FUNDAMENTALS OF MACHINE LEARNING FINAL EXAM

REPORT

PROJECT GOAL:

The Breast Cancer Wisconsin (Diagnostic) project aims to create a predictive model that, using a set of diagnostic characteristics, can accurately classify breast cancer tumors as benign or malignant.

The purpose of this analysis is to determine which features are most useful for predicting benign or malignant cancer, as well as to identify general trends that may assist us in selecting hyperparameters and models. The objective is to determine whether the breast cancer is malignant or benign. I used classification techniques from machine learning to fit a function that can predict the discrete class of new input to accomplish this.

INTRODUCTION:

This is an analysis of the Breast Cancer Wisconsin (Diagnostic) DataSet, obtained from Kaggle. This data set was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. To create the dataset Dr. Wolberg used fluid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which is capable of perform the analysis of cytological features based on a digital scan. The program uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, then it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector.

DATA DESCRIPTION:

Attribute Information:
1)ID number
2) Diagnosis (M = malignant, B = benign)

3-32) Ten real-valued features are computed for each cell nucleus:

a. radius (mean of distances from center to points on the perimeter)
b. texture (standard deviation of gray-scale values)
c. perimeter
d. area
e. smoothness (local variation in radius lengths)
f. compactness (perimeter^2 / area - 1.0)
g. concavity (severity of concave portions of the contour)
h. concave points (number of concave portions of the contour)
i. symmetry
j. fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

Summary Statistics:

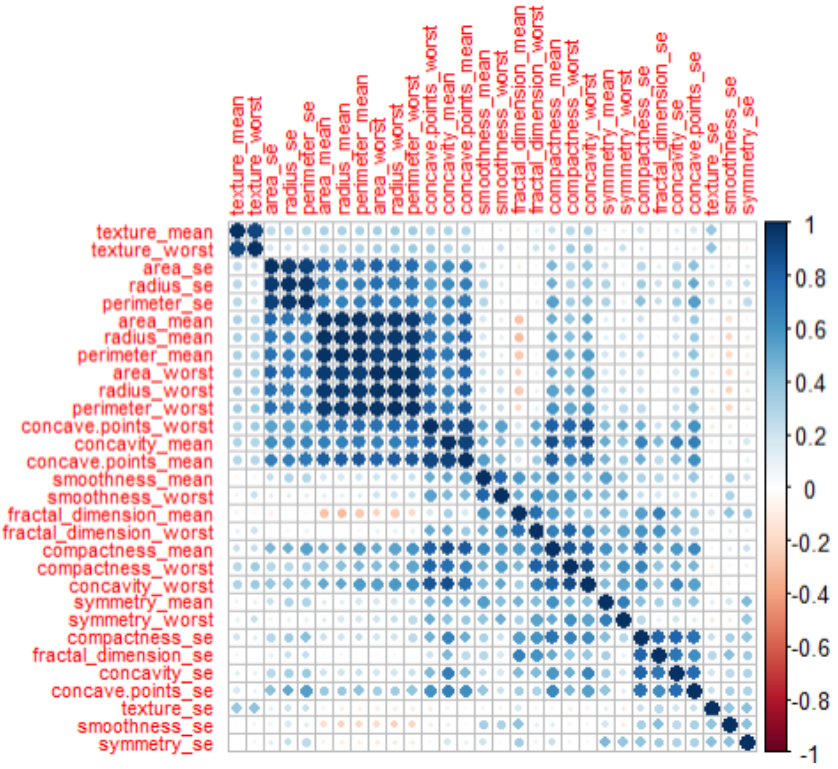
This dataset contains 569 samples of Breast Cancer which are of categories Malignant cells and Benign cells. There are 10 attributes for each of these samples which are Radius, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry and Fracture. The Mean value, Standard Error and Worst value (mean of the three largest values are calculated from the original data).

Missing Values:

There are 569 missing or Null values in this dataset in the X attributes. As the X and ID attributes do not contributes to dataset, we can remove these attributes.

GRAPHICAL REPRESENTATION:

Correlation plot



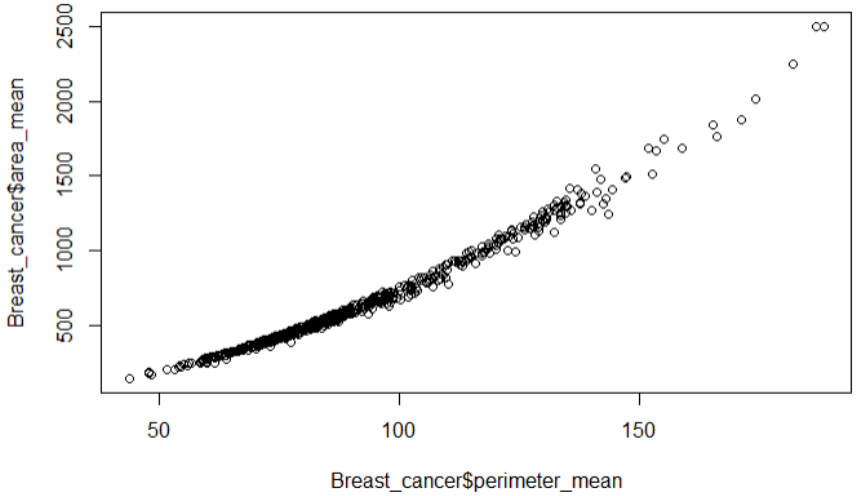
From the above graph it is concluded that we frequently have highly connected traits that offer duplicate information. To prevent a prediction bias for the data these features include by removing highly linked features. This demonstrates the need to remember that just because a feature is useful for predicting an outcome does not mean that it is causative; rather, it may merely be linked with other causal factors when making claims about the biological or medical significance of that feature.

Maintaining the feature with the lower mean while eliminating all characteristics with correlations greater than 0.9.

These are the attributes that are highly correlated so we are eliminating them.

compactness_mean , concavity_mean , texture_worst , fractal_dimension_se texture_mean, perimeter_worst , diagnosis, texture_se, perimeter_se,radius_mean.

ScatterPlot



There is a positive relationship between both the attributes i.e Area mean and Perimeter Mean. When one variable increase then the other variable also increases and vice versa.

Data Partition:

Once the data has no missing values, we partition the data using the CARET package in R in the Training and Test set. The partition index is set to "0.8", which refers to the training set of 80% of the entire dataset and the test set to 20%.

```
#Partiting the data to Train (80%) and Test (20%).
```{r}
Index <- createDataPartition(Breast_cancer_norm$diagnosis, p=0.8, list = FALSE)
Train <- Breast_cancer_norm[Index,]
Test <- Breast_cancer_norm[-Index,]
```
```

DETAILS OF MODELLING STRATEGY

I have used the Classification technique to build a model for this dataset. I used the KNN and Decision tree based on the 80% of the training data to determine the most accurate model for classify breast cancer tumors as benign or malignant.

Building the KNN Model:

```
k-Nearest Neighbors

569 samples
 30 predictor
  2 classes: 'B', 'M'

Pre-processing: re-scaling to [0, 1] (30)
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 569, 569, 569, 569, 569, 569, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.9516728  0.8960333
   2  0.9485958  0.8889557
   3  0.9507134  0.8933778
   4  0.9522491  0.8969000
   5  0.9548906  0.9024455
   6  0.9560318  0.9047146
   7  0.9585231  0.9100440
   8  0.9583626  0.9095792
   9  0.9613315  0.9160923
  10  0.9618997  0.9173466
  11  0.9621022  0.9178559
  12  0.9623009  0.9182257
  13  0.9624987  0.9186711
  14  0.9619756  0.9173835
  15  0.9619780  0.9174100

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 13.
```

From the above picture, We can conclude that the KNN model gives a largest value 96.24 when the K values is 13.

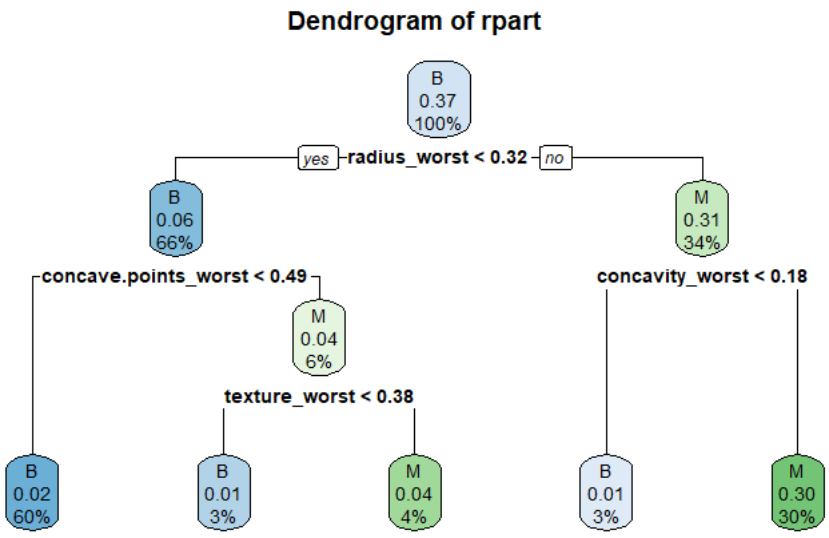
Confusion Matrix of KNN Model

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  M
##          B 71  3
##          M  0 39
##
##                Accuracy : 0.9735
##                  95% CI : (0.9244, 0.9945)
##     No Information Rate : 0.6283
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9423
##
##  Mcnemar's Test P-Value : 0.2482
##
##             Sensitivity : 1.0000
##             Specificity : 0.9286
##          Pos Pred Value : 0.9595
##          Neg Pred Value : 1.0000
##              Prevalence : 0.6283
##          Detection Rate : 0.6283
##    Detection Prevalence : 0.6549
##       Balanced Accuracy : 0.9643
##
##        'Positive' Class : B
##
```

The following conclusions have been made :-

1. Accuracy – 97.23%
2. Sensitivity – 100%
3. Specificity – 92.86%.

Building a Decision Tree Model



Confusion Matrix of Decision Tree Model

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  M
##          B 68  7
##          M  3 35
##
##                Accuracy : 0.9115
##                  95% CI : (0.8433, 0.9567)
##     No Information Rate : 0.6283
##     P-Value [Acc > NIR] : 6.062e-12
##
##                   Kappa : 0.8068
##
##  Mcnemar's Test P-Value : 0.3428
##
##             Sensitivity : 0.9577
##             Specificity : 0.8333
##          Pos Pred Value : 0.9067
##          Neg Pred Value : 0.9211
##              Prevalence : 0.6283
##          Detection Rate : 0.6018
##    Detection Prevalence : 0.6637
##       Balanced Accuracy : 0.8955
##
##        'Positive' Class : B
```

The following conclusions have been made :-

1. Accuracy – 91.15%
2. Sensitivity –95.77%
3. Specificity – 83.33%.

From the above model, the KNN Model is the optimal model for this dataset. It is the best model to use as it has higher accuracy than the Decision Tree Model. Hence, KNN Model is the right and optimal Model to use.

INSIGHTS AND CONCLUSION:

- Starting annual mammograms should be an option for women between the ages of 40 and 44.
- Mammograms should be performed annually on women ages 45 to 54.
- Mammograms should be performed every two years or yearly for women over 55.
- As long as a woman is in good general health and has a life expectancy of ten years or more, screening should continue.

Every woman should be aware of the recognized advantages and restrictions associated with breast cancer screening. They must be aware of how their breasts typically feel and seem, and they must immediately notify a healthcare professional of any changes.

Magnetic resonance imaging, generally known as MRI, should be used in addition to mammography to screen women with high lifetime risk for breast cancer due to their family history, a genetic propensity, or certain other factors.

They ought to consult a medical professional about the best screening strategy.

REFERENCE:

https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

https://rpubs.com/raviolli77/352956