Breast Cancer Diagnosis Using Machine Learning

Riya Senthil

Guided By: Dr. Badrinath Kottimukkular, Assistant Professor, George Washington University, Washington DC.

Objective

Using digitally coded histopathological image data, predict the samples that has breast cancer (malignant) cells. Identify which features in the images are highly predictive of malignancy.

Data Description

The breast cancer diagnosis data is a publicly available dataset [1]. There are 699 observations, 9 features and diagnosis indicating the breast cancer cells are malignant or benign.

After a Fine Needle Aspiration (FNA) biopsy, microscopic examination produces histopathological images of the cells. These cells are then coded into various characteristics and the breast cancer dataset is generated.

The nine features are all coded between 1 to 10.

Features are:

- clump thickness – malignant sample has thick grouping of cancer cells in multiple layers.
- uniformity of cell size – represents metastasis to lymph nodes.
- uniformity of cell shapes - cancerous cells have varying sizes.
- marginal adhesion - suggests loss of adhesion which is a sign of malignancy.
- single epithelial cell size (SECS) – Larger SECS may indicate a malignant cell.
- bare nuclei – benign cells has a smaller number of nuclei without cytoplasm. Sixteen missing values.
- bland chromatin - In cancer cells the chromatin tends to have coarse texture.
- normal nucleoli - generally very small in benign cells.
- mitoses – cell division process. Generally higher counts in malignant cells.

	Clump_Thickness	Cell_Size_Uniformity	Cell_Shape_Uniformity	Marginal_Adhesion	Single_Epithelial_Cell_Size	Bare_Nuclei	Bland_Chromatin	Normal_Nucleoli	Mitoses	Diagnosis
count	699	699	699	699	699	683	699	699	699	699
mean	4.42	3.13	3.21	2.81	3.22	3.54	3.44	2.87	1.59	0.34
std	2.82	3.05	2.97	2.86	2.21	3.64	2.44	3.05	1.72	0.48
median	4	1	1	1	2	1	3	1	1	0

Table 1. Descriptive feature statistics. Bare Nuclei has 683 non-missing values. 34% of samples have malignant diagnosis.

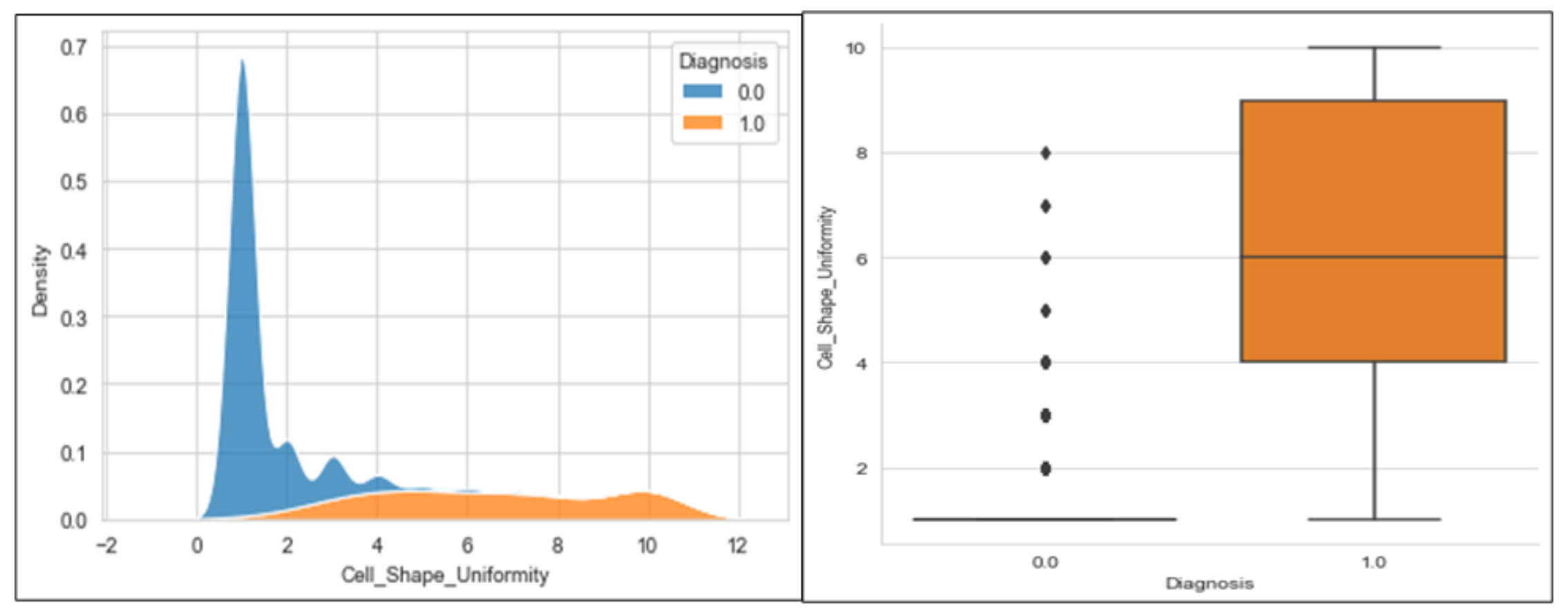

Fig. 1 Left chart shows Clump Shape Uniformity density distributions of benign (blue, diagnosis = 0) and malignant cells (orange, diagnosis = 1). Right chart shows same distributions in a box plot. There is a strong differentiation in values of Clump Shape Uniformity between benign and malignant cells.

Feature Exploration

All nine features are converted to float. The target variable Diagnosis is converted to 0 and 1 representing benign and malignant sample, respectively. Python packages such as *numpy* and *pandas* are used for computations. Packages *matplotlib* and *seaborn* are used for plotting and data visualization.

Table 1 has descriptive statistics. There are 699 observations. Bare Nuclei has 16 missing values. Mitoses, Normal Nuclei and Marginal Adhesion have most records with low values (avg: 1.6 to 2.8). All other features also have low values in general (avg: 3.1 to 4.4). In most cases, standard deviations are similar or less than mean. About 34% of observations are malignant samples and 66% are benign samples. Data is a bit unbalanced with more benign samples.

Figure 1 shows that cell shape uniformity value is low (i.e., uniformly shaped cells) in benign cells. Cancer cells have high cell shape uniformity value (irregular cell shapes) and is well differentiated from normal cells. Thus Cell Shape Uniformity is a good predictor of malignancy. Many features have similar strong differentiation between benign and malignant cells. Example: Cell Size Uniformity, Bare Nuclei, Normal Nucleoli, Single Epithelial Cell Size.

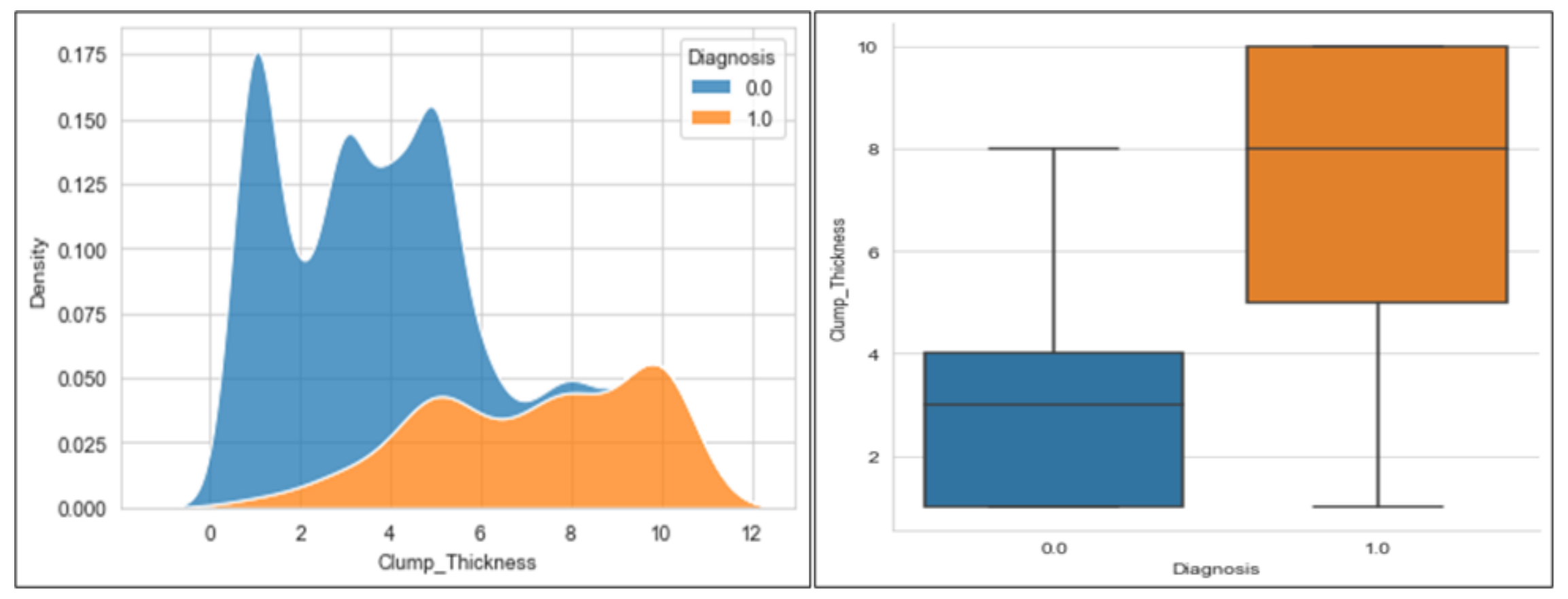

Fig. 2 Left chart shows Clump Thickness density distributions of benign (blue, diagnosis = 0) and malignant cells (orange, diagnosis = 1). Right chart shows same distribution in a box plot. There is some overlap and a medium level of differentiation in values of Clump Thickness between benign and malignant cells. Distributions deviates fairly from normal distribution.

Figure 2 shows clump thickness (grouping of cancer cells into multiple layers) is higher in malignant cells and is somewhat differentiated from benign cells. However, there is a fair overlap in clump thickness of benign and malignant cells. They are not normally distributed and has bands and tails. Few other variables show similar overlaps in their values. Example: Mitoses, Marginal Adhesion.

Many of the variables have moderate to high correlations among themselves. Highly correlated features are Cell Size, Shape and Texture uniformity (0.76 to 0.91). These uniformity values along with bare nuclei count have highest correlation (0.82) with benign vs malignant diagnosis. Mitoses has relatively lower correlation with other variables (0.34 to 0.48).

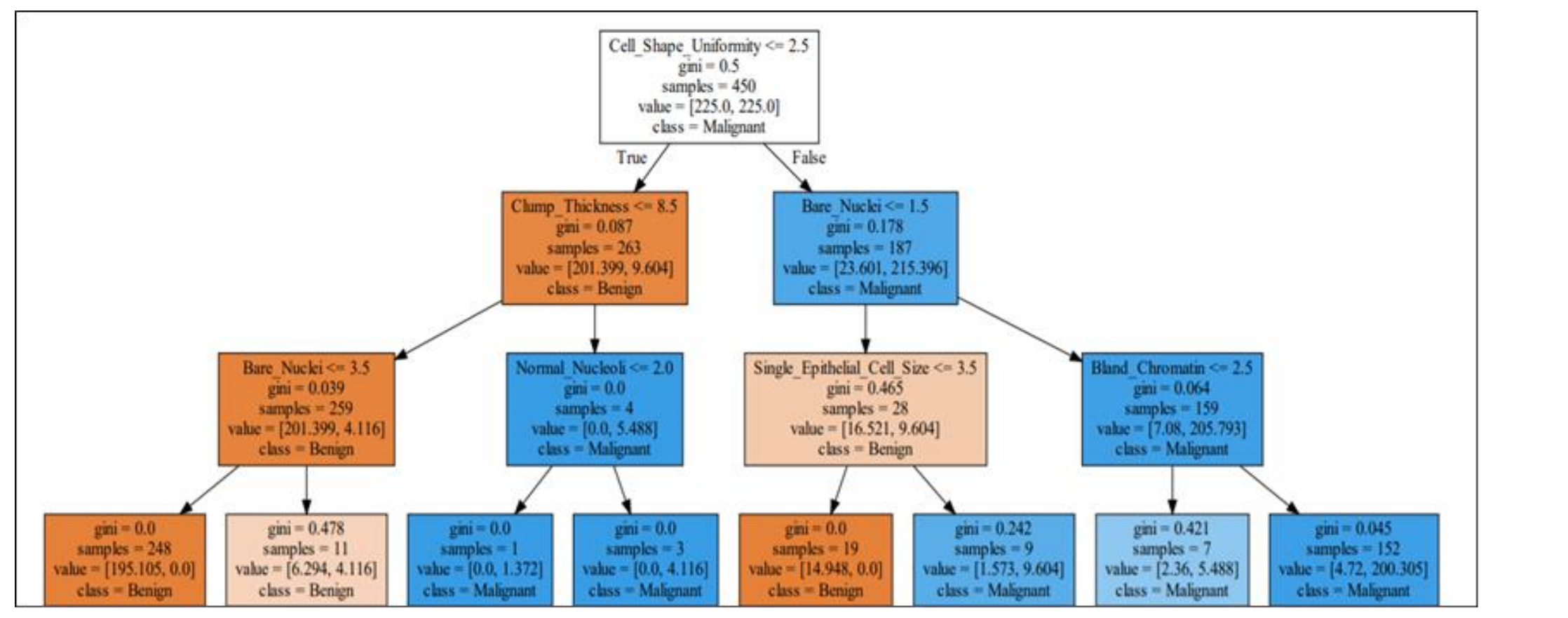

Fig. 3. Decision tree with max depth of three and using diagnosis as target. Orange nodes denote benign cell classifications, and the blue nodes denote malignant cell classification. Most important features to classify benign vs malignant cells are Cell Shape Uniformity, Clump Thickness, Bare Nuclei and Single Epithelial Cell Size.

Models and Results

Data is split into 66% for training [450 observations] and 34% for testing [283 observations]. Models are built on training test dataset and their performances are evaluated using test data. Least error in classifying the malignant cells is emphasized. Training data is balanced (i.e., malignant samples are repeated) so that the data has equal representation of benign and malignant cells to be modelled. Data balancing helps to improve model accuracy.

Decision trees classify target variable. The algorithm first goes through all the features and chooses a feature and boundary point that has maximum information to classify the target variable. Using this boundary, it subsets the data into two nodes. The algorithm is then iterated for each of the child node until leaf nodes with maximum classification accuracy could be obtained. Tree depth is adjusted to not overfit the data. Python packages sci-kit learn (*sklearn*) and *graphviz* were used in building decision trees. Figure 3 shows a tree of depth 3 with diagnoses as target variable and all nine features as inputs. Most important features to classify benign vs malignant cells are **Cell Shape Uniformity, Clump Thickness, Bare Nuclei and Single Epithelial Cell Size.**

Logistic regression classifies the binary outcomes (benign vs malignant) using regression methods that seek to fit a linear equation that minimizes the error in classification. A logit transformation is done to the output variable so that the output is the probability of a given sample to have a value of 1 (i.e., malignant). Odds ratio, that represents odds of a given feature to be predictive of malignancy, is used to choose most influential features to predict malignancy. Python package *statsmodels* was used for logistic regression models. Table 2 shows results from logistic regression model where some of the non-significant variables were removed. Higher odds ratios are seen *(1.5 to 1.8)* for ***Clump Thickness, Cell Shape Uniformity, Bare Nuclei and Bland Chromatin in predicting malignancy.***

	Coefficient	Standard Error [Std. Deviation of the coefficient]	Confidence Level	Odds Ratio
Intercept	-9.3061	1.079	100%	0.00
Clump_Thickness	0.5836	0.145	100%	1.79
Cell_Shape_Uniformity	0.405	0.184	97%	1.50
Marginal_Adhesion	0.2955	0.165	93%	1.34
Bare_Nuclei	0.4059	0.103	100%	1.50
Bland_Chromatin	0.3837	0.172	97%	1.47
Normal_Nucleoli	0.2109	0.108	95%	1.23
Mitoses	0.5369	0.36	87%	1.71

Table 2: Results from logistic regression with select features. Except for Mitoses and Marginal Adhesion all other features are statistically significant at 95% confidence level. Clump Thickness, Cell Shape Uniformity, Bare Nuclei and Bland Chromatin have higher odds (1.5 to 1.8) in predicting malignancy.

Confusion Matrix. (Cells have count of records)		Predicted		Accuracy = [TP + TN] / [TP + TN + FP + FN]
		Negative [=0] (benign)	Positive [=1] (malignant)	Precision = [TP] / [TP + FP]
Actual	Negative [=0] (benign)	True Negative [TN]	False Positive [FP]	Recall = [TP] / [TP + FN]
	Positive [=1] (malignant)	False Negative [FN]	True Positive [TP]	F1 = [2*Precision*Recall] / [Precision + Recall]

Fig. 4: A confusion matrix is used to assess the model performance. It compares actual vs predicted classification record counts in test dataset. Various metrics such as accuracy, precision, recall and F1 are used to assess the model performance. Metric values range between 0 (low performance) to 1 (perfect prediction). Here we focus more on Recall as the cost of predicting someone has no cancer when they actually have cancer is high.

Best Model and Results

A confusion matrix (Fig. 4) produces various classification accuracies. **Accuracy** defines overall predictive accuracy as % of samples that were classified correctly as either benign or malignant. **Recall** represents % of actual malignant samples that were correctly predicted. **Precision** represents % of predicted malignant samples that are actually malignant. **F1 score** is a weighted average of Recall and Precision. **Best models are chosen by analyzing the highest recall in predicting malignancy using test dataset.**

Table 3 shows performance metrics for various model configurations. **Logistic regression with select variable (Table 2) is the best predictive model with a recall of 97.8% and accuracy of 97.4%. Decision Tree model with balanced data and a max depth of 3 (Fig. 4) has almost similar recall of 97.3% and is straight forward to interpret.** We use this decision tree model to explain the results and are discussed in conclusion section.

Model Description		Recall	F1	Precision	Accuracy
Decision Trees	Unbalanced data. Unrestricted tree depth (=9)	0.93	0.93	0.94	0.96
	Balanced data. Unrestricted tree depth (=11)	0.91	0.925	0.94	0.95
	Balanced data with depth 3	**0.973**	0.948	0.924	0.965
	Balanced data with depth 5	0.933	0.933	0.933	0.957
Logistic Regression	All Features	0.978	0.967	0.957	0.974
	Non significant features removed	**0.978**	0.967	0.957	0.974

Table 3: Classification metrics from confusion matrix for various model configurations. Logistic regressions with statistically non-significant features removed has the highest recall at 97.8% and a precision of 97.4% . This is the best predictive model. It is closely followed by random forest model and decision tree with max depth of 3, with recalls at 97.3% and predictive accuracy of 96.5%. Decision tree models are more straight forward to interpret.

Conclusions

The breast cancer dataset [1,2] with 9 digitally coded features from histopathological images of the breast cancer cell samples is used to predict if the given sample is benign or malignant. This data was split into 66% for training and 34% for testing. Classification models such as decision trees and logistic regression were applied to training data with diagnosis as target variable. Logistic regression with select variables with highest recall of 97.8% is the best predictive model followed closely by a decision tree model with short tree depth and a recall of 97.3%. Based on this easily interpretable decision tree model following was observed:

A given sample is malignant when:

- Cell Shape is more uniform (Cell_Shape_Uniformity <= 2.5) BUT the Clump Thickness is very high (Clump_Thickness > 8.5)
- Cell shape is not so uniform (Cell_Shape_Uniformity > 2.5) AND there are more bare nuclei (Bare_Nuclei > 1.5) AND a coarse texture of chromatin (Bland_Chromatin > 2.5). Even if chromatin has a fairly uniform texture, there is a high chance of malignancy.
- Cell shape is not so uniform (Cell_Shape_Uniformity > 2.5) AND there are less bare nuclei (Bare_Nuclei <= 1.5) BUT the single epithelial cell size is large (Single_Epithelial_Cell_Size > 3.5). In this case, there is a fairly good chance of malignancy.

References

[1] Data Source: http://pages.cs.wisc.edu/~olvi/uwmp/cancer.html
Dr. WIlliam H. Wolberg , University of Wisconsin Hospitals, Madison, Wisconsin, USA
Donor: Olvi Mangasarian (mangasarian@cs.wisc.edu). Received by David W. Aha

[2] W.H. Wolberg, W.N. Street, and O.L. Mangasarian.
Breast Cytology Diagnosis via Digital Image Analysis. Analytical and Quantitative Cytology and Histology, Vol. 15 No. 6, pages 396-404, December 1993.