Bayesian Statistics

Project Paper

Bayesian Machine Learning in Breast Cancer Diagnosis

Author: Mohammad Ehtasham Billah

Örebro University

Master of Science in Applied Statistics

Contents

1 Introduction

1.1 An Overview of Breast Cancer

Breast cancer has been a concern for several decades among the medical practitioners. Several factors determine the breast cancer including the absence of sufficient physical exercise, not conceiving at the appropriate age, avoiding pregnancy, high alcohol consumption, obesity, hormone replacement therapy during menopause, first menstruation at a premature age, smoking, genetic reasons, ionizing radiation etc. Breast cancer can happen to both the male and female, even though its rare in males. Around 5 percent to 10 percent of breast cancers are assumed to be genetical, triggered by genes (BRCA1, BRCA2, ATM, BRIP1, CDH1, CHEK2 and several others) inherited from the parents to their children.

1.2 The Potential of Data Science

Over the years, medical practitioner and scientist have been relentlessly trying to improve the medical diagnosis to identify the cancer cells at the initial stage. With the improvement of medical diagnostics and exponentially increasing generated data for the purpose of disease analysis, the survival rate has been increased in the highly developed countries but it is still very low in the underdeveloped countries. The high amount of generated data has opened a new door in medical science. The rise of artificial intelligence especially machine learning and deep learning is now turning into a vital tool in medical science to identify diseases and eventually saving lives of millions of people.

1.3 Bayesian Machine Learning in Breast Cancer Diagnosis

So, the first question is what does Machine Learning mean actually? It means when a Machine/Computer can learn from its previous experience and then utilize that experience in future without taking instruction from human. The Machine/Computer gains the experience from the dataset we provide to it during the training. Machine learning is the intersection point of Statistics and Artificial Intelligence where they supplement each other. The second question is, can we utilize Bayesian Statistical/Machine Learning technique to predict the diagnosis result based on the sample data we have collected from the patients? If yes, how can we devise the Bayesian machine Learning algorithms to the dataset and develop a model that can predict the diagnosis condition of a patient with high accuracy? The aim of this project is to answer these questions by developing such a suitable Bayesian Machine Learning model to classify and predict the patients carrying breast cancer based on different measurements of the cell nuclei. Here different measurements of the cell nuclei are the information that we provide to the machine/computer to gain experience during training. Upon developing the model, we can deploy the model in production i.e. in the diagnosis of the breast cancer.

2 Methods

2.1 Methods Used in Data Preprocessing

2.1.1 Why Data Preprocessing is Essential?

To make an efficient estimation, data preprocessing is required prior to fitting any Statistical/Machine Learning model. Most of the time, the data we work with is not at the ideal level that one can hope for. In such cases, we need to prepare the data suitable for the task at hand. Since Supervised Machine Learning models heavily rely on the relation between the dependent variable and the independent variable, a clean dataset can significantly improve the efficiency of the model. Systematic data preprocessing will assist us in attaining the clean dataset.

2.1.2 Techniques Applied for Data Preprocessing

Several statistical methods were used at the data preprocessing stage including:

Centering and Scaling

To keep the values of a variable on the same scale and hence maintain the numerical firmness throughout the dataset Centering $(x_i - \bar{x})$ and Scaling were implemented to every independent variable. This step is required because for example, Euclidean distance between two data points for a variable can easily be dominated by Euclidean distance of some other data points of a different variable. After centering, all the independent variables have the zero mean. Scaling was performed by dividing each value of an independent variable by its standard deviation. Because of applying scaling, the standard deviation has tuned into one for all independent variables.

Reducing Skewness from the Data

One of the assumptions made on this project is that the variables are normally distributed. In order to make the assumption viable, I attempted to reduce the skewness from the variables. To do that, two different statistical techniques were applied. At the initial stage, for simple explanation, Box-Cox Transformation was preferred. During cross-validation, Yeo-Johnson transformation was deployed since Box-Cox transformation does not work for a variable if any its value is zero or negative.

Removing Outliers

Outliers refers to the values of a variable that are significantly differ with the other values. Outliers can provide irrelevant and wrong information during the model building process and thus it is important to remove the outliers before fitting a Machine Learning Model. In this project, removal of outliers took place by applying Spatial Sign Transformation.

Variable Selection

All independent variables may not play the effective role in defining the dependent variable. Prior to model fitting, one needs to take into consideration the selection of variables that are efficient in describing dependent variable.For variable selection, two techniques namely,Recursive Feature Elimination (RFE) and Least Absolute Shrinkage and Selection Operator (LASSO) were applied. Twenty-one variables out of thirty-one were selected as independent variables through Recursive Feature Elimination (RFE) while the number of selected independent variables was sixteen when LASSO was implemented. REF was applied using 10-fold cross validation 5 times and in each iteration "naive Bayes" algorithm was implemented to get the set of variables that returns the optimum performance. Here, during the cross-validation, performance was measured based on Accuracy, Sensitivity, Specificity, Kappa and Receiver Operating Characteristics (ROC). For variable selection with LASSO, 10-fold cross-validation was used and after regularization, variables that still contained non-zero coefficients were selected as the independent variables for fitting the Machine Learning model. At the final stage of this project, the performance of the Machine Learning models will be evaluated based on these two different techniques used for variable selection.

Splitting up the dataset into Training Set and Test Set

To fit a Bayesian machine learning model and then evaluate its performance, two sets of dataset are required. So, the entire dataset was divided into two parts namely, "training set" and "test set" with splitting ratio 0.70. i.e training set contained 70 percent of the data and the remaining are treated as a test set. The class proportion was equalized for both the dataset. The models were trained on the training set and its evaluation is done on out-of-sample set or test set.

2.2 Methods Used in Fitting The Machine Learning Model

After selecting the variables using two different techniques i.e. RFE and LASSO, I proceeded to fit the Naive-Bayes algorithm to the initial non-preprocessed dataset using 10-fold cross validation 10 times. At this phase, all sorts of data pre-processing including centering, scaling, Yeo-Johnson Power Transformations and Spatial Sign Transformation (Serneels et al. 2006) [2] were done at individual training fold to prevent Data Leakage (Shachar Kaufman et al. 2012) [4] and to achieve optimum model performance. The selection

of prior probabilities and probability density for individual observations are discussed in the next section.

3 Model

3.1 Models Considered

Two statistical models were considered for the distribution of variables in this project.

A)Normal Density function: At first, simple distributional assumption was made i.e. it was assumed that the data is normally distributed. Thus, the normal density for each independent variable on individual classes was considered.

The probability density for a normally distributed variable X can be expressed as,

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp[-\frac{(x-\mu)^2}{2\sigma^2}] \tag{1}$$

where $-\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$

B)Non-Parametric Kernel Density Function: In the second model, I applied the nonparametric kernel density (Gaussian) estimation to estimate the density of the independent variables.

$$p(x_0) = \frac{1}{n(2\pi\lambda^2)^{m/2}} \sum_{i=1}^{n} exp(-\frac{(||x_i - x_0||)^2}{2\lambda^2}) \tag{2}$$

3.2 Estimation of Prior Probabilities

Prior probabilities were calculated from the training dataset. It was done by estimating the proportion of two classes. For example, in this dataset, there are two classes in dependent variable diagnosis namely B for Benign and M for Malignant. Now, the prior probability for class B was estimated as,

$$P(Y = B) = \frac{B}{B+M} = \frac{250}{399} = 0.6266 \tag{3}$$

And the prior probability for M was estimated as,

$$P(Y = M) = \frac{M}{B+M} = \frac{149}{399} = 0.0.3734 \tag{4}$$

3.3 Estimation of Posterior Probabilities

After estimating the probability densities of the variables and the prior probabilities, I moved forward to estimate the posterior probabilities of the classes for on each observation. For this specific problem, the

Bayesian Theorem can be stated as,

$$P(Y = k|X = x) = \frac{P(X = x|Y = k)P(Y = k)}{\sum_{k=1}^{2} P(X = x|Y = k)P(Y = k)} \tag{5}$$

Here,

- $P(Y = k|X = x)$= Posterior probability which states that, given the observation X=x what is the probability that the observation belongs to class k.
- $P(X = x|Y = k) = \prod_{i=1}^{n} P(X = x_i|Y = k)$is the Joint density of independent variables given class k and i denotes the number of independent variable.
- $k = 1, 2$. i.e. B and M denotes the classes in the dataset that will be briefly discussed in the upcoming section.

After plugging in the corresponding probability value to equation (5), based on the posterior probabilities, appropriate classes were chosen for each individual observation. However, the prior probabilities of both classes were kept constant.

4 Data

4.1 Data Collection

The data used in this project is a Breast cancer diagnosis dataset. Using the digital image of Fine Needle Aspirates(FNA) from breast mass of 569 individual, different measurement was made for each cell nucleus. In the dataset, total 32 variables and 569 observations were counted.
A short description of the Variables is as follow:

- Variable 1: ID number
- Variable 2: Diagnosis (M = malignant, B = benign)
- Variable 3-32: Ten real-valued features were measured for each cell nucleus:
    1. radius (mean of distances from the center to points on the perimeter)
    2. texture (standard deviation of gray-scale values)
    3. perimeter
    4. area
    5. smoothness (local variation in radius lengths)
    6. compactness ($perimeter^2/area - 1.0$)

7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension (coastline approximation - 1)

For each of these 10 different measurements information about mean, standard deviation and the worst condition for every individual was assessed and stored as independent variables. Note that, the size of cell nuclei is associated with the breast cancer-causing genes (BRCA1, BRCA2, ATM, BRIP1, CDH1, CHEK2 and several others).
Variable 1 indicates the ID number of individuals. Note that, in fitting the Bayesian Machine Learning model, variable 2 i.e. Diagnosis was treated as the dependent variable which contains two classes "B" and "M". "B" stands for Benign and it indicates that the tumor is not cancerous. These types of tumor do not harm the surrounding tissues. On the other hand, "M" stands for Malignant tumor and it is cancerous. The malignant tumor grows uncontrollably and it badly harms the surrounding tissues.
This dataset was collected from University of California, Irvine (UCI) Machine Learning Repository. Further details of the dataset and list of scholarly papers that used this dataset can be found at the following link:

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

4.2 Exploratory Data Analysis

To understand the principal attributes of the dataset, Exploratory Data Analysis is very handy. It will assist us in identifying certain characteristics e.g the missing values, the distributional skewness of the variables, multicollinearity, outliers, proper selection of variables for model fitting etc. Following is the short description of the Exploratory Data Analysis performed in this project.

4.2.1 Missing Values

Dealing with missing values properly is very important for any sort of data analysis. Improper management with missing value can lead to wrong inference. The dataset used in this project is not affected by this limitation as there is no missing value in any of the variables.

4.2.2 Skewness

In this dataset, the variables are mostly right-skewed. To reduce the skewness from the variables, I applied Box-Cox transformation to the skewed data. This is done for simple illustration purpose as during the cross-validation Yeo-Johnson Power Transformation will be deployed.
The Box-Cox transformation can be expressed as,

Independent_Variables	Skewness_Before_Transformation	Skewness_After_Transformation
radius_mean	0.937	-0.018
texture_mean	0.647	-0.014
perimeter_mean	0.985	-0.018
area_mean	1.637	0.283
smoothness_mean	0.454	-0.067
compactness_mean	1.184	-0.034
concavity_mean	1.394	1.394
concave.points_mean	1.165	1.165
symmetry_mean	0.722	0.002
fractal_dimension_mean	1.298	0.151
radius_se	3.072	0.027
texture_se	1.638	0.029
perimeter_se	3.425	0.069
area_se	5.419	0.115
smoothness_se	2.302	-0.024
compactness_se	1.892	-0.004
concavity_se	5.084	5.084
concave.points_se	1.437	1.437
symmetry_se	2.184	0.055
fractal_dimension_se	3.903	0.012
radius_worst	1.097	0.026
texture_worst	0.496	-0.004
perimeter_worst	1.122	0.061
area_worst	1.85	0.068
smoothness_worst	0.413	0.026
compactness_worst	1.466	-0.221
concavity_worst	1.144	1.144
concave.points_worst	0.49	0.49
symmetry_worst	1.426	-0.057
fractal_dimension_worst	1.654	0.047

Figure 1: Comparing the skewness for each variable in the dataset (before and after transformation)

$$x^* = \begin{cases} \frac{x^\alpha - 1}{\alpha} & if \alpha \neq 0 \\ \log(x) & if \alpha = 0 \end{cases}$$

Here x^* is the transformed observation and α is determined through maximum likelihood estimation.

4.2.3 Multicollinearity

When one independent variable can be explained as a linear function of other independent variables, we take this characteristic as Multicollinearity. Multicollinearity can have a negative impact on the model by reducing its efficiency. In this dataset, high Multicollinearity was observed among the independent variables. Bayes theorem is based on the fact that all the predictors are independent of each other. Hence, Multicollinearity can significantly influence the efficiency of the Bayesian Machine Learning model. To reduce the impact

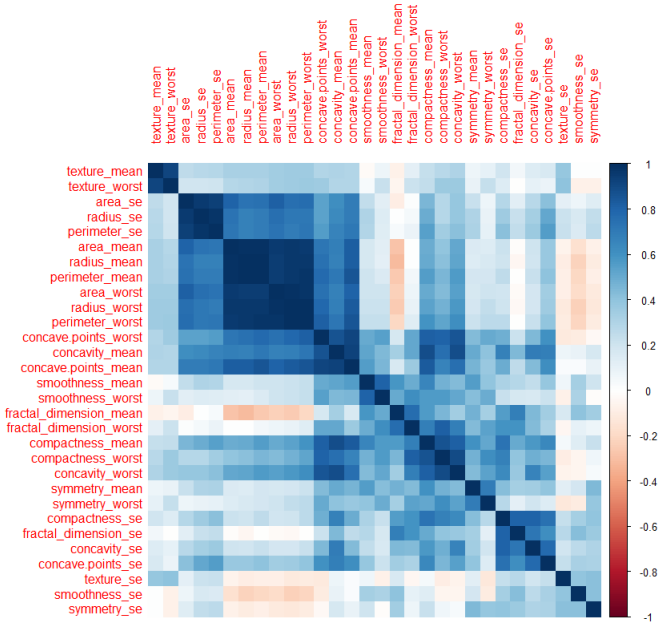

Figure 2: Depicting Multicollinearity among the independent variables. Deep red color indicates the highly negative multicollinearity and deep blue represents highly positive multicollinearity. Multicollinearity lies between -1 to +1.

of Multicollinearity, different technique Recursive Feature Elimination (RFE) and Least Absolute Shrinkage and Selection Operator (LASSO) will be applied prior to model fitting and model performance for both cases will be evaluated.

4.2.4 Outliers

When the value of a sample is relatively different from the other values of the same variables, we can treat that sample as an outlier. This is a generalization for the purpose of easy understanding. Outliers can make a model less efficient and to deal with outliers Spatial Sign Transformation (Serneels et al.,2006) [4] will be applied during cross-validation. Spatial Sign Transformation technique plots the value of independent variables into a multidimensional sphere and tries to make the distance equal for all the sample from the center of the sphere. A simple explanation of Spatial Sign Transformation is,

$$x_{ij}^{*} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^{2}}} \tag{6}$$

Here, x_{ij}^{*} is the transformed observations,i denotes the number of observations in each independent variable and j denotes the number of independent variables. The denominator is the length of vector for each individual variable or we can simply call it as the squared norm. Thus, each individual observation is divided by the length of the vector to obtain the transformed observations.

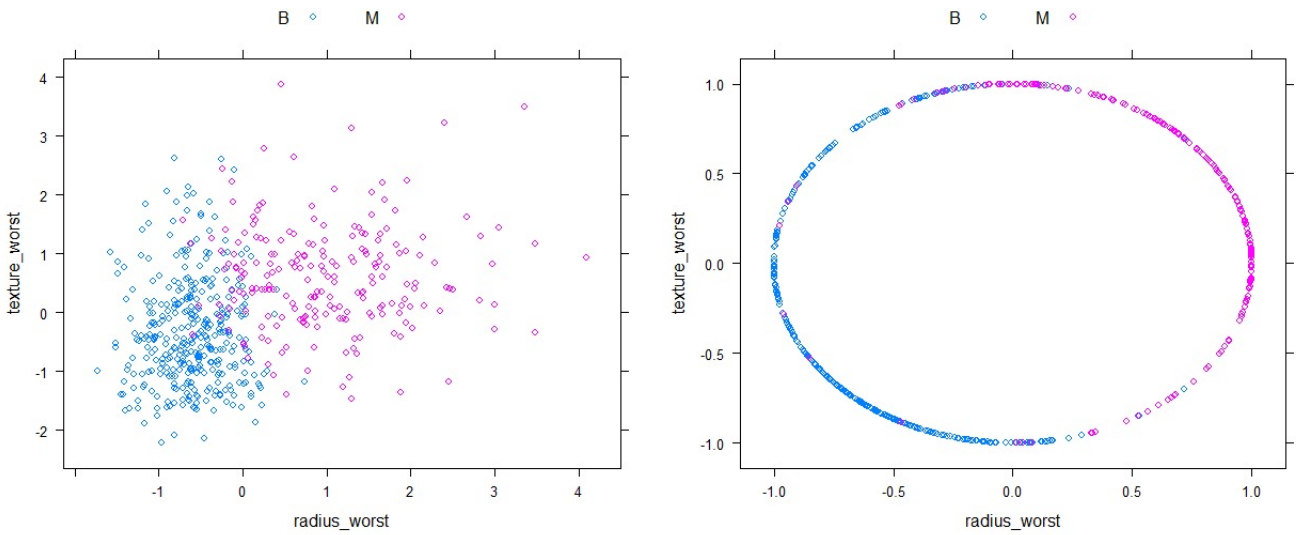

Figure 3: Image on the left portrays data points for "radius worst" and "texture worst" variables with a few "probable" outliers. On the right, after the special sign transformation, all the data points are equidistant from the center of the circle.

4.2.5 Variables Selection

Recursive Feature Elimination (RFE)

To make the model efficient twenty-one variables out of thirty-one were selected as independent variables through Recursive Feature Elimination (RFE).This technique was implemented by 10 fold cross validation 5 times and in each iteration nave Bayes algorithm was used to get the set of variables that returns the optimum performance. Here, performance was measured based on Accuracy, Sensitivity, Specificity, Kappa and Receiver Operating Characteristics (ROC).

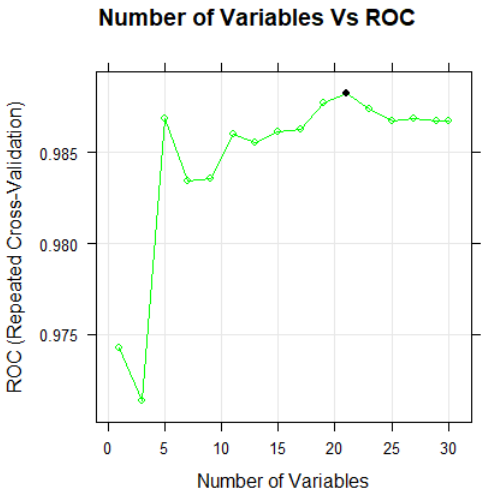

Figure 4: Set of Twenty one variables returning the highest ROC. The black circle indicates the ROC associated with the optimum number of variables.

Least Absolute Shrinkage and Selection Operator (LASSO)

By applying the Least Absolute Shrinkage and Selection Operator (LASSO) sixteen variables were selected. This technique was implemented by 10-fold cross validation 5 times and after shrinkage of the coefficients of the variables, variables that contained non-zero coefficients were selected as the independent variables for fitting the Machine Learning model. Here, performance was measured based on Misclassification Error.

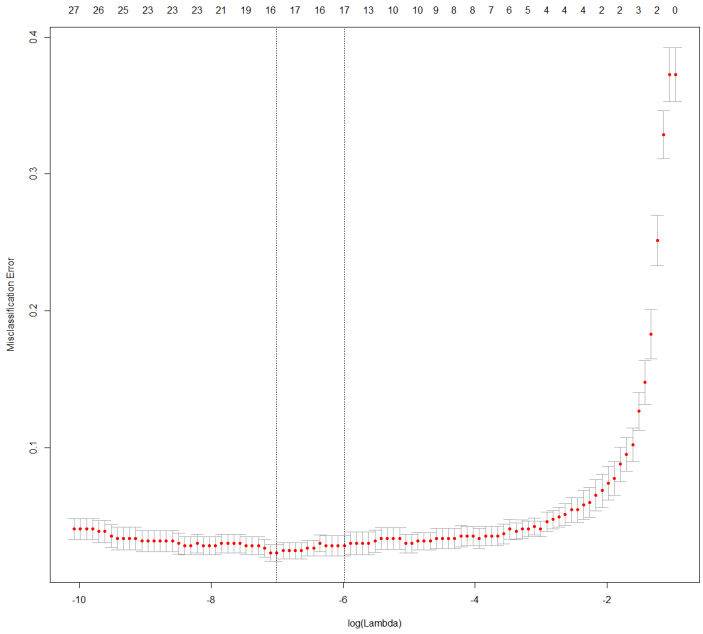

Figure 5: Variable selection by LASSO.An illustration of different values of hyperparameter and Misclassification Error. Sixteen variables were chosen (vertical line on the left) based on the value that returns the lowest misclassification error.

4.2.6 Distribution of Variables

Dependent Variable

The distribution of dependent variable diagnosis is quite imbalanced. The proportion of two classes is shown in the following table.

Benign	0.6266
Malignant	0.3734

In choosing prior probabilities for individual classes, this proportion was used.

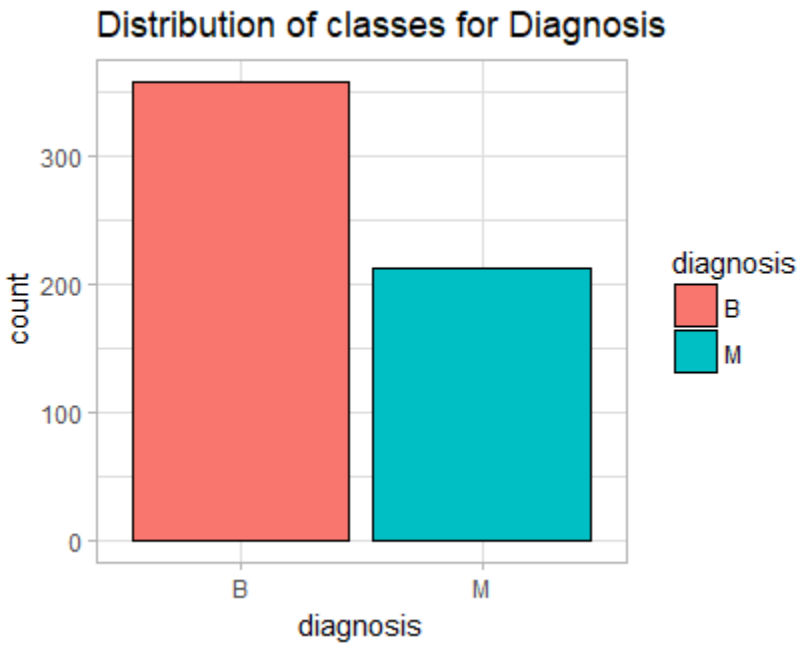

Figure 6:An Illustration of the proportion of classes "B" and "M" in response variable diagnosis.

Independent Variable

The independent variables in the dataset are not normally distributed. To remove the skewness, Box-Cox transformation was applied. Despite the fact that, the variables are not normally distributed normal density was considered for the estimation of posterior probability. For simplicity, nonparametric kernel density (Gaussian) estimation will also be considered as an alternative to normal density. This estimation technique is being used to obtain the probability estimates with more accuracy and flexibility.

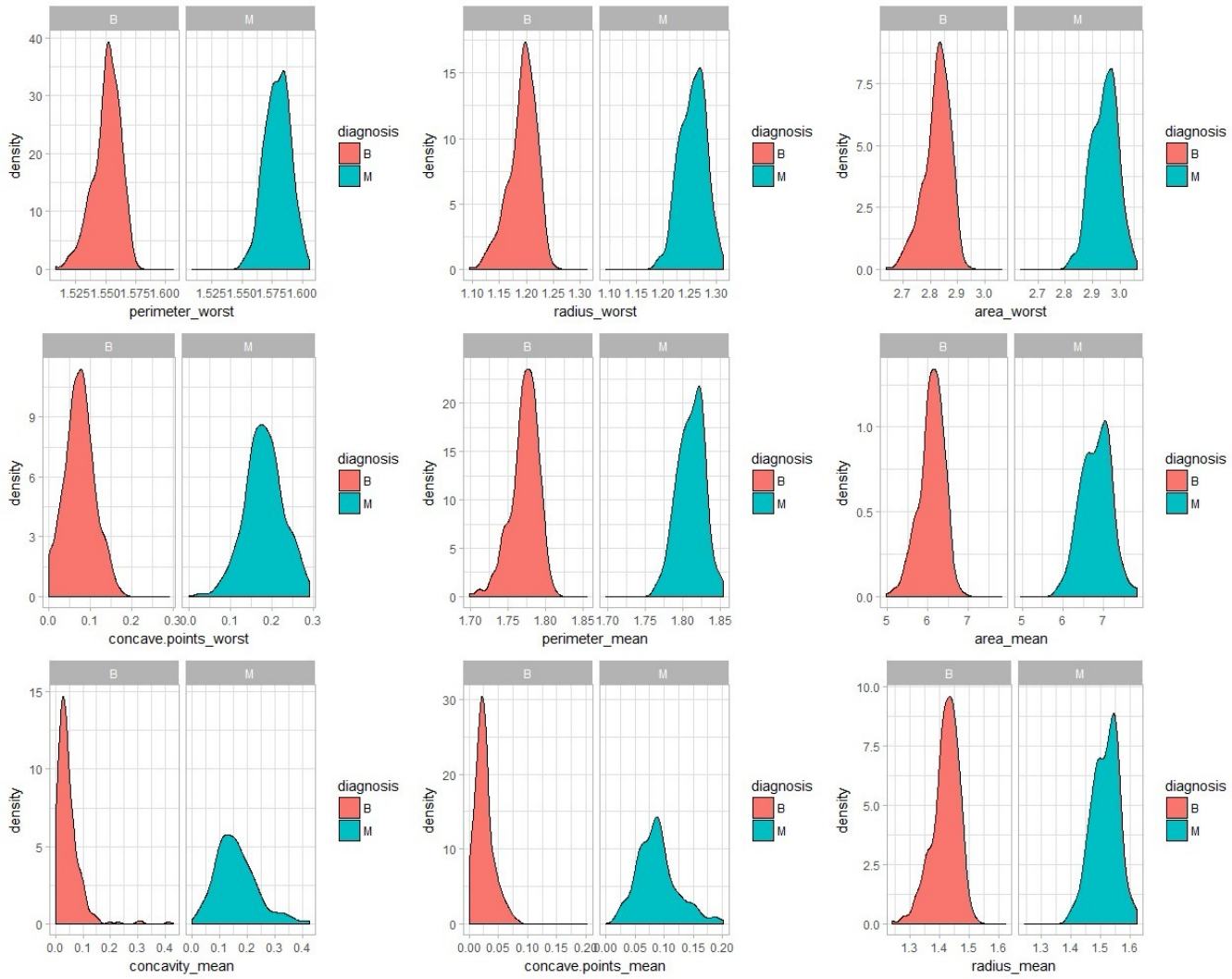

Figure 7: Distribution of a few independent variables illustrated above after Box-Cox Transformation.

4.2.7 Other General Information

Information about the variables can easily be extracted by using box plot. Box-plot provides valuable information about descriptive statistics. For example, the top-left boxplot in Fig. 7 gives us information about maximum, minimum, median, interquartile range and outliers for two classes separately. From the box-plot, it is obvious that there is a significant difference in measured values of cell nuclei between two classes i.e. Benign and Malignant. For instance, the median of the "perimeter worst" variable is different for Benign and Malignant. The data points are more centered around the median for "B" class while for "M" it is slightly scattered. For all the remaining variables this distinction existed throughout all data-points. These box plots give a clear indication of what the measured values could be for a certain class. We can also observe some outliers (big rounded dots above and below the whisker) in most of the box-plot. Spatial Sign Transformation will be utilized to deal with those outliers.

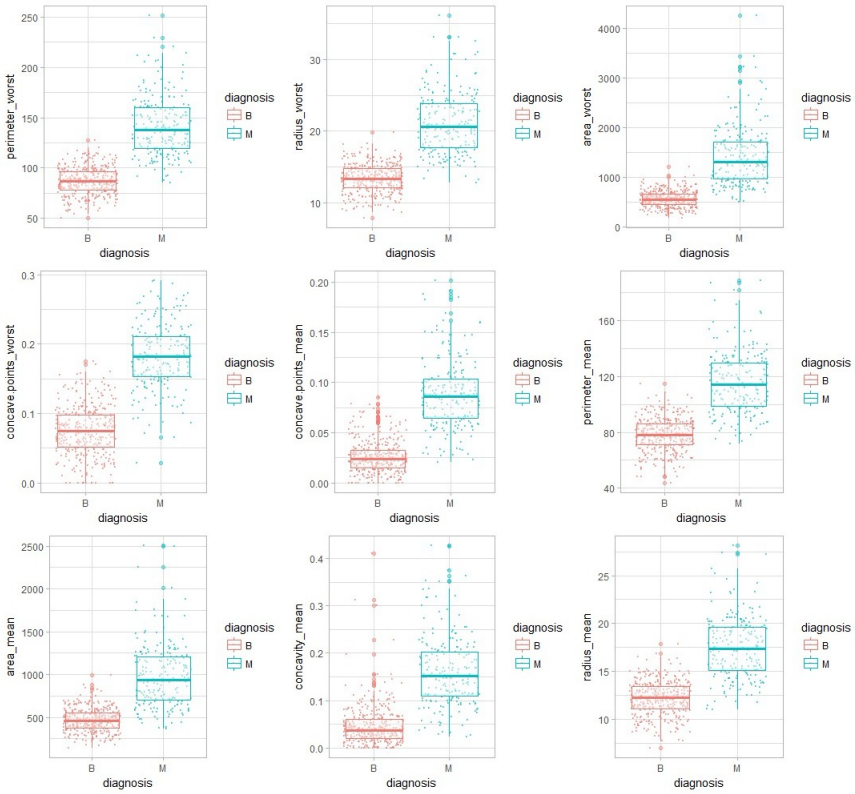

Figure 8:Box plot for a few independent variables. These box-plots provide information about median, inter-quartile range, maximum, minimum and outliers. A significant difference is observed between two classes.

5 Analysis and Results

5.1 Prior Beliefs and Probability Density Functions

5.1.1 Prior Probabilities

Before fitting the Bayesian Statistical Learning model, the dataset was split into two parts namely training set and test set. Then the model was fitted to the training set. The prior probabilities were estimated from both training set and test set and they were equal in both cases. The proportion of two classes is being considered as prior probability. Hence, the prior probability for class B in training set was estimated as,

$$P(Y = B) = \frac{B}{B + M} = \frac{250}{399} = 0.6266 \tag{7}$$

And the prior probability for M was estimated as,

$$P(Y = M) = \frac{M}{B + M} = \frac{149}{399} = 0.0.3734 \tag{8}$$

5.1.2 Probability Density Functions

For density estimation, two possibilities were considered.

- Normal Density
- Non-Parametric Kernel Density (Gaussian)

Normal Density Functions

The probability density for a normal distribution is,

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp[-\frac{(x-\mu)^2}{2\sigma^2}] \tag{9}$$

where $-\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$

The mean and variance used in normal density estimation for twenty-one variables (when RFE was used to select the variable based on two classes were estimated from the dataset and then plugged in as the value of the parameters.

	mean_benign	var_benign	mean_melignant	var_melignant
perimeter_worst	-0.613857163	0.378820509	1.029961683	0.347339189
radius_worst	-0.606196244	0.39780897	1.017107791	0.35767138
area_worst	-0.606268991	0.385704266	1.017229851	0.377637747
concave.points_worst	-0.612802422	0.284266784	1.028191983	0.512271854
concave.points_mean	-0.601271014	0.154029324	1.008843984	0.79471506
perimeter_mean	-0.577031827	0.450096353	0.968174207	0.425795559
area_mean	-0.575584543	0.378069163	0.965745877	0.55452221
concavity_mean	-0.539083048	0.283671109	0.904501758	0.897384253
radius_mean	-0.567030716	0.461581031	0.951393818	0.458229635
area_se	-0.567532086	0.436469446	0.952235044	0.497905104
concavity_worst	-0.50367986	0.441767463	0.845100437	0.798388742
perimeter_se	-0.499474584	0.575014197	0.838044605	0.593292262
radius_se	-0.492293113	0.580312996	0.825995156	0.616594629
compactness_mean	-0.464262799	0.667726761	0.778964428	0.5908108
compactness_worst	-0.450470824	0.691291912	0.755823531	0.608231154
concave.points_se	-0.309809468	0.894978251	0.519814544	0.74928427
texture_worst	-0.343121054	0.873095806	0.575706466	0.687715389
concavity_se	-0.175651155	1.220767912	0.294716703	0.49576774
texture_mean	-0.319784033	0.940593142	0.536550391	0.644133708
smoothness_worst	-0.308797491	0.863163098	0.518116595	0.805642838
symmetry_worst	-0.29482409	0.66632554	0.494671292	1.174962002

Figure 9:Mean and variance for twenty one variables based on two different class. These values were inputted as the parameter value for normal density estimation.

For instance, the density for 1st observation(x=1.593870) of "perimeter worst" variable given class "B" can be estimated as,

$$p(x = 1.593870|Y = B, \mu = -0.6138, \sigma^2 = 0.3788) = \frac{1}{\sqrt{2\pi * 0.3788}} exp[-\frac{(1.593870 + 0.6138)^2}{2 * 0.3788}]$$

And, the density for 1st observation(x=1.593870) of "perimeter worst" variable given class "M" can be estimated as following,

$$p(x = 1.593870|Y = M, \mu = 1.0299, \sigma^2 = 0.3473) = \frac{1}{\sqrt{2\pi * 0.3473}} exp[-\frac{(1.593870 - 1.0299)^2}{2 * 0.3473}]$$

Similarly, the density for 1st observation(x=1.280061) of "radius worst" variable given class "B" can be estimated as,

$$p(x = 1.280061|Y = B, \mu = -0.6062, \sigma^2 = 0.3978) = \frac{1}{\sqrt{2\pi * 0.3978}} exp[-\frac{(1.280061 + 0.6062)^2}{2 * 0.3978}]$$

Similarly, the density for 1st observation(x=1.280061) of "radius worst" variable given class "B" can be estimated as,

$$p(x = 1.280061|Y = B, \mu = 1.0171, \sigma^2 = 0.3577) = \frac{1}{\sqrt{2\pi * 0.3577}} exp[-\frac{(1.280061 - 1.0171)^2}{2 * 0.3577}]$$

We find the densities for all observations as above and take the product to obtain the Joint density given class "B" and Joint density given the class "M".

For simple explanation, I restate the Bayes Theorem in terms of two classes as follows. Bayes theorem for class "M" can be written as,

$$P(Y = B|X = x) = \frac{P(X = x|Y = B)P(Y = B)}{P(X = x|Y = B)P(Y = B) + P(X = x|Y = M)P(Y = M)} \tag{10}$$

And Bayes theorem for class "M" is,

$$P(Y = M|X = x) = \frac{P(X = x|Y = M)P(Y = M)}{P(X = x|Y = B)P(Y = B) + P(X = x|Y = M)P(Y = M)} \tag{11}$$

Here,

- $P(Y = B|X = x)$=Posterior probability which states that, given the observation X=x what is the probability that the observation belongs to class "B".
- $P(Y = M|X = x)$= Posterior probability of given the observation X=x what is the probability that the observation belongs to class "M".
- $P(X = x|Y = B) = \prod_{i=1}^{n} P(X = x_i|Y = B)$ states given Y="B" what is the Probability of observing the predictor values.
- $P(X = x|Y = M) = \prod_{i=1}^{n} P(X = x_i|Y = M)$ states given Y="M" what is the Probability of observing the predictor values.

After plugging in the corresponding joint density and prior probabilities to equation (10) and (11),

- If equation (10) is greater than 0.5 then we assign that observation to class "B", otherwise to class "M".
- If equation (11) is greater than 0.5 then we can assign that observation to class "M", otherwise to class "B".

Non-Parametric Gaussian Kernel Density Function

Let us consider the Parzen estimate as,

$$p(x) = \frac{1}{n}\sum_{i=1}^{n}\varphi_\lambda(x - x_i) \tag{12}$$

φ_λ is the Guassian Density with Mean zero and Standard Deviation λ. For k-dimensional dataset, joint Gaussian kernel density can be estimated as,

$$p(x_0) = \frac{1}{n(2\pi\lambda^2)^{m/2}}\sum_{i=1}^{n} exp(-\frac{(||x_i - x_0||)^2}{2\lambda^2}) \tag{13}$$

As before, the prior probabilities were kept same as the proportion of the two classes.
After estimating the Gaussian kernel density, I multiply it with prior probability to obtain the posterior probability from equation (10) and (11),

- If equation (10) is greater than 0.5 then we can assign that observation to class "B", otherwise to class "M".
- If equation (11) is greater than 0.5 then we can assign that observation to class "M", otherwise to class "B".

5.2 Fitting The Bayesian Machine Learning Model

Cross-validation technique was used to evaluate the model performance at each iteration.10-fold cross-validation was applied 10 times and out of 10 folds, 9 folds were used as training set and the remaining fold was the test set. Fitted model in the training set fold was evaluated at the test set fold and the performance measures at each iteration were stored. In the end, the average performance was returned.
Confusion matrix will assist us in understanding the model performance better. Here is a simple illustration of the confusion matrix.

		Actual Class			
		Benign	Malignant		
Predicted Class	Benign	P	R	Positive Predictive Value	$\frac{P}{P+R}$
	Malignant	Q	S	Negative Predictive Value	$\frac{S}{Q+S}$
		Sensitivity /True Positive Rate /Recall	Specificity	1-Specificity = False Positive Rate	
		$\frac{P}{P+Q}$	$\frac{S}{R+S}$		

Figure 10: Confusion Matrix

5.2.1 Model Fitting with Normal Density

Prior to fitting the model, I selected two sets of variables through RFE and LASSO. Based on that, model performance has been evaluated twice. With the RFE variable sets, during the cross-validation, the Receiver Operating Characteristics (ROC) was 0.9857 with sensitivity 0.9404 and Specificity 0.9472. When LASSO variable sets were used during cross-validation, the Receiver Operating Characteristics (ROC) was 0.9890 with sensitivity 0.9472 and Specificity 0.9485.

	RFE	**LASSO**
ROC	0.9857	0.9890
Sensitivity	0.9404	0.9472
Specificity	0.9472	0.9485

Figure 11: Model performance during cross-validation in two different datasets with normal density

Variable Importance

Variable importance feature helps us to assess the relative importance of the individual variables to the model. For example, Perimeter worst, concave.points worst, radius worst, area worst, concave points mean, area mean, radius mean, area se, concavity worst, perimeter se, radius se, compactness mean, compactness worst including several others had the most significant effect on determining classes.

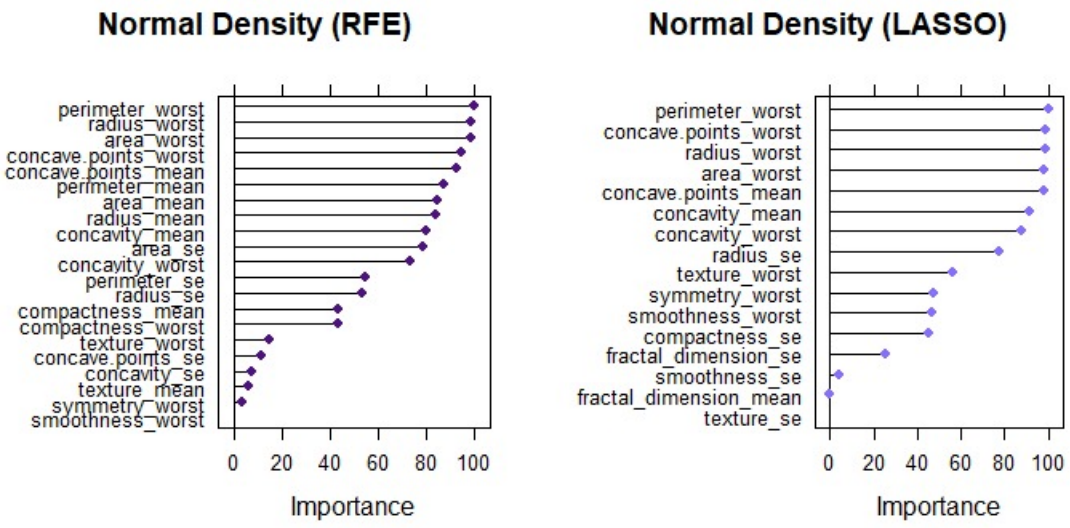

Figure 12:Variable importance in Hierarchical order. Top most variables have the most significant effect on determining class.

5.2.2 Model Fitting with Non-Parametric Kernel Density

As before, model performance has been evaluated twice based on RFE and LASSO. With the RFE variable sets, during the cross-validation, with the optimum imputation of hyperparameters, the Receiver Operating Characteristics (ROC) was 0.9885 with sensitivity 0.9528 and Specificity 0.9512. When LASSO variable sets were used during cross-validation, with optimum values of hyperparameters, the Receiver Operating Characteristics (ROC) was 0.9877 with sensitivity 0.9544 and Specificity 0.9250.

	RFE	***LASSO***
ROC	*0.9885*	*0.9877*
Sensitivity	*0.9528*	*0.9544*
Specificity	*0.9512*	*0.9250*

Figure 13:Model performance during cross-validation in two different datasets with kernel density.

Variable Importance

This feature describes the effect of individual independent variables on dependent variables. For example, Perimeter worst, concave.points worst, radius worst, area worst, concave points mean, area mean, radius mean, area se, concavity worst, perimeter se, radius se, compactness mean, compactness worst including several others had the most significant effect on determining classes in both models.

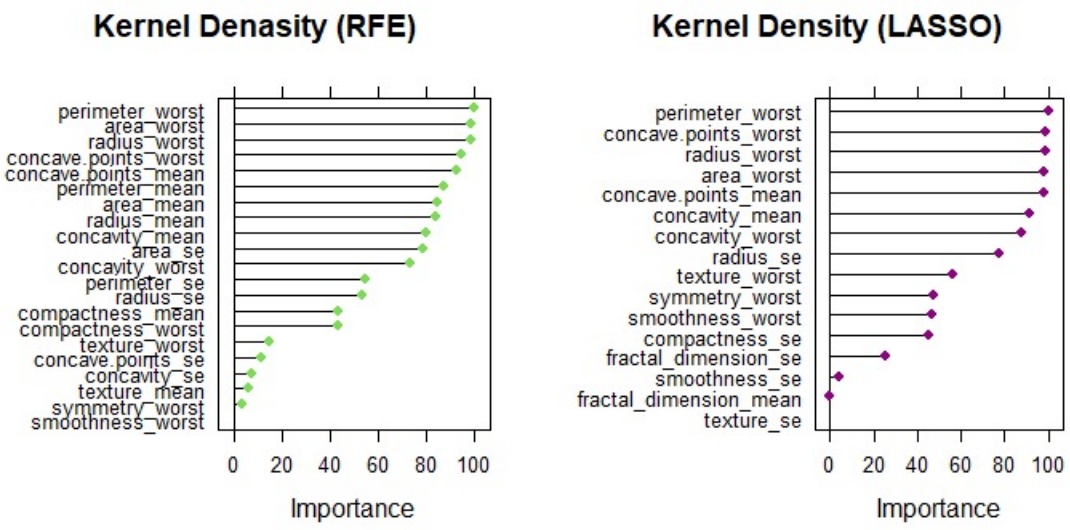

Figure 14:Variable importance in Hierarchical order. Topmost variables have the most significant effect on determining class.

5.3 Prediction(When Normal Density Function was Applied)

After fitting the Bayesian Machine Learning model, we evaluate the model on unseen dataset to understand its effectiveness and accuracy. By utilizing the knowledge the model has gained during training, it will try to predict the classes for each observation of the test set.The prediction will be made twice, one for RFE variable set and another one is for LASSO variable set.

With the RFE dataset, the model predicted the classes with 94.71% accuracy. The sensitivity was 94.39% while specificity was 95.24%.

On the other hand, while the LASSO dataset was used, the prediction accuracy was improved by almost 2 percent. Here, the model predicted classes with 96.47% accuracy with sensitivity 96.26% while specificity was 96.83%.

	RFE	***LASSO***
Accuracy	*0.9471*	*0.9647*
AUROC	*0.9482*	*0.9654*
Sensitivity	*0.9439*	*0.9626*
Specificity	*0.9524*	*0.9683*

Figure 15:Model evaluation on two different test set.

5.4 Prediction(When Non-Parametric Gaussian Kernel Density Function was Applied)

Similar to the previous section, the prediction was performed twice, one for the test set from RFE variable set and another one for test set from LASSO variable set.

With the RFE dataset, the model predicted the classes with 94.71% accuracy. The sensitivity was 93.46% while specificity was 96.83%.

With the LASSO dataset, the prediction accuracy was improved by 3 percent. Here, the model predicted classes with 97.65% accuracy with sensitivity 98.13% while specificity was 96.83%.

	RFE	LASSO
Accuracy	0.9471	0.9765
AUROC	0.9514	0.9748
Sensitivity	0.9346	0.9813
Specificity	0.9683	0.9683

Figure 16:Model evaluation on two different test set.

5.5 AUROC Curve

One effective way of measuring model performance is the Area Under Receiver Operating Characteristics (AUROC) curve. Here, True positive rate(Sensitivity) is plotted on the y-axis and False positive rate (1-Specificity) is plotted on x-axis. A model that can classify the observations without any error has a ROC value of 1 and a random model without much capability of making prediction has ROC values of 0.5.When ROC value is 1, the curve will touch the top left corner of the plot. Thus, the closer the curve gets to the top left corner, the higher the performance of the model.

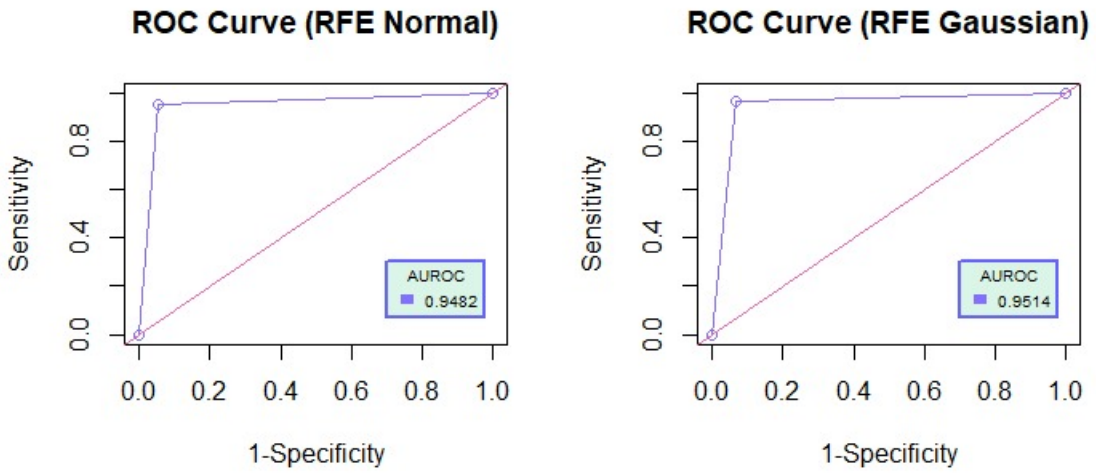

Figure 17:An illustration of the Area Under the ROC curve. The closer the curve reaches the top left corner the higher the performance of the model. With RFE, AUROC is 0.9482 and 0.9514 for normal density and kernel density respectively.

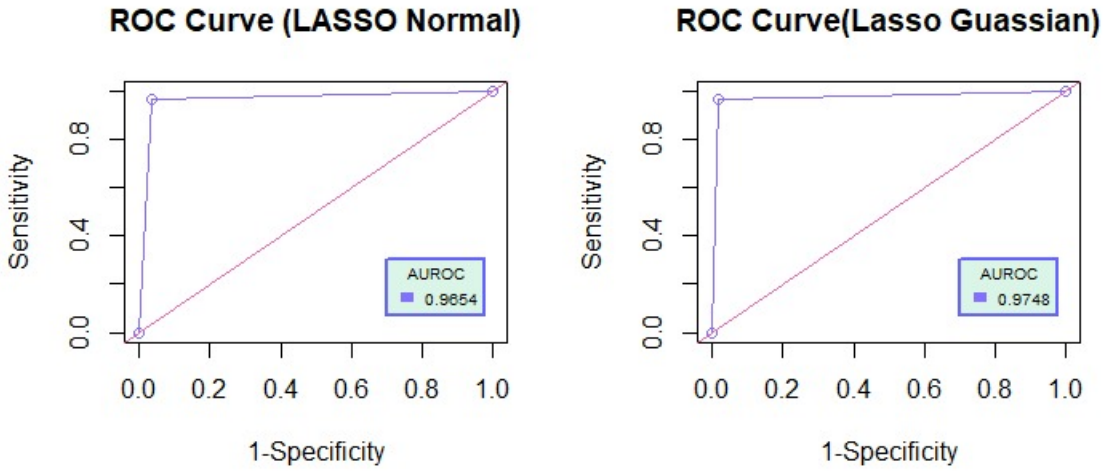

Figure 18:When variables were selected through LASSO, AUROC was 0.9654 and 0.9748 with normal density and kernel density respectively.

5.6 Model Performance on Raw data

To investigate the model performance, I fitted the models on the raw dataset and evaluated the performance. When the models were fitted with the raw data, the model performance slightly reduced.

Model Evaluation on Test Set (Prediction)

Normal Density

	Transformed Data		Raw Data	
	RFE	LASSO	RFE	LASSO
Accuracy	0.9471	0.9647	0.9462	0.9647
AUROC	0.9482	0.9654	0.9401	0.9556
Sensitivity	0.9439	0.9626	0.9523	0.9907
Specificity	0.9524	0.9683	0.9317	0.9206

Kernel Density

	Transformed Data		Raw Data	
	RFE	LASSO	RFE	LASSO
Accuracy	0.9471	0.9765	0.9471	0.9412
AUROC	0.9514	0.9748	0.9449	0.9337
Sensitivity	0.9346	0.9813	0.9533	0.9626
Specificity	0.9683	0.9683	0.9365	0.9048

Figure 18:Comparison of model performance between Transformed data(centering,scaling,Spatial Sign Transformation,Yeo-Johnson Power Transformation) and Raw data

6 Conclusion

To predict the diagnosis condition of individuals, eight Bayesian Machine Learning model was fitted. In the first two models, I considered the RFE technique for variable selection and normal density, as well as non-parametric kernel (Gaussian) density was implemented. For the third and fourth models, I considered the LASSO technique for variable selection and normal density, as well as non-parametric kernel (Gaussian) density, was implemented. Next, I applied the same techniques on the raw dataset to observe the magnitude of improvement due to data preprocessing. The Accuracy, Sensitivity, Specificity and Area Under Receiver Operating Characteristics (AUROC) curve was higher when variables were selected through LASSO.

The prediction results can be considered as robust since all eight models were fitted through 10-fold cross validation 10 times and the outcome was measured by several statistics.

Using the normal density, the prediction was pretty accurate even though the data was not "exactly" normally distributed. Bayesian Machine Learning algorithm is computationally fast and sometimes it can outperform other "high-profile" machine learning algorithms. Finding the "perfect" density might be difficult sometimes, but in such cases, other methods could be tried to make the algorithm more efficient. One alternative could be the Mixture Models. In such case, we can consider different distribution and combine them to fit a more robust model. It is imaginable that Accuracy, Specification, Sensitivity, and AUROC can be improved by experimenting with the mixture models using this same training dataset.

References

[1] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[2] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):15, 2012.

[3] Max Kuhn and Kjell Johnson. *Applied predictive modeling*, volume 26. Springer, 2013.

[4] Sven Serneels, Evert De Nolf, and Pierre J Van Espen. Spatial sign preprocessing: a simple way to impart moderate robustness to multivariate estimators. *Journal of Chemical Information and Modeling*, 46(3):1402–1409, 2006.