**Predicting the Cancerous Nature of a Breast Cancer Cell**

Akshay Chauhan

September 2020

**1. Introduction**

**1.1 Background**

Breast Cancer is one of the most fatal diseases affecting women, especially so in developing countries where almost 50% of women pass away within one year of being diagnosed. A substantial percentage of women are also never diagnosed. This has also to do with limited medical resources in developing countries and those resources allowing only discriminatory access to women. In general, this disease has widespread impacts on the lives of women where 1 in 28 women are diagnosed with this disease at least once in their lifetime.

**1.2 Problem**

Data will contribute to the predictive classification of cancerous cells based on the features of the cell. This will have the effect to improve accuracy of testing of cells of patients. A problem having emerged around Breast cancer has been that a number of cases do not receive early diagnosis and cases are traced only in Stage 3 of Cancer. This study will hence have the effect of providing more equitable and accurate testing.

**1.3 Interest**

This study will interest medical professionals all around the world who can now use E-Medicine along with traditional diagnostic techniques to make their diagnoses more accurate to cure their patients. In addition, this study will also interest medical researchers who study the nature of Cancer cells and drug research.

**2. Data acquisition and cleaning**

**2.1 Data sources**

The Data-set has been procured by Sci-kit Learn as a pre-set from the University of Wisconsin, Clinical Sciences Centre. This data was collected and curated in the year 1995 by a team led by Dr. William H. Wolberg and Nick Street. It contains a large number of attributes to determine the status of malignancy or benignity of Cancerous cell.

**2.2 Feature selection**

The data set is pre-cleansed, it has 569 samples and 30 features in the data.



This indicates that the data is representative for training and testing for our model, the target value of 0.0 represents Benign cells which has 212 values. While the target value of 1.0 represents Malignant cells with 357 values.

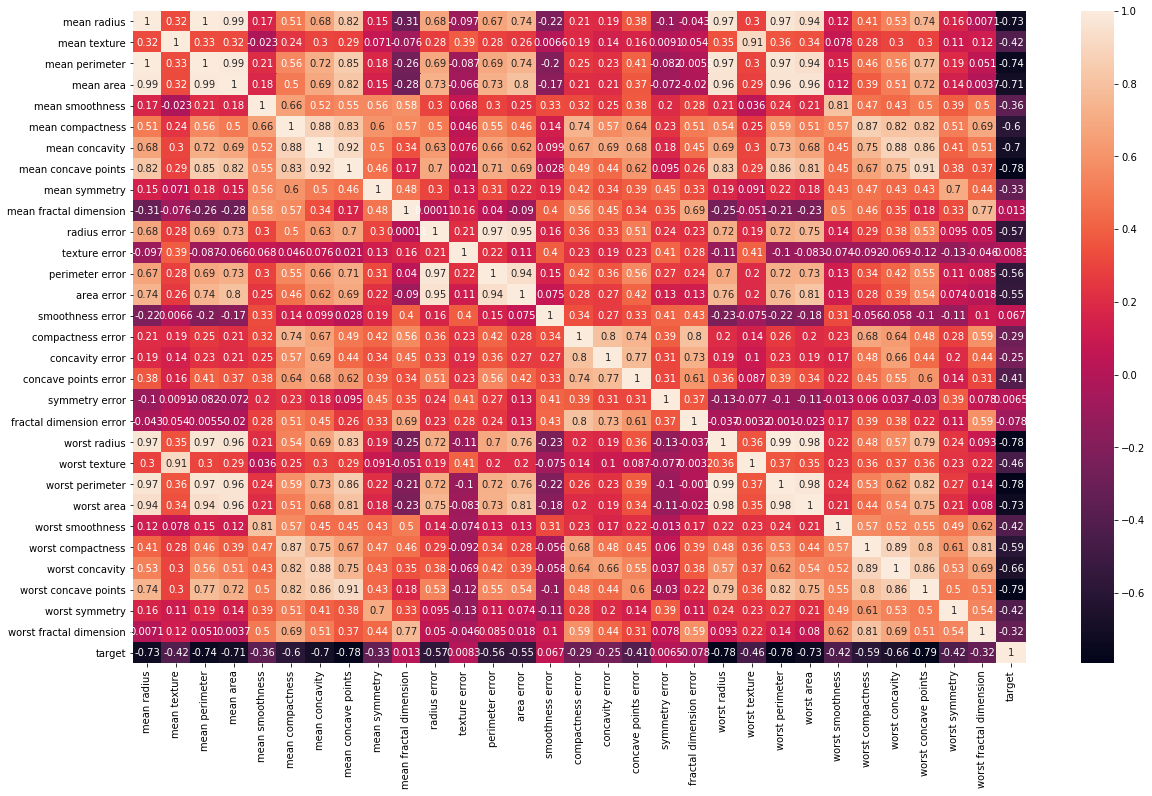
**3. Exploratory Data Analysis**

**3.1 Data Analysis Methodology**

The study first visualised and analysed the relationship of the Independent variables in themselves which was followed by plotting of Dependent variable with the Independent variables.

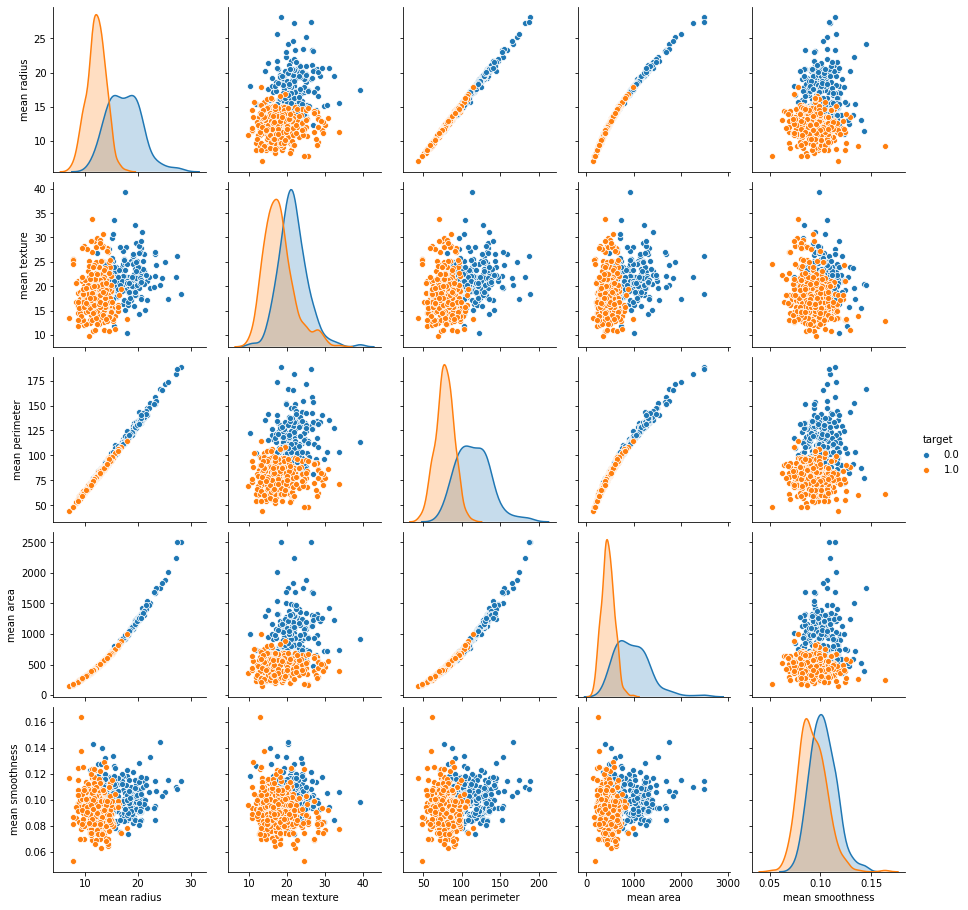
**3.2 Calculation of target variable**

The target variable for this study is classification of the status of cells into Malignant and benign status. It has already been noted for the large part of the data. The future classification based on attributes is modelled through our study. The relation around the target is depicted best through the correlation of the Independent variables to the target classification and their impact on the status of the cell.



*The heatmap represents the correlation of all variable amongst themselves as well as with the Target Variable of cell status.*

**3.2 Relationship between the Mean Radius, Mean Texture, Mean Area, Mean Perimeter, Smoothness**



*This diagram indicates the values of independent variables and how they affect the prediction of Target variables.*

*A number of relationships in the*

**4. Predictive Modelling**

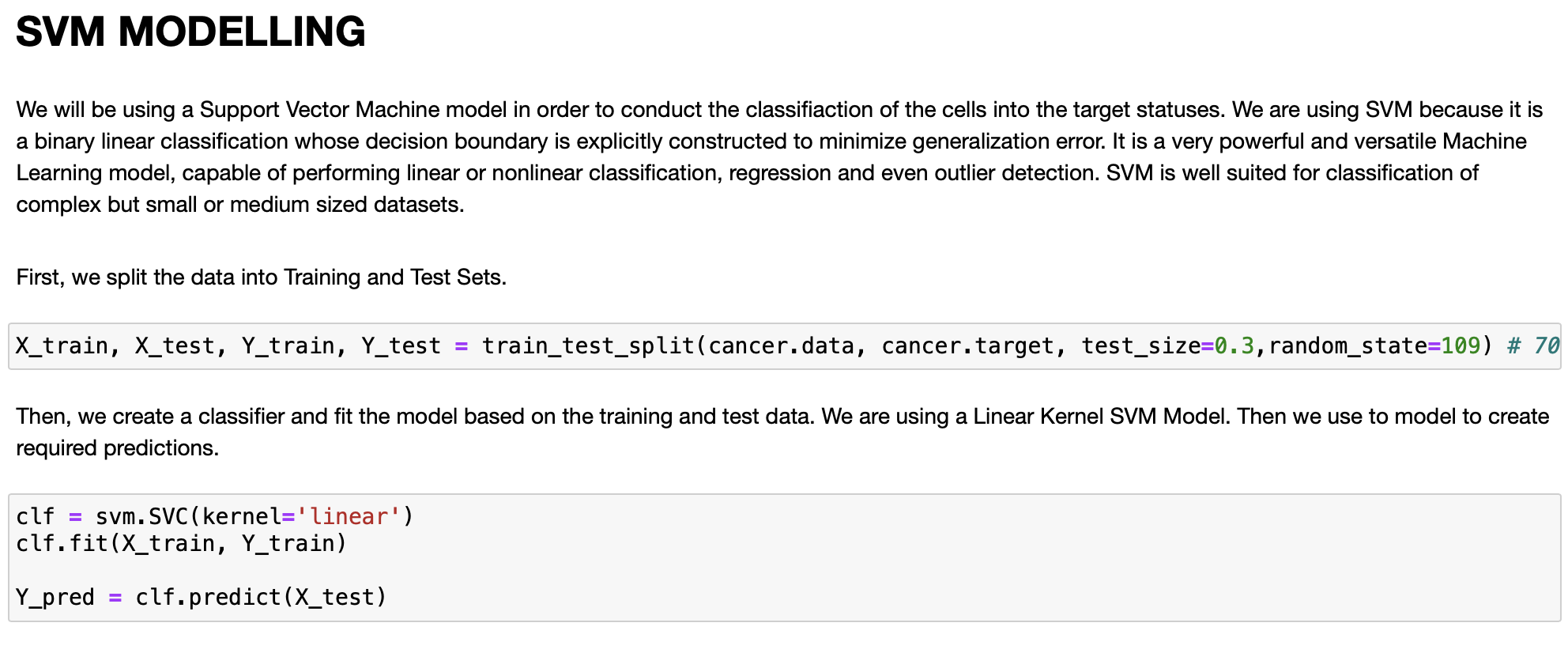
**4.1 Choosing Models**

We use the classification model for predicting the cell statuses . We will be using the Support vector machine for the classification of Target variables to be predicted. There are a number of advantages of using SVM modelling. These advantages are:-

1. SVM works relatively well when there is a clear margin of separation between classes.
2. SVM is more effective in high dimensional spaces.
3. SVM is effective in cases where the number of dimensions is greater than the number of samples.
4. SVM is relatively memory efficient.
5. The data-set suited our requirements, hence our choice was an SVM model.

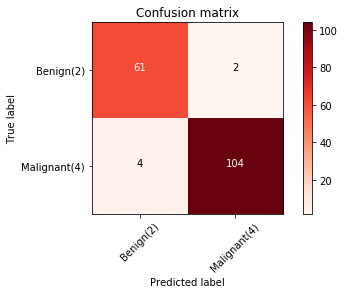
**4.2 Classification models**

The SVM machine is created through the Sci-kit Learn Library. We are using SVM because it is a binary linear classification whose decision boundary is explicitly constructed to minimize generalization error. It is a very powerful and versatile Machine Learning model, capable of performing linear or nonlinear classification, regression and even outlier detection. SVM is well suited for classification of complex but small or medium sized datasets.



**4.3 Model Evaluation**

* **The Confusion matrix gives us a Visual representation of the True and Predicted Labels. This depicts a high level of accuracy for our model.**
* **The Accuracy score for the model is 96.49%**
* **The Precision Score for the model is 98.11%**
* **The Recall Score for the model is 96.29%**
* **The F-Score for the model is 96.5%**
* **The Jaccard Score for the model is 94.54%, which sums up a very high level of accuracy for our model.**

****

**5. Conclusions**

In this study, we used a number of Data Science methods to create an SVM model for the classification cells as Benign or Malignant based on the cell attributes or features. We have received a 94.54% accuracy score on the Jaccard score and can be rest assured our model will work accurately.