BUDT758T: Data Mining & Predictive Analytics

Project Title: **Breast Cancer Severity detection using Predictive Analytics**

Team Members: Liheng Wang

Sahil Pathan

Seo Yuhyeon

Yimin Cui

Nirmali Das

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	**Signature**
	Liheng Wang	Liheng Wang
	Sahil Pathan	Sahil Pathan
	Seo Yuhyeon	Seo Yuhyeon
	Yimin Cui	Yimin Cui
	Nirmali Das	Nirmali Das

TABLE OF CONTENTS

Executive Summary:

The primary goal of this project was to build a predictive model using an extensive breast cancer dataset to help medical professionals make more accurate diagnoses and in turn improve patient care. We are excited to share the details of our approach and the outcomes we achieved.

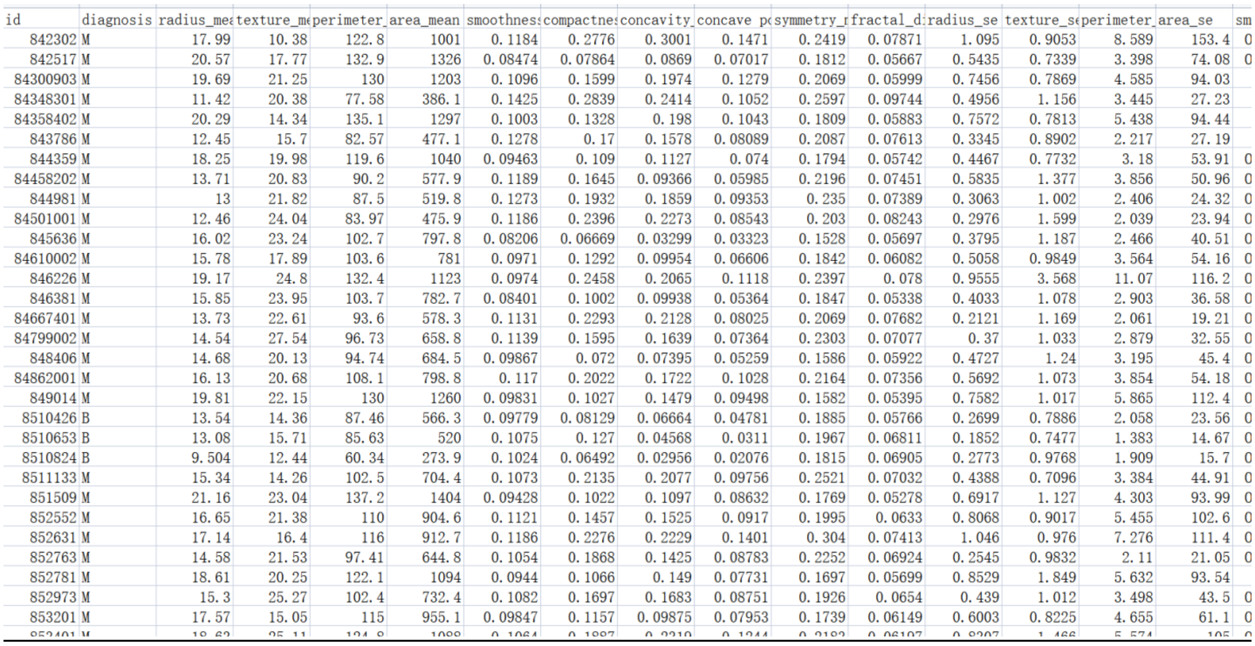
To ensure a strong foundation for our predictive models, we preprocessed the data by transforming the target variable, which represented the cancer diagnosis as either malignant or benign. We also scaled the numeric variables, which allowed the models to perform better.

The report compares the predictive performance of various machine learning algorithms for breast cancer diagnosis. The results indicate that the random forest algorithm outperforms other models with an accuracy of 0.98, which is higher than KNN, Naive Bayes, and Classification Tree. However, other models like ridge regression and Classification Tree still provide useful insights into the data. These findings suggest that the random forest algorithm can be a valuable tool for accurate breast cancer diagnosis. The report concludes that healthcare professionals and data scientists can use these results to improve patient outcomes and advance cancer research.

Data Summary:

The data utilized in this study is the Breast Cancer Wisconsin (Original) Data Set. This data set was sourced from the UCI Machine Learning Repository, a recognized online database of machine learning datasets utilized by researchers and scholars worldwide. The direct link to the data set is:

http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29



id	diagnosis	radius_mea	texture_me	perimeter_	area_mean	smoothness	compactnes	concavity_	concave pc	symmetry_r	fractal_di	radius_se	texture_se	perimeter_	area_se	sm
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0
842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	
844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0
844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0
84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0
845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0
84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0
846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0
846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0
84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0
84799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0
848406	M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0
84862001	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0
849014	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0
8510426	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0
8510653	B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477	1.383	14.67	0
8510824	B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.9768	1.909	15.7	0
8511133	M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.4388	0.7096	3.384	44.91	0
851509	M	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	0.6917	1.127	4.303	93.99	0
852552	M	16.65	21.38	110	904.6	0.1121	0.1457	0.1525	0.0917	0.1995	0.0633	0.8068	0.9017	5.455	102.6	0
852631	M	17.14	16.4	116	912.7	0.1186	0.2276	0.2229	0.1401	0.304	0.07413	1.046	0.976	7.276	111.4	0
852763	M	14.58	21.53	97.41	644.8	0.1054	0.1868	0.1425	0.08783	0.2252	0.06924	0.2545	0.9832	2.11	21.05	0
852781	M	18.61	20.25	122.1	1094	0.0944	0.1066	0.149	0.07731	0.1697	0.05699	0.8529	1.849	5.632	93.54	
852973	M	15.3	25.27	102.4	732.4	0.1082	0.1697	0.1683	0.08751	0.1926	0.0654	0.439	1.012	3.498	43.5	0
853201	M	17.57	15.05	115	955.1	0.09847	0.1157	0.09875	0.07953	0.1739	0.06149	0.6003	0.8225	4.655	61.1	0
853401	M	18.63	25.11	124.8	1088	0.1064	0.1887	0.2319	0.1244	0.2183	0.06197	0.8307	1.466	5.574	105	0

The data set comprises 699 instances, each consisting of 10 attributes. These attributes are represented by integers, and they are as follows:

1. Sample code number: a unique identifier for each instance.
2. Clump Thickness: ranging from 1 - 10, indicating how densely packed the cells are.
3. Uniformity of Cell Size: ranging from 1 - 10, assessing the consistency in the size of cells.
4. Uniformity of Cell Shape: ranging from 1 - 10, assessing the consistency in the shape of cells.
5. Marginal Adhesion: ranging from 1 - 10, assessing how much the cells stick to the border.
6. Single Epithelial Cell Size: ranging from 1 - 10, assessing the size of individual cells.
7. Bare Nuclei: ranging from 1 - 10, indicating the proportion of cells without a nucleus.
8. Bland Chromatin: ranging from 1 - 10, indicating the texture of the nucleus in a light microscopic view.
9. Normal Nucleoli: ranging from 1 - 10, assessing the size and shape of the nucleoli.
10. Mitoses: ranging from 1 - 10, indicating the level of cell division.

In addition, there is a classification attribute that indicates the severity of the cancer: 2 for benign and 4 for malignant. All attributes are treated as numerical for this study.

A few sample observations from the data are as follows:

- Instance 1: 1000025,5,1,1,1,2,1,3,1,1,2

- Instance 2: 1002945,5,4,4,5,7,10,3,2,1,2
- Instance 3: 1015425,3,1,1,1,2,2,3,1,1,2

The data is of considerable interest due to its implications in breast cancer diagnosis and understanding. The data set's breadth and depth of attributes allow for comprehensive analysis and pattern recognition, key in developing accurate diagnostic algorithms.

Research Questions:

Breast cancer is the most common cancer in women worldwide, with an estimated 2.3 million new cases diagnosed in 2020 alone. While significant advances have been made in the treatment of breast cancer, early detection remains critical to improving outcomes and reducing mortality rates. Detecting breast cancer early is essential because it can lead to more treatment options, less invasive treatments, and a better chance of survival. Late detection, on the other hand, can lead to more aggressive therapies, poorer outcomes, and reduced quality of life.

To better understand which elements in breast cancer medical reports have the most significant impact on the cause of breast cancer, several research questions can be explored:

1. What are the most common risk factors associated with breast cancer? This research question will help identify the factors that increase a person's risk of developing breast cancer in the report.
2. How can breast cancer medical reports be used to predict the risk of developing breast cancer? This research question will investigate the use of models such as KNN, Naive Bayes, regularization, and random forest, to identify people at higher risk for developing breast cancer. Based on the report, it will explore how these reports can help patients better understand the risk factors for breast cancer.

By exploring these research questions, we can gain a better understanding of the factors that contribute to the development of breast cancer and identify strategies to reduce the risk of developing the disease. Additionally, we can identify the most effective detection method and improve outcomes for people diagnosed with breast cancer. Ultimately, this research can help us improve breast cancer prevention, detection, and treatment and reduce the significant burden of this disease on individuals, families, and society.

Methodology:

The models we considered for performing breast cancer study are:

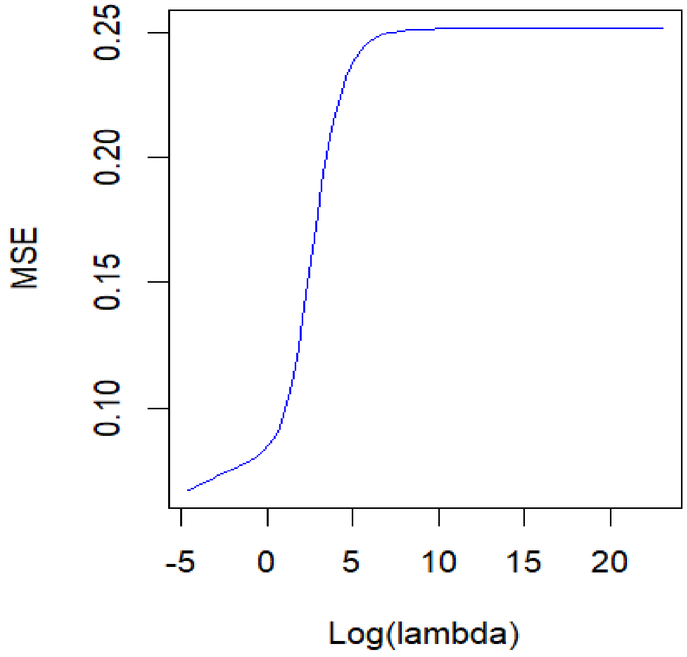
1. Random Forest
2. Classification Tree
3. Ridge Regression
4. KNN Model
5. Naive bayes model

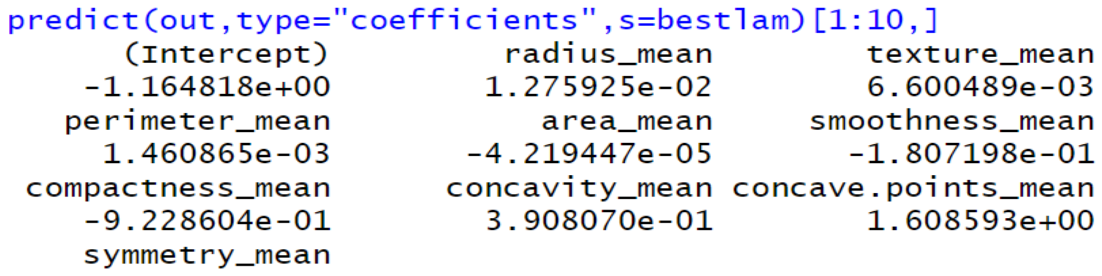
Result and Finding:

Ridge Regression:

In this project, we aim to explore Ridge Regression as a machine learning model to predict breast cancer diagnosis. To achieve this we collected data from breast cancer medical exam reports, and the dataset contained 569 observations with 31 features. We preprocessed the data by removing the ID column, converting the diagnosis column to a factor variable, and creating dummy variables for categorical variables. We then split the data into training and test sets and performed Ridge Regression using the glmnet package in R. We used a grid of 100 lambda values ranging from 10^-2 to 10^10 to fit the model. We also used the validation set approach to select the optimal lambda value. We compared the performance of the Ridge Regression model with the linear regression model using mean squared error (MSE).

Results:





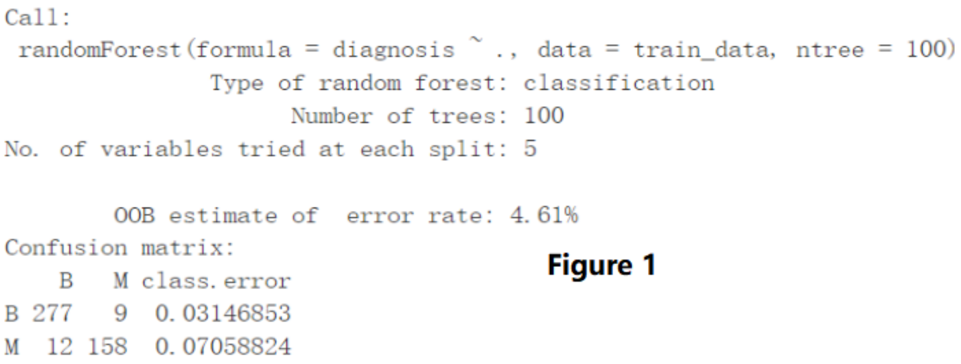
```
predict(out,type="coefficients",s=bestlam)[1:10,]
      (Intercept)          radius_mean         texture_mean 
    -1.164818e+00         1.275925e-02         6.600489e-03 
   perimeter_mean            area_mean      smoothness_mean 
     1.460865e-03        -4.219447e-05        -1.807198e-01 
 compactness_mean       concavity_mean concave.points_mean 
    -9.228604e-01         3.908070e-01         1.608593e+00 
    symmetry_mean 
```

We found that Ridge Regression did not perform better than linear regression in predicting breast cancer diagnosis. The MSE difference between the Ridge Regression model and the linear regression model was too small to conclude that Ridge Regression was a better model. Since we collected the data from breast cancer medical exam reports, the Ridge Regression model's regularization was not very useful. However,Ridge Regression produced the top 10 most important elements that affect breast cancer, which gave us very useful information.

In conclusion, Ridge Regression did not perform better than linear regression in predicting breast cancer diagnosis in our dataset. However it does provide valuable information on the top 10 most important factors affecting breast cancer.

Random Forest:

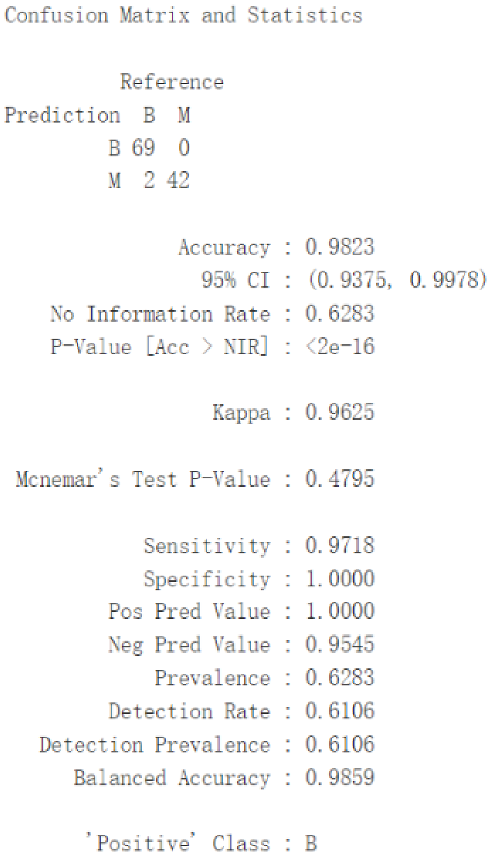
In the first part for data partitioning, we divide the dataset into training and testing sets. We utilize the "createDataPartition" function from the caret package in R to split the data. This function creates a balanced split of the data based on the target variable, 'diagnosis,' ensuring that both the training and testing sets have a similar distribution of the target variable. We have chosen a split of 80% of the data for the training set and the remaining 20% for the testing set, which is a common ratio in machine learning. This is achieved by setting the parameter 'p' equal to 0.8. A seed is also set for reproducibility purposes.



```
Call:
 randomForest(formula = diagnosis ~ ., data = train_data, ntree = 100)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 5

        OOB estimate of  error rate: 4.61%
Confusion matrix:
    B   M class.error
B 277   9  0.03146853
M  12 158  0.07058824
```

Figure 1

In the second part of Random Forest Model Training, we train a Random Forest model on the training data. The target variable is 'diagnosis', and all other variables are used as predictors. We use the "randomForest" function from the randomForest package in R. The 'ntree' argument is set to 100, which means we are creating an ensemble of 100 decision trees. Each tree is trained on a different subset of the training data and makes its own predictions. The final prediction of the Random Forest model is determined by the majority vote among the predictions made by the individual trees (Figure 1).



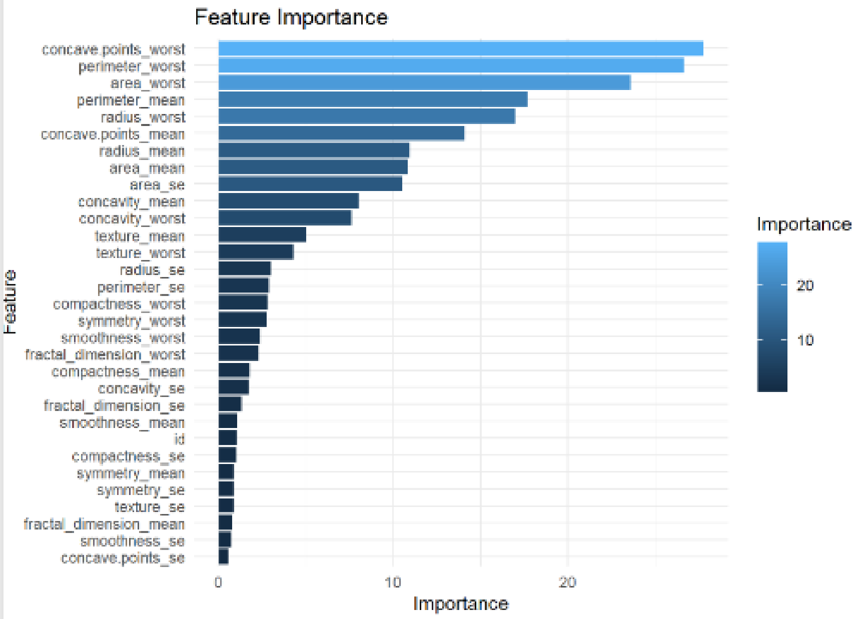
```
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 69  0
         M  2 42

               Accuracy : 0.9823
                 95% CI : (0.9375, 0.9978)
    No Information Rate : 0.6283
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9625

 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 0.9718
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9545
             Prevalence : 0.6283
         Detection Rate : 0.6106
   Detection Prevalence : 0.6106
      Balanced Accuracy : 0.9859

       'Positive' Class : B
```

Figure 2

In the third part of model evaluation, we evaluate the performance of the trained Random Forest model on the testing data. We use the 'predict' function to generate predictions on the test data and then create a confusion matrix to compare these predictions with the actual 'diagnosis' values in the test data. The 'confusionMatrix' function from the caret package provides a detailed summary of the model's performance, including metrics such as accuracy, sensitivity, and specificity (Figure 2).

In the final part of feature importance, we compute and plot the feature importance based on the trained Random Forest model. The importance of a feature is determined by the 'MeanDecreaseGini' metric, which measures the total decrease in node impurity (weighted by the probability of reaching that node) that results from splits over that variable, averaged over all trees. The 'ggplot2' package is used to create a bar plot of the feature importances. This provides insights into which variables are most influential in predicting the 'diagnosis' (Figure 3).

Figure 3



Classification Tree：

Classification trees are a popular machine learning technique used for making predictions in various domains, including healthcare. In the case of breast cancer, a classification tree can be used to predict whether a given patient is likely to have malignant or benign breast cancer.

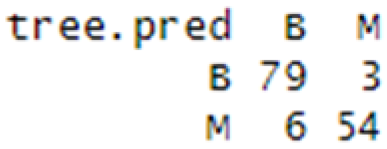
The process of building a classification tree typically involves recursively splitting the data into smaller and smaller subsets, in our case, we divided the dataset into training, test and validation set based on the features that are most informative for distinguishing between the different classes. In the context of breast cancer, some of the relevant features might include concave mean, radius mean, and so on.

We begin our analysis by loading the necessary R libraries and reading in the cancer dataset. This dataset contains various characteristics of tumor cells, such as mean radius of the tumor cell, mean area of the tumor cell, and so on. After importing the data, we convert the ‘diagnosis’ column into a factor and scale the numeric features to ensure they are on a comparable scale.

To evaluate the performance of our model, we split the dataset into a training set, test and validation dataset. This division enables us to train our model on one portion of the data and assess its performance on a separate, unseen dataset.

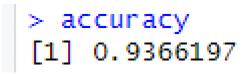
Next, we train a classification tree model on the training data.

Once our model is trained, we evaluate its performance on the test set. To do this, we use the trained model to make predictions and compare them to the actual diagnoses. We then construct a confusion matrix to visualize the model’s performance. The confusion matrix shows the number of true positive, true negative, false positive, and false negative predictions. We got the following confusion matrix.



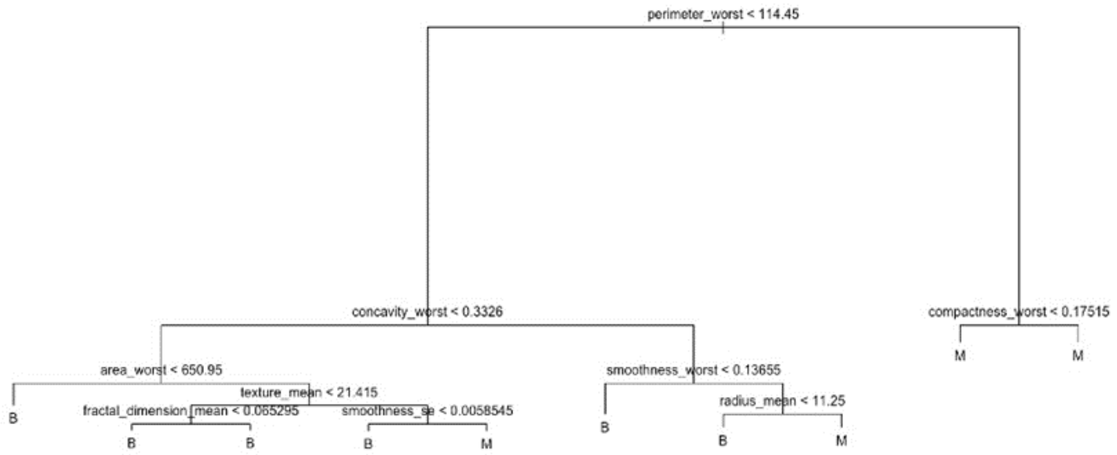
```
tree.pred  B  M
        B 79  3
        M  6 54
```

We calculated the accuracy too.



```
> accuracy
[1] 0.9366197
```

We plotted the Classification tree to look into the characteristics of tumor cell in a more understandable way:



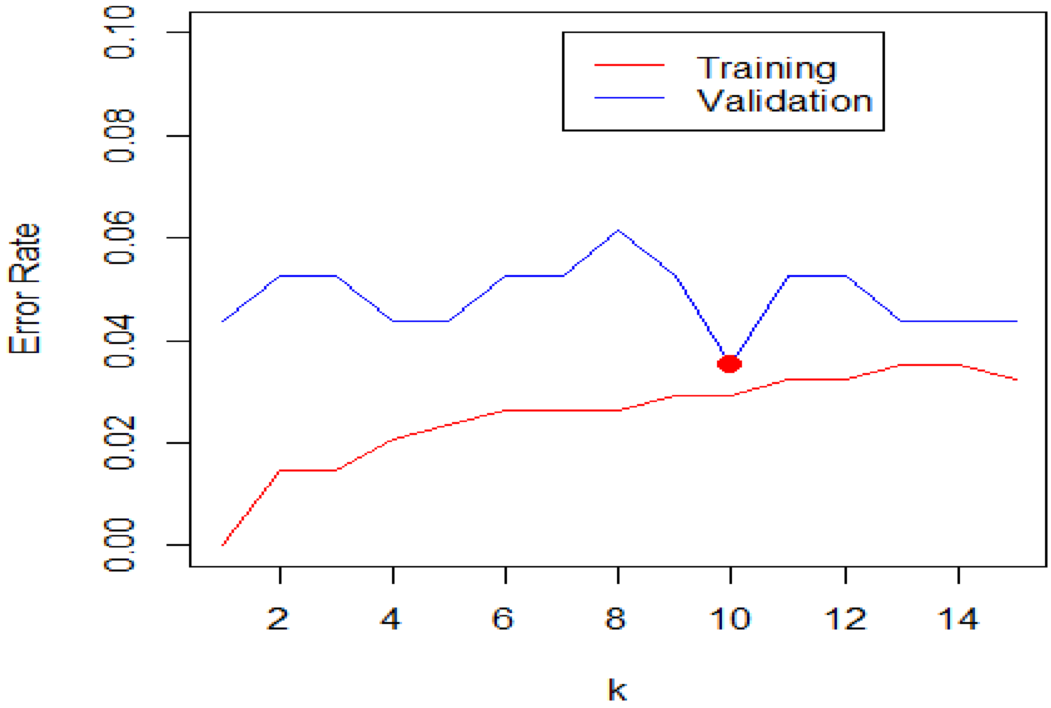
In conclusion, the decision tree analysis demonstrated the potential of this machine learning technique for breast cancer prediction. We can effectively classify breast cancer cases as malignant or benign by constructing a decision tree model based on tumor cell characteristics. The decision tree visualization helps to interpret the model's decision-making process, providing valuable insights to medical practitioners and researchers.

KNN :

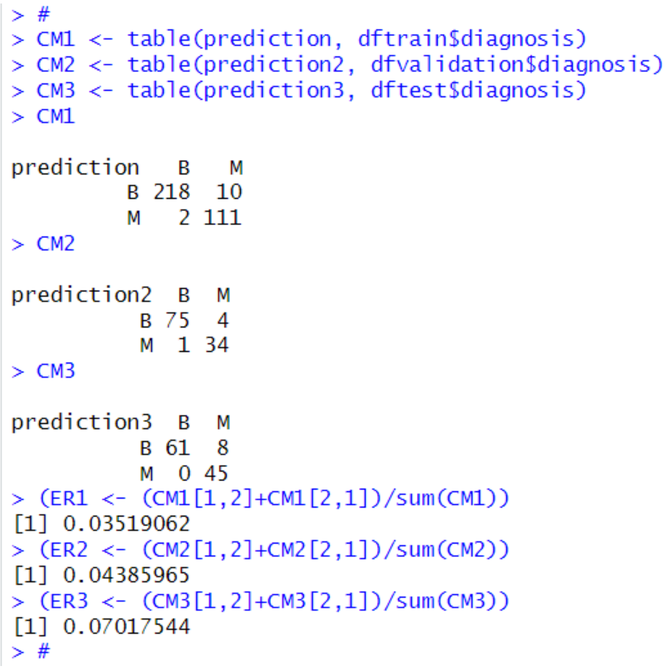
KNN method is a versatile machine learning algorithm in which class labels of nearest neighbors are used to determine the class label of an unknown record.In this model , we are trying to classify a medical record as malignant or benign. The “diagnosis” attribute was made into a factor variable with “M” as the success class by using the factor() function. “M” stands for malignant and “B” stands for benign.The “id” column in the cancer dataset was first made null as it has no significance in predicting a diagnosis.

The data was then normalized by defining a function that deducts the mean from the variable and then divides the result by the standard deviation.The data is then partitioned into training(60%), test(20%) and validation(20%) data sets.

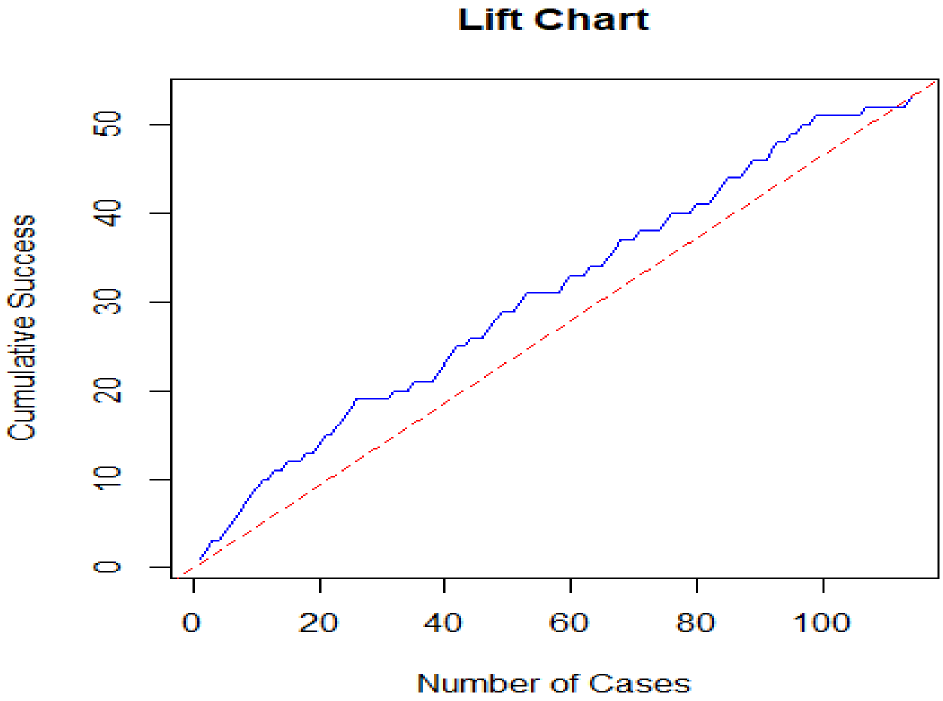
We set the maximum value of k to 15 , that is, we compute the misclassification rate for a range of values of k from 1 to 15.It is found that the minimum validation error occurs at k= 10.Therefore, 10 was chosen as the optimum value of k in the KNN model .



Results



```
> #
> CM1 <- table(prediction, dftrain$diagnosis)
> CM2 <- table(prediction2, dfvalidation$diagnosis)
> CM3 <- table(prediction3, dftest$diagnosis)
> CM1

prediction   B   M
         B 218  10
         M   2 111
> CM2

prediction2  B  M
          B 75  4
          M  1 34
> CM3

prediction3  B  M
          B 61  8
          M  0 45
> (ER1 <- (CM1[1,2]+CM1[2,1])/sum(CM1))
[1] 0.03519062
> (ER2 <- (CM2[1,2]+CM2[2,1])/sum(CM2))
[1] 0.04385965
> (ER3 <- (CM3[1,2]+CM3[2,1])/sum(CM3))
[1] 0.07017544
> #
```

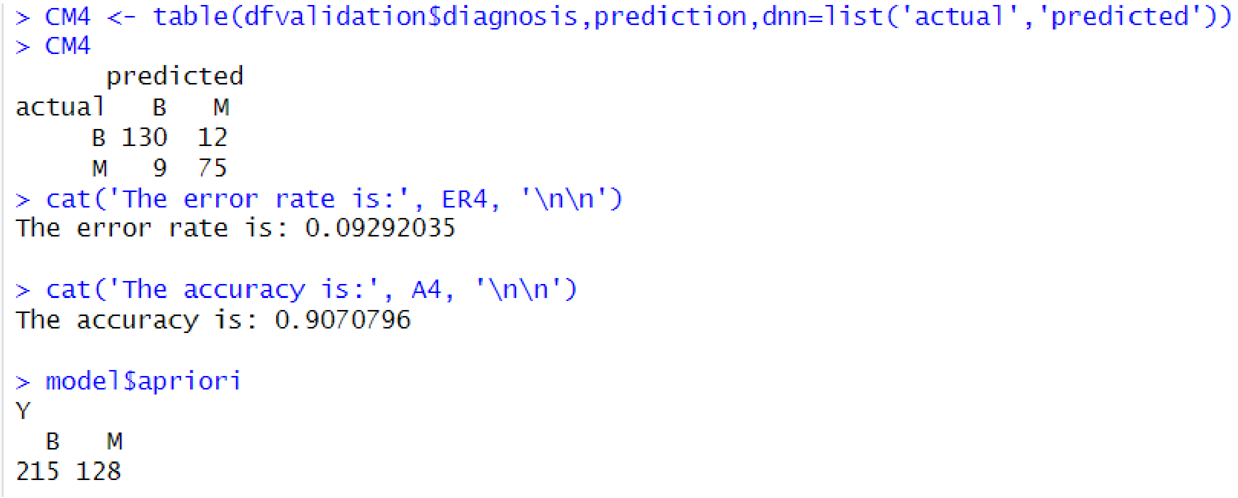


From our findings, we know that the optimum k value for generating predictions in our model is 10. As expected , the error rate on the training data set is the lowest and error rate on the test data set is the highest.The test error rate gives us an idea of how well our KNN model performs.In this case, the error rate of the test data set is 0.07 , which means the accuracy is 0.93.This is relatively high and we can thus conclude that the KNN model performs decently well.The lift chart tells us that our model is better than the baseline.

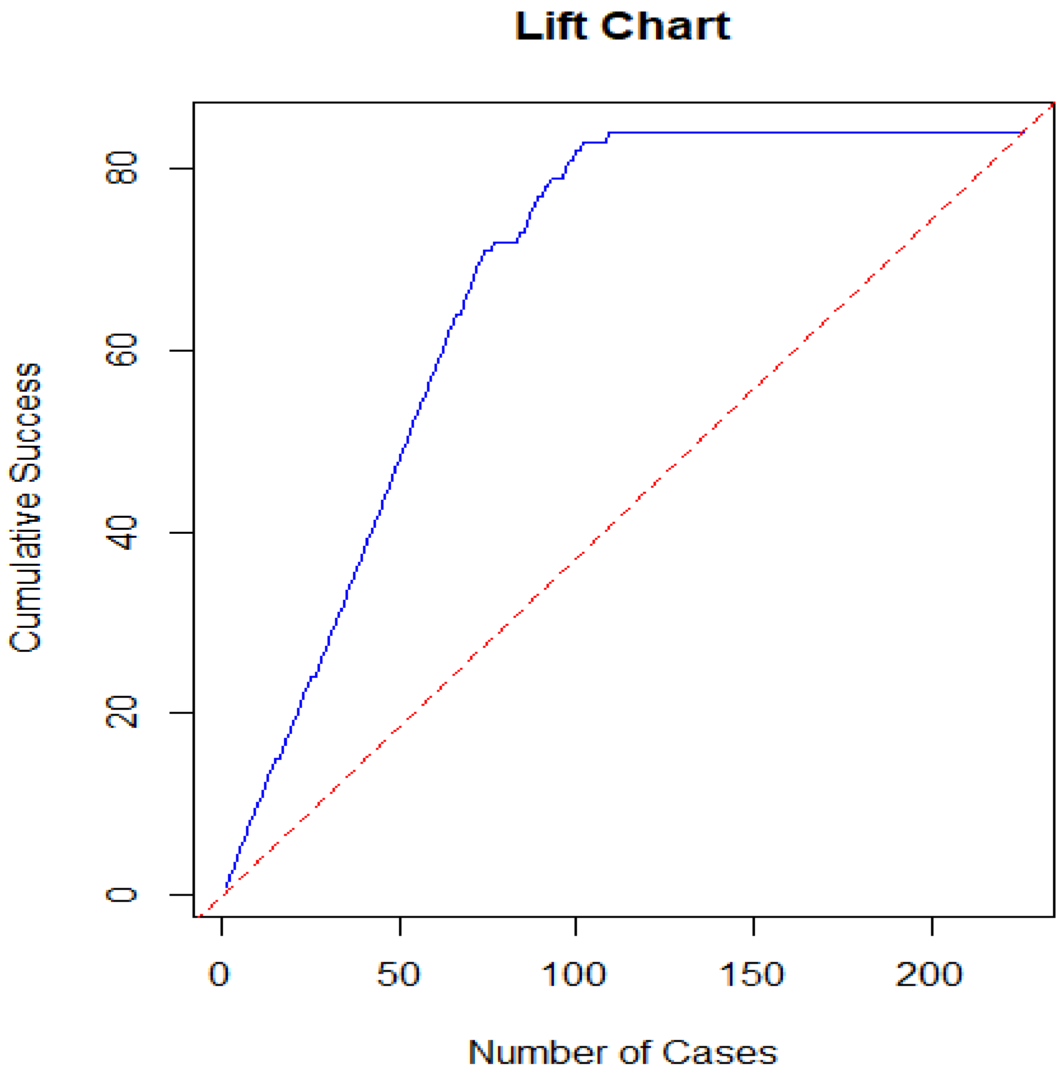
Naive- Bayes

The Naive-Bayes algorithm is another classification technique based on Bayes' Theorem on conditional probability.The library "caret" is loaded.Seed is set to 12345.The data is partitioned into training(60%) and validation(40%) data sets.To create the Naive- Bayes Model, naiveBayes() function is used from the e1071 library.Predict() function is used to generate class predictions that can later be tabled into a confusion matrix which can give us other performance metrics like the error rate, specificity, etc.The model$apriori function gives us the prior probabilities of the class distributions before the conditioning information.

Results:



```
> CM4 <- table(dfvalidation$diagnosis,prediction,dnn=list('actual','predicted'))
> CM4
      predicted
actual   B   M
     B 130  12
     M   9  75
> cat('The error rate is:', ER4, '\n\n')
The error rate is: 0.09292035

> cat('The accuracy is:', A4, '\n\n')
The accuracy is: 0.9070796

> model$apriori
Y
  B   M
215 128
```



From the above output, it can be seen that the model performs reasonably well as the number of errors in the confusion matrix are few.The accuracy rate for the validation data set is 0.907. From the lift chart, we can see that the model performs much better than the baseline.

Therefore, the Naive-Bayes model proved to be effective in classifying records in the validation data set as malignant or benign.

Conclusion

In conclusion, the analysis shows that the Random Forest model is the best predictor of breast cancer with an accuracy of 0.98. This accuracy is higher than that of other models such as KNN which achieved an accuracy of 0.93, Naive Bayes with an accuracy of 0.907 and Classification Tree with an accuracy of 0.94. Therefore, the Random Forest model can be relied upon to accurately predict the presence or absence of breast cancer in patients.

However, it is worth noting that other models such as Ridge Regression and Classification Tree still provide insightful information that can help patients better understand which parts of the data to focus on. These models may not be as accurate as the Random Forest model, but they can still provide valuable insights that can be used to make informed decisions about a patient's health.

In summary, while the Random Forest model is the best predictor of breast cancer, other models can also provide useful information that can be used to improve patient outcomes. Therefore, it is important to carefully consider the strengths and limitations of each model before making any decisions about patient care.

Appendix

UCI Machine Learning Repository: Breast Cancer wisconsin (diagnostic) data set. (n.d.).
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29