in csv file, we will be able to see that the model is doing correct predictions.

```
# Checking the model's accuracy
mean(predicted.classes == testData$Class)
```

```
> mean(predicted.classes == testData$Class)
[1] 0.9562044
>
```

Now, we are checking the model's accuracy. The model's accuracy is about 95%, which a good accuracy. 5% is the misclassification error, where the model is unable to predict the class correctly. However, 95% is a good accuracy percentage in front of the misclassification error.

2. Random Forest:

We have built model using Random Forest using the same set of predictors as we used for the logistic regression model. After fitting the model, as we run it, we are able to see the confusion matrix, where the model has done 337 correct predictions for benign cancer and 193 correct predictions for malignant cancer. In all, the model was unable to predict 16 records correctly, which is negligible when we compare it with correct predictions the model has made.

```
# We are ignoring Sample.code.number from predictors as it is not contributing towards prediction.
bcancer_RF <- randomForest(as.factor(Class)~Clump.Thickness+Uniformity.of.Cell.Size+Uniformity.of.Cell.Shape+Marginal.Adhesion+Single.Epithelial.Cell.Size+
                           Bare.Nuclei+Bland.Chromatin+Normal.Nucleoli+Mitoses,data=train, mtry=4,ntree=5000)
bcancer_RF
```

```
> bcancer_RF

Call:
 randomForest(formula = as.factor(Class) ~ Clump.Thickness + Uniformity.of.Cell.Size +
l.Nucleoli + Mitoses,      data = train, mtry = 4, ntree = 5000)
               Type of random forest: classification
                     Number of trees: 5000
No. of variables tried at each split: 4

        OOB estimate of  error rate: 2.93%
Confusion matrix:
    2   4 class.error
2 337  10  0.02881844
4   6 193  0.03015075
>
```

Now, we wanted to see which attribute acted as an important attribute while doing the prediction. We use the important() function here and we were able to see that Uniformity of cell shape, with the highest value of 75.8 amongst others, acted as an important attribute.

```
# Checking which variable is the most important variable for making predictions about breast cancer status
importance(bcancer_RF)
```

```
> importance(bcancer_RF)
                           MeanDecreaseGini
Clump.Thickness                   13.706163
Uniformity.of.Cell.Size           73.777706
Uniformity.of.Cell.Shape          75.818618
Marginal.Adhesion                  5.346117
Single.Epithelial.Cell.Size       23.085644
Bare.Nuclei                       24.069083
Bland.Chromatin                   20.903960
Normal.Nucleoli                   13.974805
Mitoses                            1.643178
>
```

After fitting the model, we are checking if the model is able to predict the data correctly. So we use predict() function over our testing set. We get a set of 10 predicted values. If we cross check these predictions with our dataset present in the csv file, we will be able to see that the model has done correct predictions.

```
# Predicting the Class status on the test set based on the fitted random forest model.

rForest_pred=predict(bcancer_RF,testData)
rForest_pred[1:10]
```

```
> rForest_pred=predict(bcancer_RF,testData)
> rForest_pred[1:10]
10 12 13 14 23 25 27 31 39 52
 2  2  4  2  2  2  2  2  4  4
Levels: 2 4
>
```

We have successfully built the model using Logistic Regression and Random Forest.

Challenges:

The main challenges we faced during the project are as follows:

1. Initially, there were headers missing in the original dataset file and we had to manually input them from the names text file which is available on the UCI website.

2. The “Bare Nuclei” column had 16 records that had special character of ‘?’, which we had to remove using the R script.

Conclusion:

From our modelling, analysis and visualization, we were able to see that the Uniformity of cell shape and Uniformity of cell size are the two variables that are highly related to the type of breast cancer (Benign/Malignant). Also, with respect to our dataset, the number of Benign cases is higher than that of the malignant ones, which means with respect to our dataset, since benign cancer is the one that occurred the most, there are higher chances of going through a surgical operation and removing the tumor, thereby saving lives and giving people a second chance at life.

References:

1. “Breast Cancer Wisconsin (Original) Data set”. [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/. [Accessed December 6,2022].