STAD95 REPORT

Prognostic Analysis in Breast Cancer

Yuxi Cao

Le Yao Feng

Xinyi Guo

Jessica Hiu Wing Lau

University of Toronto Scarborough

Summer 2019

TABLE OF CONTENT

ABSTRACT

Using dataset from Kaggle, we will explore the nucleus features that contribute to different types of tumours in breast cancer and determine whether the patient has a benign or malignant tumour based on the given nucleus features. This analysis is to classify which variables are the most helpful when it comes to predicting benign or malignant cancer. We aim to observe the relationship between benign and malignant tumours, as well as other variables that will be discussed in the project.

INTRODUCTION

BACKGROUND & SIGNIFICANCE

Breast cancer is the most commonly diagnosed cancer among women. It can also be found in men; however, it's a very rare occurrence (Pace et al., 2016). In 2019, it's estimated that about 30% of newly diagnosed cancer among women are to be breast cancer, and it is the second leading cause of cancer death in women (Public Health Agency of Canada, 2017). Breast cancer is when cells in the breast tissue start to grow out of control, usually referred to as a tumour ("What Is Breast Cancer?", n.d). However, having a tumour does not necessarily mean it's cancerous. In general, there are 3 types of tumours, which are Benign, Premalignant and Malignant (Sinha, 2018). Benign tumours are not cancerous and they cannot spread, while Premalignant are usually not yet cancerous but seem to exploit properties of cancer; and malignant tumours are ones that are cancerous, as they keep growing, spreading and get worse over time (Sinha, 2018). The precise causes of breast cancer right now are unclear, but some factors that can increase the development of the disease are age, family history of breast cancer, high breast tissue density, previous abnormal breast biopsy, overweight, alcohol-consumption and hormone usage (Kamińska et al., 2015). Tests such as MRI, mammogram ultrasounds and biopsies are commonly used to determine the existence of breast cancer ("Breast Cancer Risk Factors and Prevention Methods",n.d.).

DATASET DEMOGRAPHIC

The dataset used in this analysis was created on November 1st, 1995, and contains 569 digitized nucleus images gathered on patient breast mass with their diagnosis status and ten real-valued features (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension for each cell nucleus). Furthermore, the mean, the largest (worst) value and standard error of each of the ten nuclear features are computed. These values are "numerically modelled such that larger values will typically indicate a higher likelihood of malignancy" (Nick Street, Wolberg, Mangasarian, 1993).

HYPOTHESIS

Our hypothesis is that individuals who exhibit high intensity in terms of the texture of the cell nucleus will be more likely to have malignant tumour than individuals who have low intensity.

METHODS

The data is analyzed using Rstudio statistical computing software. The logistic regression model will be computed using glm(). StepAIC() will be used to remove the less important variables. confusionMatrix() will be used to find accuracy for cross-validation of the models. Further analysis may be performed depending on the results of the data.

EXPLORATORY DATA ANALYSIS

This data analysis consists of ten nuclear features for each sample nucleus image. One breast mass may exist multiple tumours. Each of the following features is accompanied by a mean, largest / worst and standard error value of that feature. The mean values are the average value across all the values of the specific feature. The largest or "worst" values are the largest numerical value of the specific feature. The standard errors are the standard error from all the values of the specific feature.

Diagnosis

Tumour status of the imaged nucleus. 62.7% of the sample is diagnosed as benign while 37.3% are malignant. Malignant tumours have the ability to invade surrounding body tissues while benign tumours do not spread to surrounding tissues. One or more tumour(s) being malignant in the patient's body will result in the tumour status of that sample to be malignant.

Radius

The radius distance of each individual nucleus measured from the center to points on the perimeter.

Texture

The appearance of the cell nucleus computed using the variance of the intensity or brightness in gray scale that appear from the pixels of the nucleus image.

Perimeter

The size of the core tumour measured by the total average distance between adjacent points that make up the nuclear perimeter.

Area

Measurement of the nuclear area obtained by counting the number of pixels that reside on the interior boundaries of the cell nucleus and also the individual boundary points.

Smoothness

The difference between radius lengths and the average length of lines surrounding the cell nucleus.

Compactness

Measurement of the cell nuclei involving both the perimeter and area, it is calculated by $perimeter^2 / area$.

Concavity

Measurement of the extent of which the actual boundary of the nucleus lies on the inside of each chord, where the chords are between non-adjacent boundaries of the cell nucleus.

Concave points

The number of contour concavities that appear on the cell nucleus.

Symmetry

The length difference between perpendicular lines to the longest chord along the center of the nucleus and the cell boundary in both directions.

Fractal dimension

The coastline approximation of the fractal dimension of a cell. Computed by plotting values of the nucleus perimeter using a log scale and measuring the downward slope. Higher value of fractal dimension represents a less contour and a higher chance that the tumour is malignant.

DATA TIDYING

In order to access the dataset more conveniently, we first remove the unnecessary columns and keep the columns of the response variable (diagnosis) and the predictor variables (nuclear features). The unnecessary columns in the dataset are the first and last column which are the id and an extra column X.

Since each nuclear feature has the mean, standard error and largest / worst value, we then separate the dataset into three datasets according to mean, standard error and largest / worst value in figure 1.1, 1.2 and 1.3.

```
head(clinical_mean)

##   diagnosis radius_mean texture_mean perimeter_mean area_mean
## 1         M       17.99        10.38         122.80    1001.0
## 2         M       20.57        17.77         132.90    1326.0
## 3         M       19.69        21.25         130.00    1203.0
## 4         M       11.42        20.38          77.58     386.1
## 5         M       20.29        14.34         135.10    1297.0
## 6         M       12.45        15.70          82.57     477.1
##   smoothness_mean compactness_mean concavity_mean concave.points_mean
## 1         0.11840          0.27760         0.3001             0.14710
## 2         0.08474          0.07864         0.0869             0.07017
## 3         0.10960          0.15990         0.1974             0.12790
## 4         0.14250          0.28390         0.2414             0.10520
## 5         0.10030          0.13280         0.1980             0.10430
## 6         0.12780          0.17000         0.1578             0.08089
##   symmetry_mean fractal_dimension_mean
## 1        0.2419                0.07871
## 2        0.1812                0.05667
## 3        0.2069                0.05999
## 4        0.2597                0.09744
## 5        0.1809                0.05883
## 6        0.2087                0.07613
```

Figure 1.1 Mean Dataset

```
head(clinical_se)

##   diagnosis radius_se texture_se perimeter_se area_se smoothness_se
## 1         M    1.0950     0.9053        8.589  153.40      0.006399
## 2         M    0.5435     0.7339        3.398   74.08      0.005225
## 3         M    0.7456     0.7869        4.585   94.03      0.006150
## 4         M    0.4956     1.1560        3.445   27.23      0.009110
## 5         M    0.7572     0.7813        5.438   94.44      0.011490
## 6         M    0.3345     0.8902        2.217   27.19      0.007510
##   compactness_se concavity_se concave.points_se symmetry_se
## 1        0.04904      0.05373           0.01587     0.03003
## 2        0.01308      0.01860           0.01340     0.01389
## 3        0.04006      0.03832           0.02058     0.02250
## 4        0.07458      0.05661           0.01867     0.05963
## 5        0.02461      0.05688           0.01885     0.01756
## 6        0.03345      0.03672           0.01137     0.02165
##   fractal_dimension_se
## 1             0.006193
## 2             0.003532
## 3             0.004571
## 4             0.009208
## 5             0.005115
## 6             0.005082
```

Figure 1.2 Standard Error Dataset

```
head(clinical_worst)
```

```
##   diagnosis radius_worst texture_worst perimeter_worst area_worst
## 1         M        25.38         17.33          184.60     2019.0
## 2         M        24.99         23.41          158.80     1956.0
## 3         M        23.57         25.53          152.50     1709.0
## 4         M        14.91         26.50           98.87      567.7
## 5         M        22.54         16.67          152.20     1575.0
## 6         M        15.47         23.75          103.40      741.6
##   smoothness_worst compactness_worst concavity_worst concave.points_worst
## 1           0.1622            0.6656          0.7119               0.2654
## 2           0.1238            0.1866          0.2416               0.1860
## 3           0.1444            0.4245          0.4504               0.2430
## 4           0.2098            0.8663          0.6869               0.2575
## 5           0.1374            0.2050          0.4000               0.1625
## 6           0.1791            0.5249          0.5355               0.1741
##   symmetry_worst fractal_dimension_worst
## 1         0.4601                0.11890
## 2         0.2750                0.08902
## 3         0.3613                0.08758
## 4         0.6638                0.17300
## 5         0.2364                0.07678
## 6         0.3985                0.12440
```

Figure 1.3 Worst Dataset

DATA SUMMARY

In our given data set, there are 569 patients including 357 benign diagnosis (“B”) and 212 malignant diagnosis (“M”).

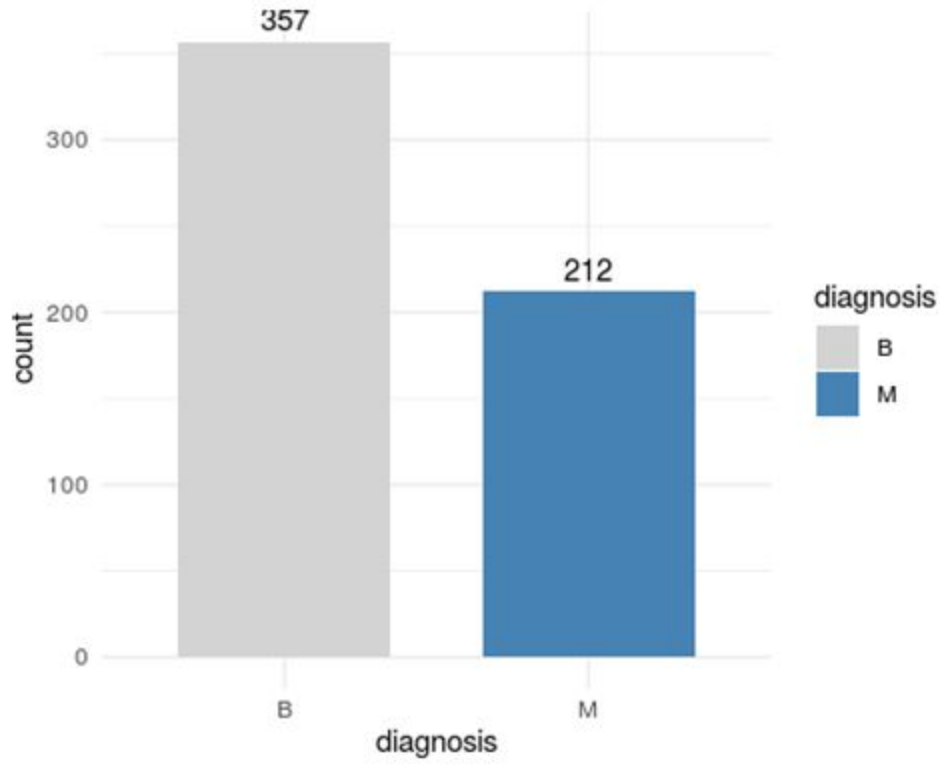

Figure 2.1 Bar chart for the count of patients with “B” (benign) and “M” (malignant) diagnosis results

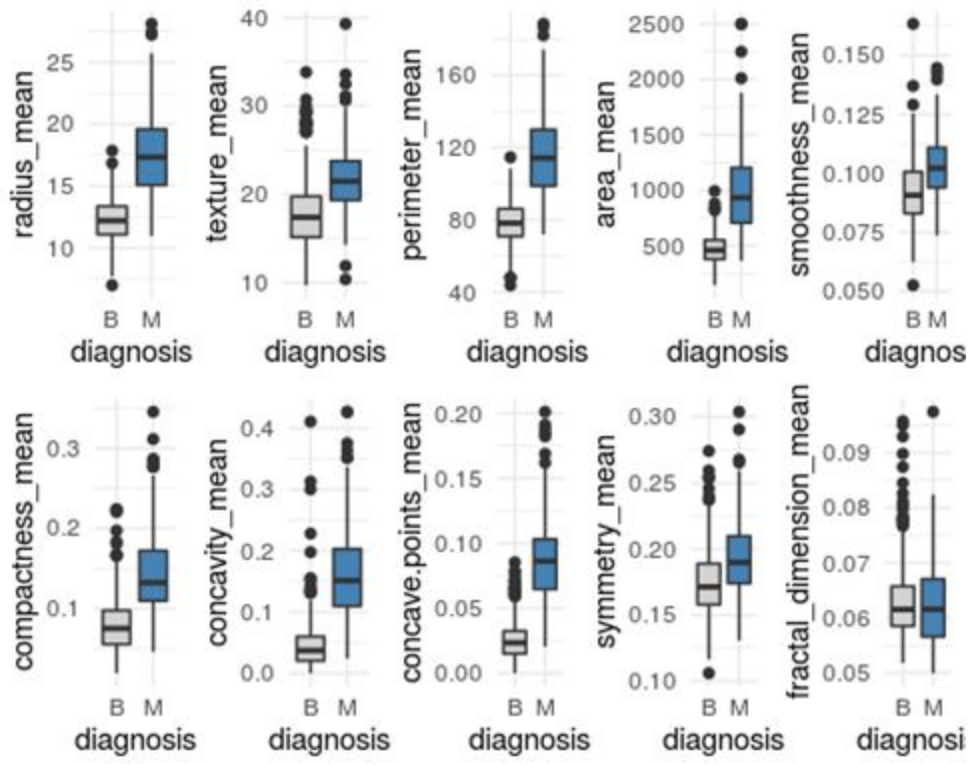

Figure 2.2 Boxplots of mean dataset comparing “B” (benign) and “M” (malignant) diagnosis results

The boxplots above compare the spread of each category of mean separated by the given two types of diagnosis result. Most diagnosis of “M” (malignant) has a greater value among different categories of mean in general than diagnosis of “B” (benign), only except for fractal dimension. We notice that there exists a positive relationship of opportunity getting malignant tumour if the value of spread of means are higher in radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, and symmetry.

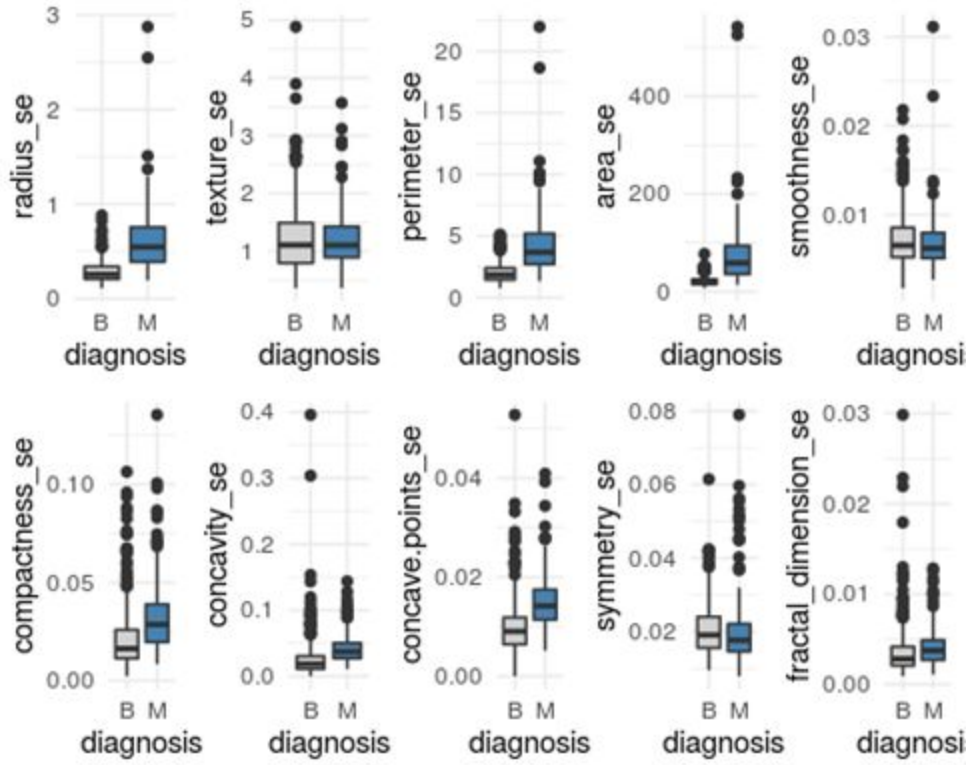

Figure 2.3 Boxplots of se dataset comparing "B" (benign) and "M" (malignant) diagnosis results

The set of boxplots above shows the spread of different category of standard error separated by two types of diagnosis result. The medians of standard error in radius, perimeter, area, compactness, concavity, and concave points with a diagnosis of "M" (malignant) are greater than those with diagnosis of "B" (benign). The medians in standard error of texture, smoothness, and symmetry are similar in both diagnosis results. Moreover, there are significant amount of outliers beyond upper whiskers as well, while there is nothing below the lower whiskers in these standard errors.

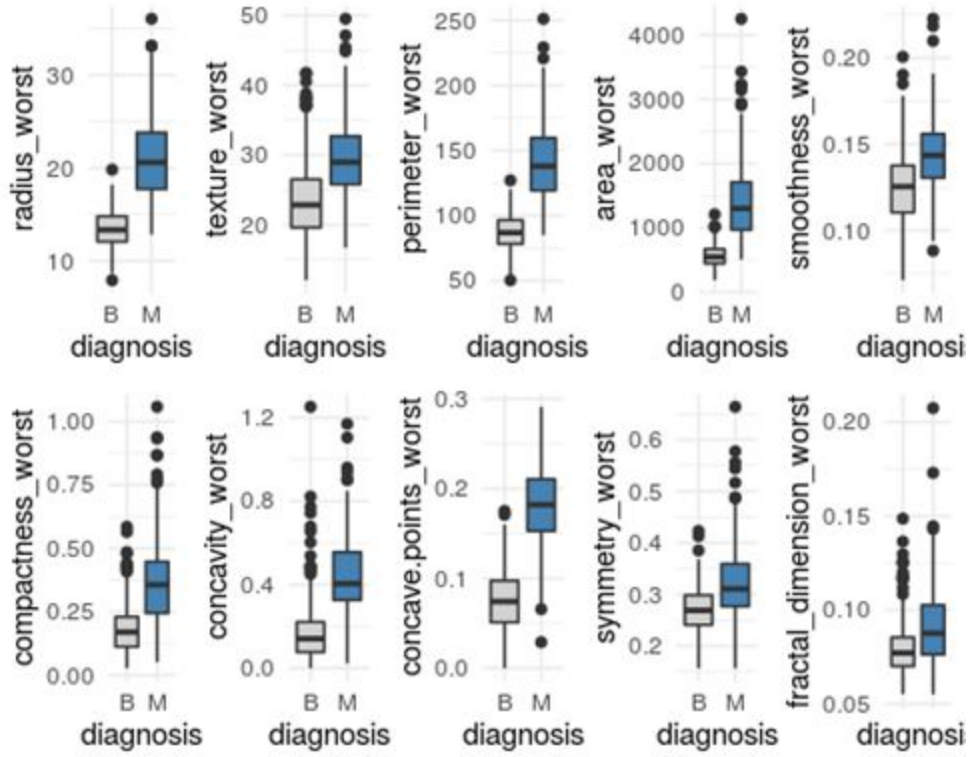

Figure 2.4 Boxplots of worst dataset comparing “B” (benign) and “M” (malignant) diagnosis results

The plot above displays the spread of the “worst” or largest (mean of the three largest values) of these features. Every value in this dataset with diagnosis of “B” is less than the ones of diagnosis of “M”, with respect to lower whiskers, Q1’s, medians, Q3’s, upper whiskers respectively.

MODEL

Model Selection & Validation

We first randomly selected 50% observations as the model-building dataset and the rest 50% observations as validation dataset. After that, we chose the important predictor variables by using the stepwise selection function. In this function, it dropped the variables which are less significant and contributed to a large AIC value to the model. The final selected model from the function will have the smallest AIC value (Akaike's Information Criterion). We repeated these steps on the mean, se and worst datasets and get the three final models as shown in Figure 3.1, 3.2, 3.3 respectively.

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + area_mean +
##     smoothness_mean + compactness_mean + concavity_mean +
concave.points_mean +
##     symmetry_mean + fractal_dimension_mean
##
## Final Model:
## diagnosis ~ texture_mean + area_mean + smoothness_mean + concavity_mean
```

Figure 3.1 Final model for mean dataset selected by stepwise selection function

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## diagnosis ~ radius_se + texture_se + perimeter_se + area_se +
##     smoothness_se + compactness_se + concavity_se + concave.points_se +
##     symmetry_se + fractal_dimension_se
##
## Final Model:
## diagnosis ~ radius_se + area_se + compactness_se + fractal_dimension_se
```

Figure 3.2 Final model for se dataset selected by stepwise selection function

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## diagnosis ~ radius_worst + texture_worst + perimeter_worst +
##     area_worst + smoothness_worst + compactness_worst + concavity_worst +
##     concave.points_worst + symmetry_worst + fractal_dimension_worst
##
## Final Model:
## diagnosis ~ texture_worst + perimeter_worst + smoothness_worst
```

Figure 3.3 Final model for worst dataset selected by stepwise selection function

We applied the confusionMatrix function to calculate all possible outcomes of the predictions based on the above selected final model from the model-building dataset and the validation dataset. It provided an accuracy rate based on the matrix with true positive, true negative, false positive and false negative as shown in Figure 4.1, 4.2, 4.3 respectively. We validate the models by checking the accuracy rate between the 50% model-building dataset and 50% validation dataset. The model with the highest accuracy rate become the best model and we will have some further analysis on the best model. Since worst dataset have the highest accuracy, we chose it as the best model.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   B   M
##          B 174  12
##          M   4  94
##
##                Accuracy : 0.9437
```

Figure 4.1 Result of Confusion Matrix for mean dataset

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   B   M
##          B 168  19
##          M  10  87
##
##                Accuracy : 0.8979
```

Figure 4.2 Result of Confusion Matrix for se dataset

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   B   M
##          B 176   5
##          M   2 101
##
##                Accuracy : 0.9754
```

Figure 4.3 Result of Confusion Matrix for worst dataset

From the summary of the best model selected in Figure 5, we can see that the estimated regression coefficients are all positive. This suggests texture, perimeter and smoothness features tend to be more likely to lend to malignant tumour than the other features. Texture_worst and perimeter_worst are almost the same in terms of the estimated coefficient. Meanwhile, smoothness_worst shows a very large estimated coefficient compared to the other two features. The effectiveness of smoothness is around 300 times more than the other two features.

```
summary(model_worst.final)

##
## Call:
## glm(formula = diagnosis ~ texture_worst + perimeter_worst +
smoothness_worst,
##     family = "binomial", data = clinical_worst)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2430  -0.0931  -0.0114   0.0050   3.9787
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -42.64473    5.24505  -8.130 4.28e-16 ***
## texture_worst      0.26032    0.05098   5.106 3.29e-07 ***
## perimeter_worst    0.23178    0.02973   7.796 6.40e-15 ***
## smoothness_worst  78.42053   14.11905   5.554 2.79e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.44  on 568  degrees of freedom
## Residual deviance: 103.61  on 565  degrees of freedom
## AIC: 111.61
##
## Number of Fisher Scoring iterations: 9
```

Figure 5 Summary of the best model

Multicollinearity

```
VIF

##    texture_worst  perimeter_worst smoothness_worst
##            1.334            1.642            1.373
```

Figure 6 VIF values for the remaining variables in the best model

From figure 6, there is no indication of serious multicollinearity because none of the VIF values is greater than 10, and the mean VIF is 1.450 which is not considerably much larger than 1.

Probability of the significant variables

The three probability plots between text_worst, perimeter_worst and smoothness_worst are displayed below. All of the plots show that the diagnosis of malignant tumours tend to be concentrated when the values of features are large. This characteristic is very obvious and clearly shown on Figure 7.2 while Figure 7.1 and 7.3 have few overlaps when the values of features are not extreme. Overall, the tendency on the three plots below is demonstrated clearly.

It's more likely for the diagnosis to be malignant when the values are large and it's more likely to have a benign diagnosis when the values are small.

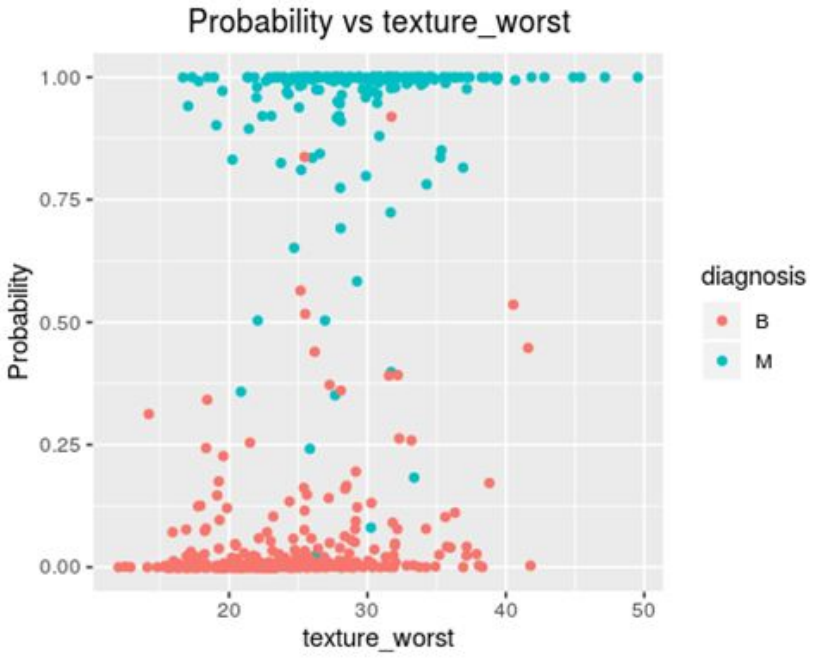

Figure 7.1 Probability vs texture_worst

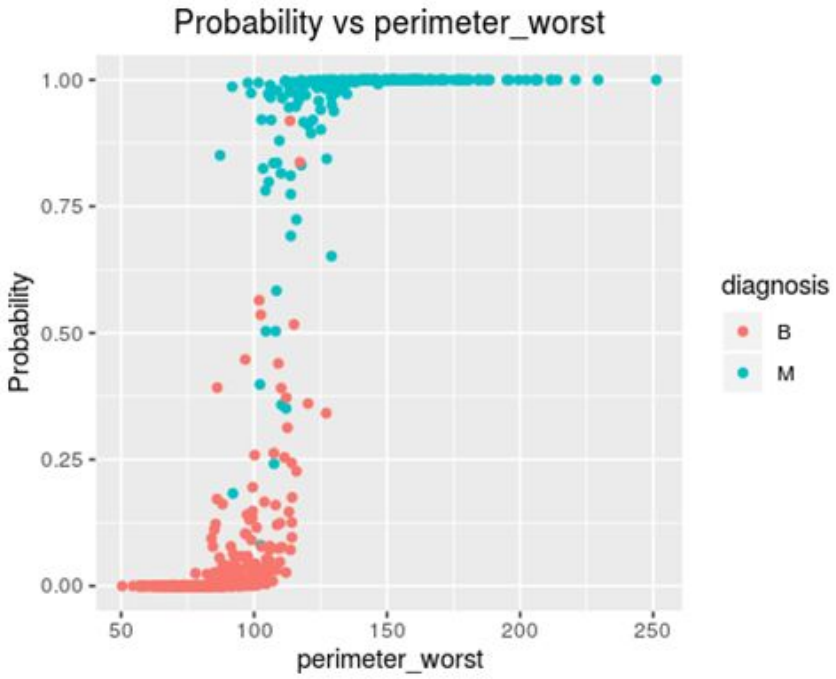

Figure 7.2 Probability vs perimeter_worst

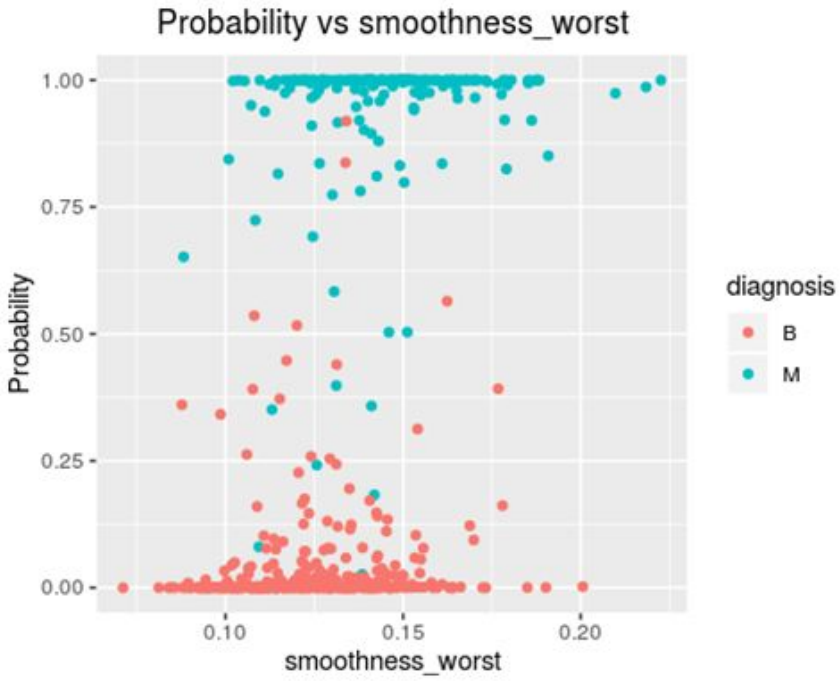

Figure 7.3 Probability vs smoothness_worst

DISCUSSION & CONCLUSION

The purpose of this study is to find out a model that can predict whether the breast tumour is benign or malignant based on the nucleus features we obtain from the examination results, in order to prognosticate the probability of getting breast cancer for a patient. In our dataset, six of the ten predictor variables were removed from the mean and se datasets while the seven predictor variables in the worst datasets were being omitted. The final model that we chose to focus on has highest accuracy during the cross-validation process. We believe that it can give us more insight on what nucleus features influence the diagnosis of benign or malignant the most. Our result shows that the texture_worst, perimeter_worst, and smoothness_worst are the three key factors that can have a generalized linear relationship to the diagnosis result among all the variables in our given dataset. The largest / "worst" value of the texture of the cell nucleus, the size of the core tumour, and the difference between radius lengths and the average length of lines surrounding the cell nucleus are the characteristics we need to focus on for prognostic analysis in breast cancer.

Limitations of our study does exist due to the restriction of our dataset. It only contains the features of the breast tumours on patients but it does not include other biological data and information of the patient. Some other studies also investigating on the prognostic analysis of breast cancer involved variables such as gene symbol, protein ID, gender, and survival duration in their dataset (Clinical Proteomic Tumor Analysis Consortium (NCI/NIH), 2016). These may also be influential factors to the diagnosis of breast tumour, however they are not recorded in the dataset we applied in this study.

The intention of our study is to support raising the awareness, early detection and screening of breast cancer. As the quality of diagnostic technologies for cancer is getting better, our project results can provide more accurate information on what to focus more when detecting the characteristic of the tumour, which can help the government institutes, health related companies and mostly especially the individuals. Providing all patients with a better diagnosing, allocating resources and treatment options will make sure they are able to access and afford the types of services in their standings in all kinds of needs.

REFERENCES

Abdelsamea, M. M., Mohamed, M. H., & Bamatraf, M. (2019). Automated Classification of Malignant and Benign Breast Cancer Lesions Using Neural Networks on Digitized Mammograms. *Cancer Informatics,18*, 117693511985757. doi:10.1177/1176935119857570

Breast Cancer. (n.d.). Retrieved from http://www.bccancer.bc.ca/health-info/types-of-cancer/breast-cancer

Breast Cancer Risk Factors and Prevention Methods. (n.d.). Retrieved from https://www.cancer.org/cancer/breast-cancer/risk-and-prevention.html

CDC - What Is Breast Cancer? (n.d.). Retrieved from https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm

Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). (2016). Breast Cancer Proteomes, Version 3. Retrieved August 1, 2019 from https://www.kaggle.com/piotrgrabo/breastcancerproteomes

Kamińska, M., Ciszewski, T., Łopacka-Szatan, K., Miotła, P., & Starosławska, E. (2015). Breast cancer risk factors. *Przeglad menopauzalny = Menopause review*, *14*(3), 196–202. doi:10.5114/pm.2015.54346

Pace, L. E., Dusengimana, J. V., Hategekimana, V., Habineza, H., Bigirimana, J. B., Tapela, N., Mpunga, T. (2016). Benign and Malignant Breast Disease at Rwanda's First Public Cancer Referral Center. *The Oncologist,21*(5), 571-575. doi:10.1634/theoncologist.2015-0388

Public Health Agency of Canada. (2017, October 23). Government of Canada. Retrieved from https://www.canada.ca/en/public-health/services/chronic-diseases/cancer/breast-cancer.html

Sinha, T. (2018). Tumors: Benign and Malignant. *Cancer Therapy & Oncology International Journal,10*(3). doi:10.19080/ctoij.2018.10.555790

Street, Nick & H. Wolberg, William & L Mangasarian, O. (1999). Nuclear Feature Extraction For Breast Tumor Diagnosis. Proc. Soc. Photo-Opt. Inst. Eng.. 1993. 10.1117/12.148698.

UCI Machine Learning. (2016, September). Breast Cancer Wisconsin (Diagnostic) Data Set, Version 2. Retrieved July 28, 2019 from https://www.kaggle.com/uciml/breast-cancer-wisconsin-data