STAT 406: Project Report
Tumour Severity Classifier

Stephanie Chan 41113101 stephaniechan219@gmail.com
Bang Chi Duong 97632939 bangchi.duong.20193@outlook.com
Belrouz Salehipour 33186123 salehipour.behrouz@gmail.com

1 Introduction

The most common and effective way to screen for breast cancer today is mammography. If we can predict with high accuracy whether a tumour is malignant, we can create an aid for doctors to use when deciding whether to recommend a biopsy to patients or not.

The dataset we chose to analyze is a mammographic dataset from the UCI Machine Learning Repository. It contains breast cancer patient data consisting of a BI-RADS assessment, the age of the patient, three other BI-RADS attributes, and the true classification of 516 benign and 445 malignant masses. BI-RADS is short for Breast Imaging, Reporting and Data System. It has a set of attributes for breast tissue masses, and these attributes include mass shape, mass margin, and mass density.

The goal of this project is to analyze which statistical learning methods would be effective for classifying the severity of a mammographic mass lesion from the age of the patient and their BI-RADS attributes. We have the following predictors that are available for us to use in our experiment: age of patient, shape, margin, and density of the tumour.

2 Data Description

The mammographic data used in our project was compiled by Dr. Matthias Elter and Dr. Rdiger Schulz-Wendtland and obtained from the UCI Machine Learning Repository's Mammographic Mass Data Set.

Below shows the attribute information:
1. BI-RADS assessment: 1 to 5 (ordinal, not a predictor)
2. Age: Patient's age in years (integer)
3. Shape (mass shape): Round, oval, lobular, and irregular (nominal)
4. Margin (mass margin): Circumscribed, microlobulated, obscured, ill-defined, and spiculated (nominal)
5. Density (mass density): High, iso, low, and fat-containing (ordinal)
6. Severity: Benign or malignant (binomial, response)

Our challenges come from having missing values in our data. Precisely, the missing values include 2 values from BI-RADS assessment, 5 values from age, 31 values from shape, 48 values from margin, and 76 values from density. There were no missing values from the column severity. Since the data set only has four potential predictors age, shape, margin, and density, we can look at the correlation or association between the response (severity) and each of the predictors (i.e. marginally).

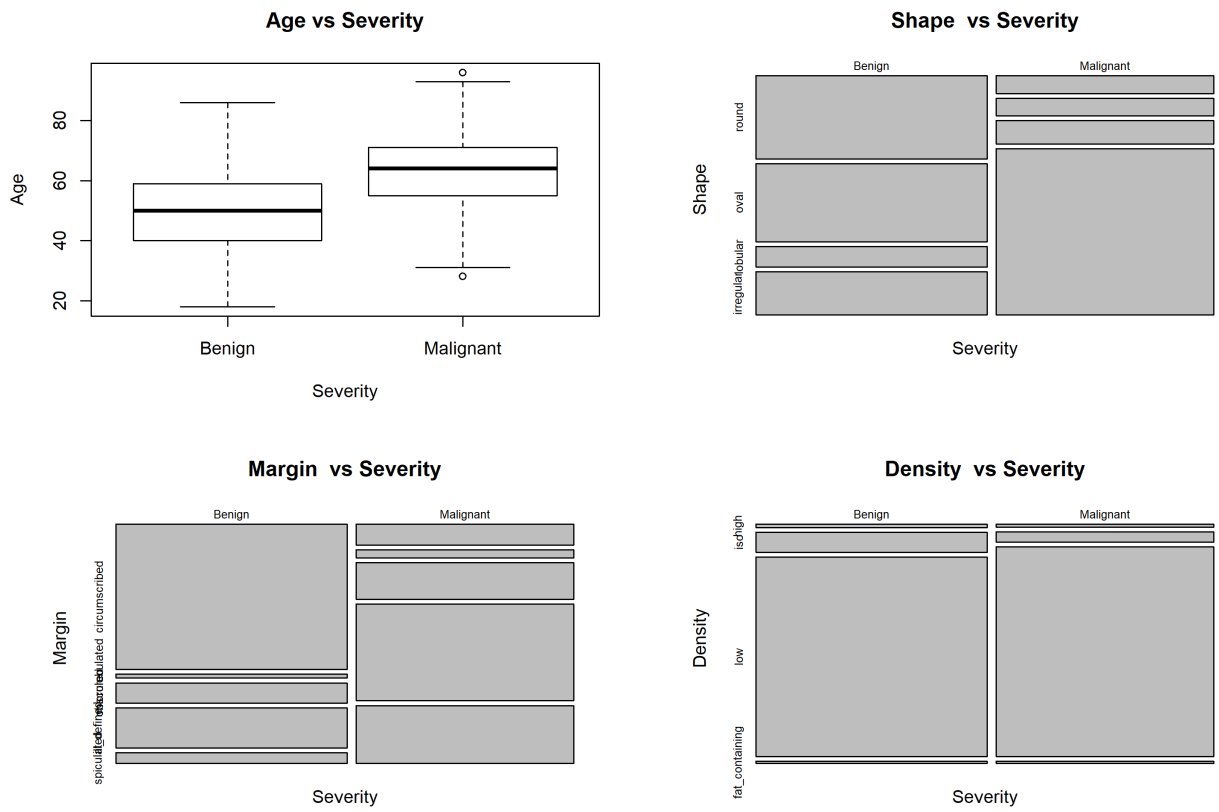

Figure 2.1: Relationships of Variables are shown. The plot suggests that except for density, all the other variables seem to be useful predictors for the classification of severity.

3 Methodology

There are 12 methods that were used in this project, which could be broadly classified into parametric (methods 1-5) and non-parametric (methods 6-12).

1. Logistic Regression
2. Linear Discriminant Analysis (LDA)
3. Quadratic Discriminant Analysis (QDA)
4. Support Vector Machine (SVM) (with linear kernel)
5. Support Vector Machine (SVM) (with radial basis kernel)
6. Random Forest
7. Adaptive Boosting (AdaBoost) (1 split - Decision Stump)
8. Adaptive Boosting (AdaBoost) (2 splits)
9. Adaptive Boosting (AdaBoost) (3 splits)
10. 1-Nearest Neighbour (1-NN)
11. 3-Nearest Neighbours (3-NN)
12. 5-Nearest Neighbours (5-NN)

4 Analysis

There are several changes that we made to preprocess the data. First, we removed the "BI-RADS" assessment from our data because according to the original description of the dataset, it is a non-predictor. Age was treated as a numeric variable, whilst the others were treated as categorical variables. Finally, we removed all rows containing missing data from our dataset. We ended up from 961 observations with 6 variables to 831 observations with 5 variables, so we still have a reasonable amount of data left to analyze.

The *Age vs Severity* boxplot from Figure 2.1 in the Data Description Section shows that the age's median for benign class is not within the age interquartile range for malignant class and vice versa, therefore we kept age as a predictor. The *Shape vs Severity* and *Margin vs Severity* mosaic plots suggest that there seems to be a difference in levels of severity based on both shape and margin, so they were kept as predictors.

From our mosaic plots in Figure 2.1, it seems density does not have an effect on severity. Thus, we have decided to run all of our methods with and without the density feature, in order to check for differences in misclassification error rate. We expect that including density would increase the error rate of the classifiers.

For methods 1-9 on the list, we ran a 5-fold cross-validation 50 times, and for the three K-Nearest Neighbour methods we ran 20 times due to time constraints, and averaged the resulting errors. We ran this entire procedure twice, one procedure included density as a predictor and the other did not. In general, we expect to see better performance of the non-parametric methods over parametric ones. To avoid the "rank deficiency" problem due to multicollinearity from fitting QDA, a small amount of noise was added to the age variable when fitting QDA. K-Nearest Neighbours were considered with Gower's coefficient (Gower, 1971), expressed as a dissimilarity, instead of the usual Euclidean distance because we have a mixture of numerical and categorical variables.

We also utilized SVM with linear and radial basis kernels to further optimize the true separating hyperplane between the two classes of severity. SVMs were developed by Cortes and Vapnik (1995) for binary classification. The method maximizes the margin between the classes' closest points, the middle of the margin is the optimal separating hyperplane, and the points on the two boundaries of the margin are called support vectors. It also deals with overlapping classes by weighing down the influence of the data points that are on the "wrong" side of the discriminant margin. When a linear separator cannot be found, the data points are projected into a higher-dimensional space (via kernel techniques), where they become linearly separable. This can be solved as a quadratic optimization problem.

Tree bagging was not included because random forest is already an improved version of bagging, where both variance and bias are reduced. Variance is reduced because the trees in random forest are more independent due to random draws of predictors. Bias is reduced because a very large number of predictors can be considered; however, in this particular case there is not much of a difference since we only have a maximum of 4 predictors.

5 Results and Conclusion

We ran the tests using the methodology we described in the previous section and recorded our results into Table 4.1 shown below. Let us call the case without density as case 1, and with density as case 2.

Method	Error (without Density)	Error (with Density)
Logistic Regression	0.1917	0.1963
LDA	0.1938	0.1974
QDA	0.2024	0.2218
SVM (linear kernel)	0.2066	0.2067
SVM (radial basis kernel)	0.2227	0.2189
Random Forest	0.1935	0.2008
AdaBoost (1 split)	0.1911	0.1930
AdaBoost (2 splits)	0.1926	0.1924
AdaBoost (3 splits)	0.1952	0.1960
1-NN	0.3016	0.3212
3-NN	0.2545	0.2691
5-NN	0.2068	0.2227

Table 4.1: Misclassification error rate from 50 runs of 5-fold cross-validation for the first 9 methods, and 20 runs of 5-fold cross-validation for 3 of the K-NN methods, with and without density as a predictor.

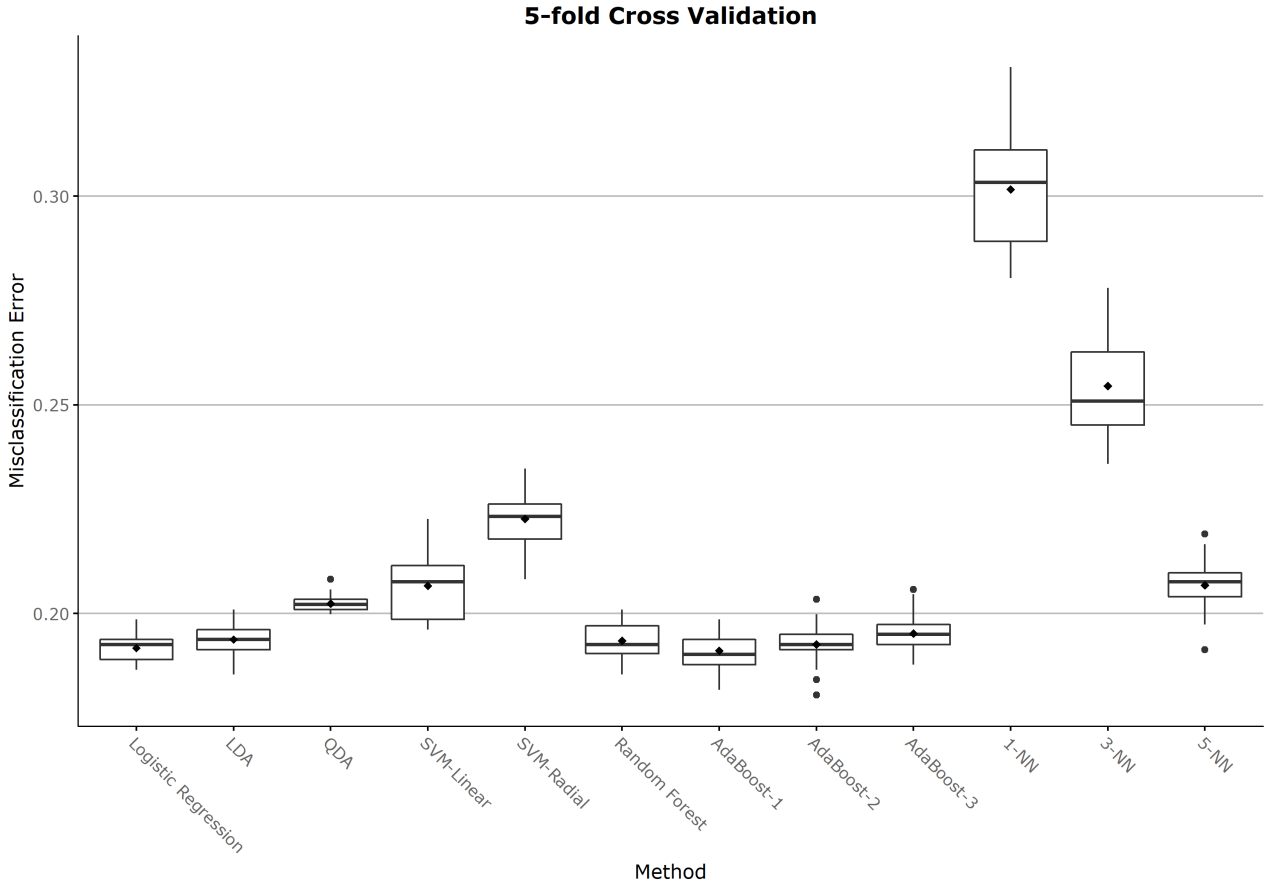

Figure 4.1: Boxplot of 5-fold cross-validation with 12 statistical learning methods, excluding the density feature. The black diamond shapes represent the averages of misclassification rate for each method.

Within each case of whether density was included or not, among the parametric methods, logistic regression was the best classifier, LDA came very close at second place, QDA performed worse than LDA, and SVM with linear kernel performed better than SVM with radial basis kernel, which was the poorest classifier. These results point towards one conclusion that the separating hyperplane between the two classes of severity is very likely to be linear.

Among the non-parametric methods, AdaBoost with a decision stump was the best classifier for case 1.

Again, this suggests a linear separating hyperplane. For case 2, AdaBoost with a decision stump came second place, but was very close to the first place, which was AdaBoost with 2 splits (difference in error = 0.006). This behavior could be due to the fact that we added information (by including density) and increased richness/depth of trees (2 splits). AdaBoost with 3 splits (and AdaBoost with 2 splits in case 1) suffered from overfitting since we only have 3 predictors in case 1, and 4 in case 2. For the K-NN methods, increasing K from 1 to 3 vastly increased the performance (error reduced from 0.30 to 0.21), and decreased variance (i.e. more stabilized according to Figure 4.1 with smaller interquatile range). Though with 3-NN, the error rate was still worse than other non-parametric methods. Overall, AdaBoost with a decision stump seems to perform best among all the classifiers. There are three apparent reasons: one, it is a non-parametric method, hence it is more flexible than parametric ones and thus has lower bias; two, it avoids overfitting; and three, it supports linear separation.

Comparing case 1 and case 2, Table 4.1 shows that including density reduced the performance of all classifiers except the case of SVM with radial basis kernel. This could be the case because including density as a predictor might have changed the linear separation to a non-linear separation, increasing the misclassification rate of those classifiers that support linear separating hyperplane, as aforementioned in the Analysis Section.

Our results generally show that the parametric models which rely on certain distributional assumptions performed worse than the non-parametric methods. For example, logistic regression assumes that severity variable follows Bernoulli distribution with the logit link function from the generalized linear models. For LDA, the distribution of the predictors given a class of severity is assumed to be multivariate Gaussian, with a constant covariance matrix for every class of severity. QDA has a similar assumption to LDA but each class has a different covariance matrix.

The models that performed well with misclassification rate under 0.20 were logistic regression, LDA, random forest, and adaptive boosting with all three splits. By not removing the density feature, more noise ended up in the data, which showed up as higher error rates. Our results showed that without density, almost all models we tested had lower error rates. Therefore, we should not use density as a predictor.

6 References

Cortes, C. & Vapnik, V. (1995). Support-vector network. Machine Learning, 20, 125.

Elter, Matthias and Rdiger Prof. Dr. Schulz-Wendtland. Mammographic Mass Data Set. 29 10 2007. Document. 30 10 2017. http://archive.ics.uci.edu/ml/datasets/mammographic+mass.

Gower, J. C. (1971) A general coefficient of similarity and some of its properties, Biometrics 27, 857–874.

James, G; Witten, D; Hastie, T; Tibshirani, R. (2015). An Introduction to Statistical Learning with Applications in R. Retrieved from www-bcf.usc.edu/gareth/ISL/.

Kaufman, L. and Rousseeuw, P.J. (1990) Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Struyf, A., Hubert, M. and Rousseeuw, P.J. (1997) Integrating Robust Clustering Techniques in S-PLUS, Computational Statistics and Data Analysis 26, 17–37.