**DATA MINING PROJECT**

**Classification of Breast Cancer**

****

**Diksha Yadav (**[**yadavd@umail.iu.edu**](mailto:yadavd@umail.iu.edu)**)**

**Gaurav Derasaria (**[**gderasar@iu.edu**](mailto:gderasar@iu.edu)**)**

**Abstract:**

In this project, we have done a detailed study of how an algorithm performs on a multivariate dataset. We have used feature reduction before applying the algorithm to provide a computationally efficient algorithm, so that the algorithm can be scalable.

**Introduction:**

Machine Learning has been a game changer in the field of healthcare. Doctors have gained insights that have helped them make better decisions. Supercomputers like IBM Watson are raising the bar, paving a way for the others to follow. In healthcare, Watson's natural language, hypothesis generation, and evidence-based learning capabilities are being investigated to see how Watson may contribute to clinical decision support systems for use by medical professionals [1] ­. Thus, healthcare has generated a lot of datasets in the recent years. Few of these datasets have been made accessible to everyone. Wisconsin Breast Cancer dataset [2] is one such dataset that was created by University of Wisconsin.

**Dataset Description:**

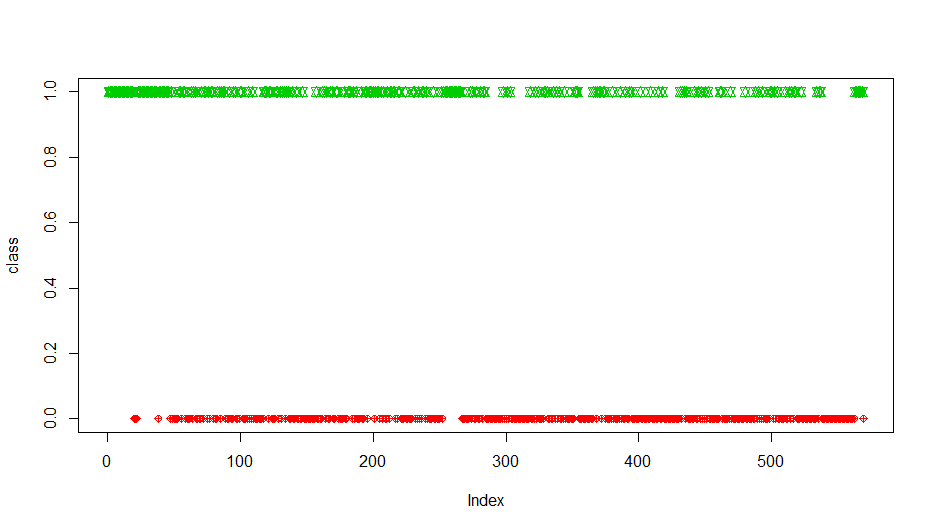
Breast Cancer Wisconsin dataset:

* Number of features :30 (real valued features)
* Class variable : 1(Diagnosis - M= Malign , B = Benign)
* ID number : 1(ID)
* Number of records: 569
* Length of Train set : 398
* Length of Test set : 171

**Visualizations:**

1. **Class Visualization**

The plot shows that the dataset is balanced as neither of the two classes is highly dominating inthe dataset and classes are in 60-40 ratio.

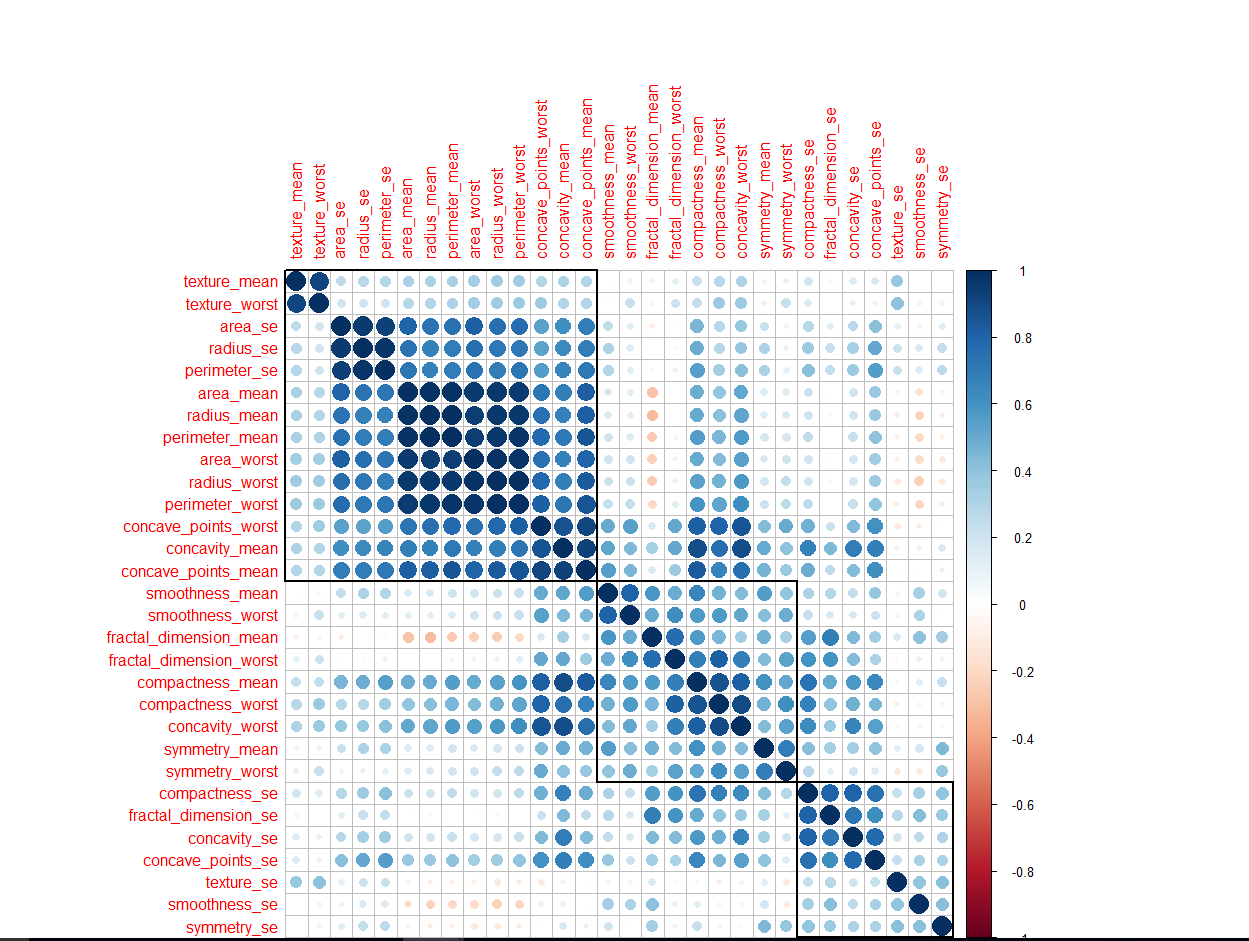
****

1. **Correlation Graph**

The plot shows that correlation between each variable of the dataset.

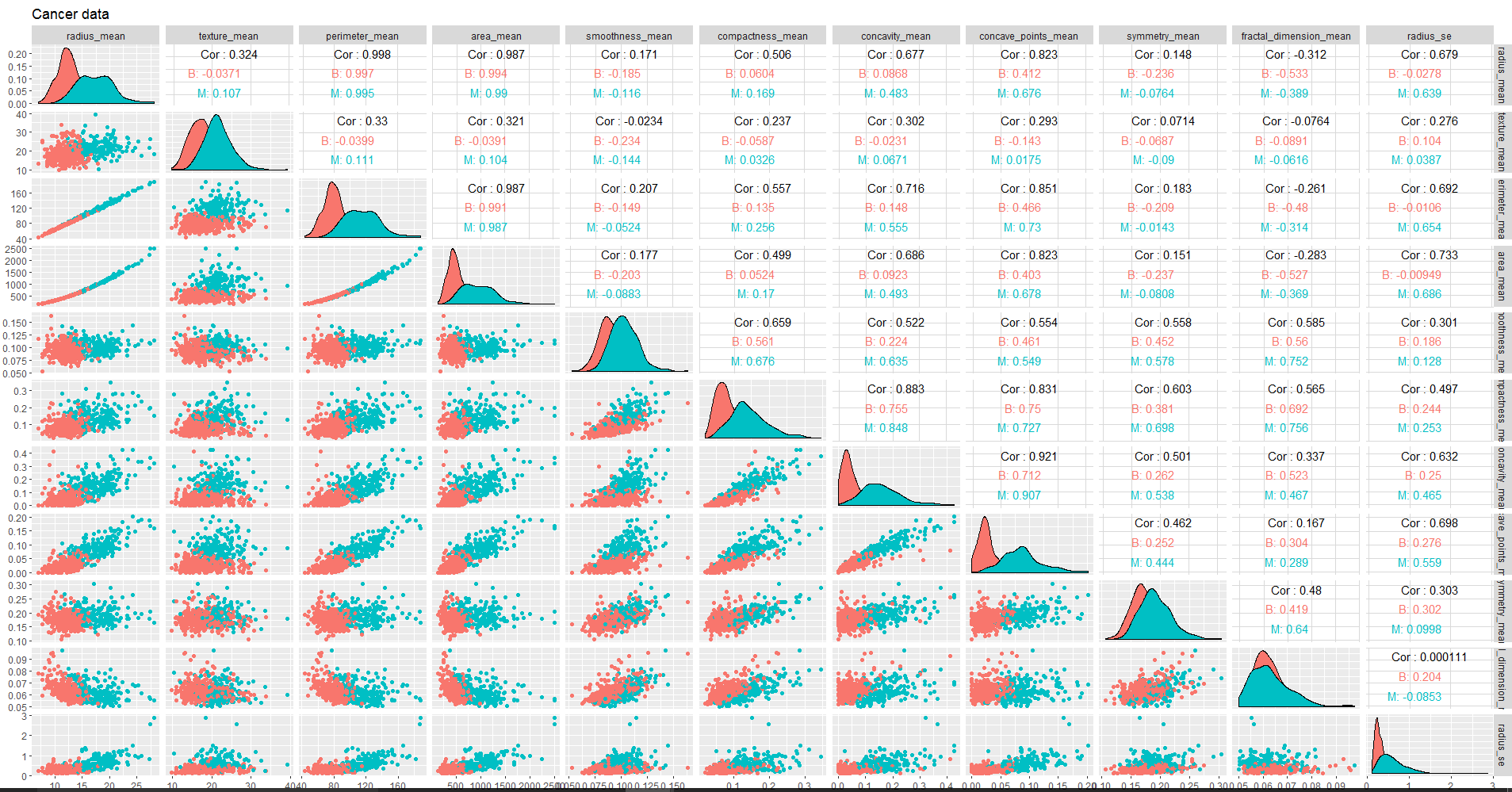
* Dark blue color signifies highly positive correlation among variables.
* Light blue color signifies lower positive correlative.

Orange color signifies negative correlation among variables respectively.

****

1. **Visualization to show that data is linearly separable**

Orange and Green color signifies two different classes of the dataset on which classification is done and the whole graph signifies that the data is linearly separable.



**RESULT:**

After applying **Support Vector Machine along without Principal Component Analysis**, we got an **ACCURACY** of **0.9883 (using Accuracy=TP+TN/TP+TN+FP+FN)**

After applying **Support Vector Machine along with Principal Component Analysis**, we got an **ACCURACY** of **0.9766 (using Accuracy=TP+TN/TP+TN+FP+FN)**

**Confusion Matrix of final model (Support Vector Machine along with Principal Component Analysis)**

**TP=True positive**

**TN=True Negative**

**FN=False Negative**

**FP=False Positive**

|  |  |  |
| --- | --- | --- |
|  | **Predicted=1** | **Predicted=0** |
| **Actual=1** | **TP=52** | **FN=2** |
| **Actual=0** | **FP=2** | **TN=115** |

Based on above confusion matrix,

**PRECISION=TP/TP+FP**

=0.9629

**RECALL=TP/TP+FN**

=0.9629

**F MEASURE=2\*PRECISION\*RECALL / (PRECISION+RECALL)**

= 0.9629

**CONCLUSION:**

We have applied 2 models, first is Support Vector Machine without PCA and second is Support Vector Machine with PCA. We applied Support Vector Machine because the data is linearly separable. Then, we applied Support Vector Machine with PCA because we had large number of variables and we wanted to reduce the dimensionality. We observed that there was just a small decrease in accuracy after applying PCA which can be ignored as our dimensionality was reduced to almost 1/4th. From our results we can conclude that there is a tradeoff between computational time/space and accuracy. Depending upon the demand of our problem, we can give preference to one accordingly. In our case we could reduce computational time/space considerably without affecting our accuracy much.