Data Science - **Data Modelling Showcase**

By Brandon Tran

Table of Contents

Executive Summary

Breast cancer is the second largest cause of cancer-related death among women. Methods of early detection can dramatically improve the probability of survival as targeted cures during this stage are most effective. This paper will observe several clinical attributes of 116 participates to hypothesize and develop machine learning models to predict the presence of breast cancer as early as possible. The focus of this prediction will include two classification models, which are K-Nearest Neighbours (KNN) and Decision Tree, to determine at-risk participants. The report concludes that although the created models have a good level of accuracy in predicting the presence of breast cancer, they are not quite at a sufficient level given the severity of a misdiagnosis. It is recommended that more data is collected to minimise the error in subsequent models.

Introduction

Cells in the human body are in constantly changing states depending on the needs of the body. Cells grow if the body demands it or die if they are no longer required. The behaviour of cancer cells does not follow this tradition. Instead, they keep growing and dividing to eventually form a tumour (Cancer Research UK, 2019).

Breast cancer is a type of cancer most commonly diagnosed among women and is ranked the second largest in the number of cancer-related deaths in women. Approximately over 200,000 breast cancer diagnosis are expected in the United States in the year 2005. Early detection, through procedures such as breast screening, can significantly increase the chance of survival (Training.seer.cancer.gov, 2019).

This report will discuss the development of models based on compounds found in the bloodstream to predict the presence or absence of breast cancer in a patient.

Methodology

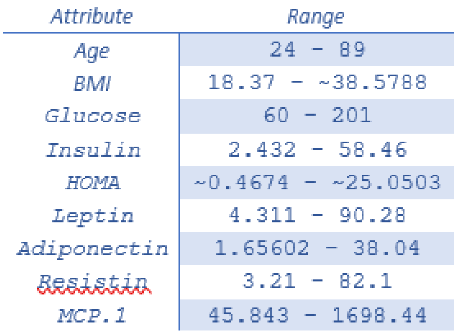
Data Retrieval

The Breast Cancer Coimbra dataset was obtained from the UCI Machine Learning Repository. The set contains information of 116 participants and a measure of 10 quantitative attributes including Age, Body Mass Index (BMI), Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1 and classification; where 1 is labelled as healthy controls and 2 is patients. The measure of these attributes was gathered through routine blood analysis.

Data Preparation

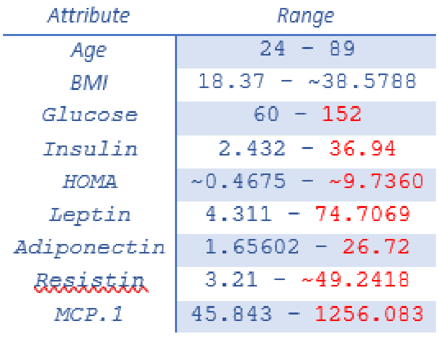
The dataset was represented as a CSV file format and a pandas python package was used to interpret the contents into a readable structure. The attributes were found to be a combination of float and integer datatypes and did not contain any missing or null values. Using the matplotlib package, the values of the attributes were visualised as a box plot to show if any outliers exist. In addition, the z-score is calculated for each value in the attribute relative to the mean and standard deviation. Any z-scores higher than 3 is considered an outlier and is consequently dropped. A total of 14 entries were removed to negate bias. Impossible values were only checked for Age and BMI attributes. The remaining attributes were left unchecked as it required domain medical expertise to determine an acceptable range. The range of each attribute after data preparation is shown in (Table 1.1).

Table 1.0: Pre-Prepared Breast Cancer Attribute Range



Attribute	Range
Age	24 - 89
BMI	18.37 - ~38.5788
Glucose	60 - 201
Insulin	2.432 - 58.46
HOMA	~0.4674 - ~25.0503
Leptin	4.311 - 90.28
Adiponectin	1.65602 - 38.04
Resistin	3.21 - 82.1
MCP.1	45.843 - 1698.44

Table 1.1: Post-Prepared Breast Cancer Attribute Range



Attribute	Range
Age	24 - 89
BMI	18.37 - ~38.5788
Glucose	60 - 152
Insulin	2.432 - 36.94
HOMA	~0.4675 - ~9.7360
Leptin	4.311 - 74.7069
Adiponectin	1.65602 - 26.72
Resistin	3.21 - ~49.2418
MCP.1	45.843 - 1256.083

- ● Affected
- ● Unaffected

Data Exploration

For each variable in the dataset, a graph was visualised using the matplotlib library or the pyplot extension of matplotlib. Each variable's graph type was chosen to be either a histogram, a box plot or a density plot based on what was deemed most appropriate for it. The variables where it was important to see the frequency for all of the different ranges it covered, such as Age, were represented using a histogram. The variables that seemed to have an approximately normal distribution of values, such as BMI, were represented using box plots. The variables that had a large spike in frequency around a certain range of values were represented using a density plot.

Figure 2.0: Age Histogram

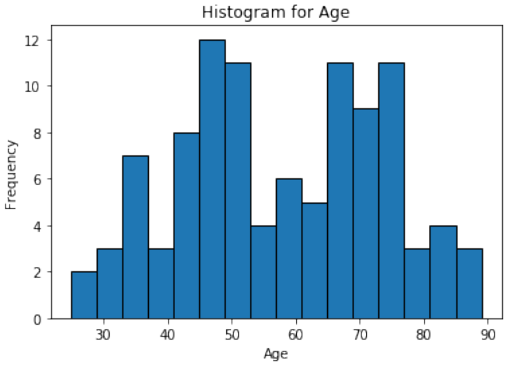


Figure 2.1: Body Mass Index (BMI) Boxplot

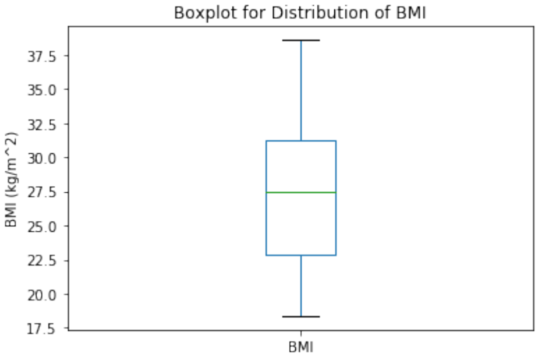


Figure 2.2: Glucose Boxplot

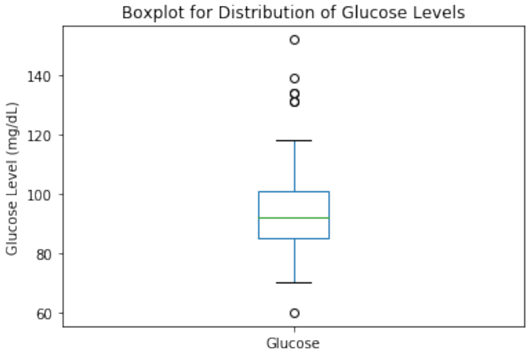


Figure 2.3: Insulin Density Graph

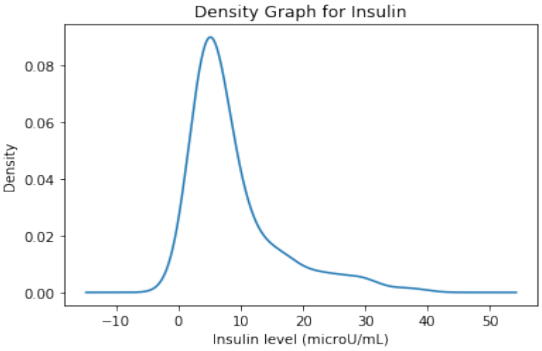


Figure 2.4: HOMA Density Graph

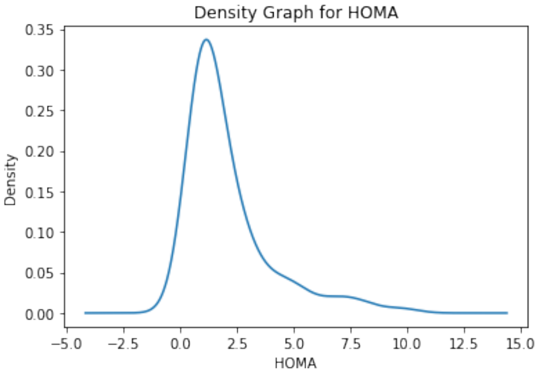


Figure 2.5: Leptin Histogram

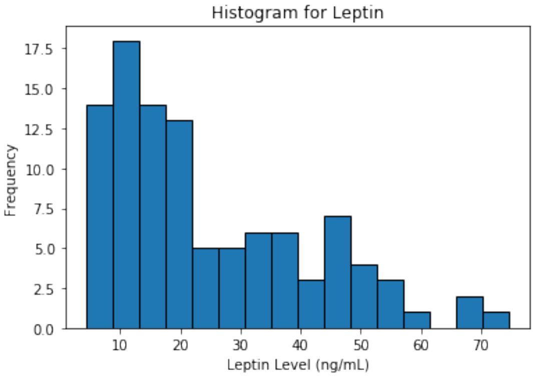


Figure 2.6: Adiponectin Histogram

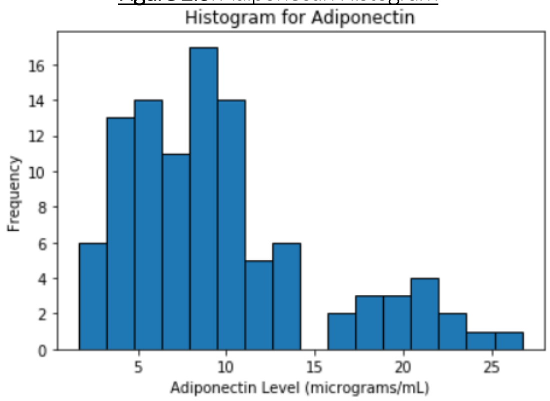


Figure 2.7: Resistin Density Graph

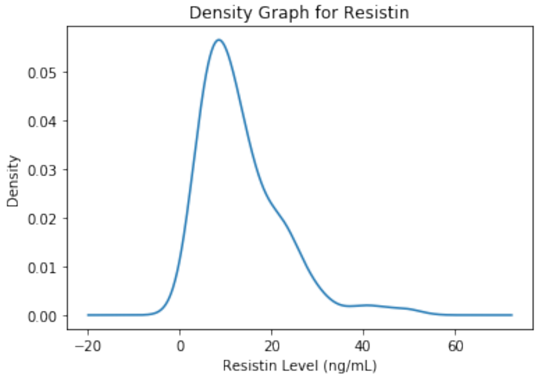


Figure 2.8: MCP.1 Boxplot

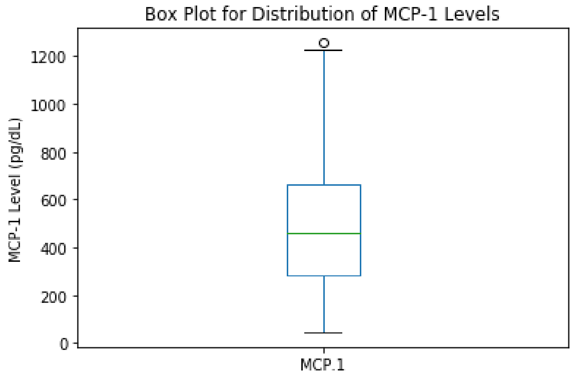
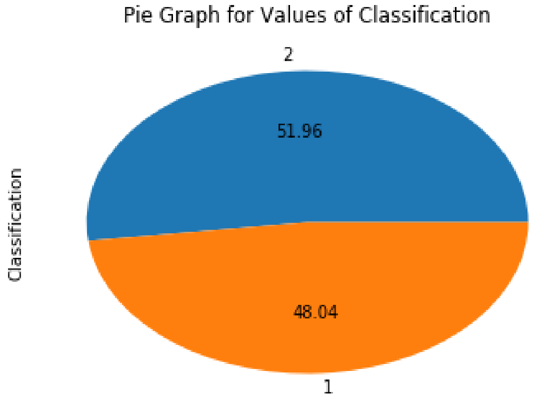


Figure 2.9: Classification Pie Graph



Ten pairs of variables were visualised (Figures 2.0 to 2.9), using the aforementioned libraries, to explore the relationships between them. A scatter plot, hexagonal binning plot, or grouped box plot was used based on what was deemed most appropriate. A scatter plot was used when the relationship looked close to linear or had no discernible pattern. A hexagonal binning plot was used when a large number of data points were grouped in a certain range. A grouped box plot was used when one of the variables being compared was binary (i.e. the variable Classification).

It was hypothesized that assuming an unbiased sample, there would be no relationship between age and BMI. This remained to be true as seen in (Figure 3.0). Therefore, the hypothesis is accepted.

Figure 3.0: BMI vs Age Scatter Plot

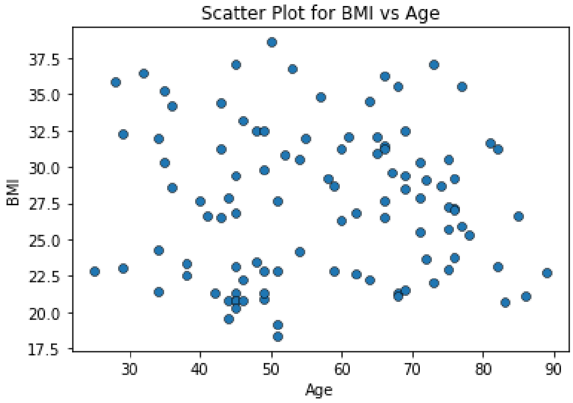
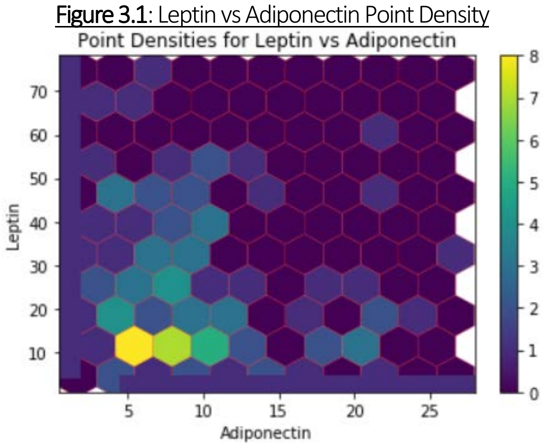


Figure 3.1: Leptin vs Adiponectin Point Density



It was predicted that leptin and adiponectin would have a negative linear relationship as they are inversely related. Evidently as shown above (Figure 3.1), most of the datapoints have low levels of both leptin and adiponectin, thus disproving the hypothesis.

It was predicted that healthy patients (Classification = 1) and patients with breast cancer (Classification = 2) would have the same distribution based on insulin. From Figure 3.2 below there is a drastic difference in the variance of the two groups, as well as a slight change in the median value. Thus, the hypothesis is rejected.

Figure 3.2: Insulin Boxplot Grouped by Classification

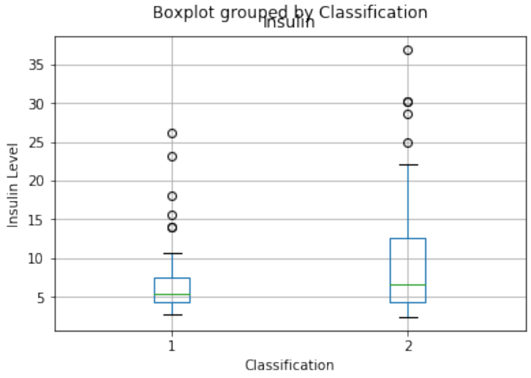
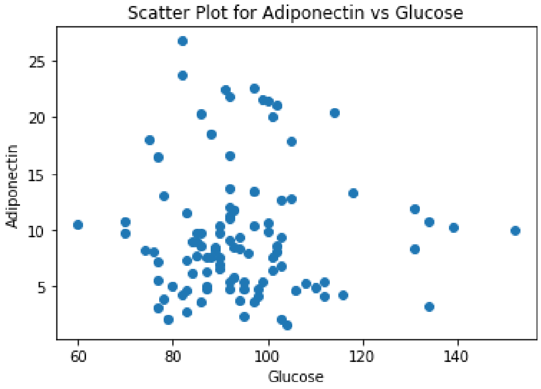


Figure 3.3: Adiponectin vs Glucose Scatter Plot



It was predicted that a higher glucose level would coincide with a higher adiponectin level, as they are both involved in breaking down food. Figure 3.3 shown above that there are a significant number of data points with a high glucose value and a low adiponectin value, or vice versa, disproving the hypothesis.

It was predicted that there would be no linear relationship between age and resistin as age doesn't change how active the immune system is. Although Figure 3.4 below suggests that the resistin levels for lower ages may be more spread out, there is no significant relationship to be seen. The hypothesis is then accepted.

Figure 3.4: Age vs Resistin Point Density

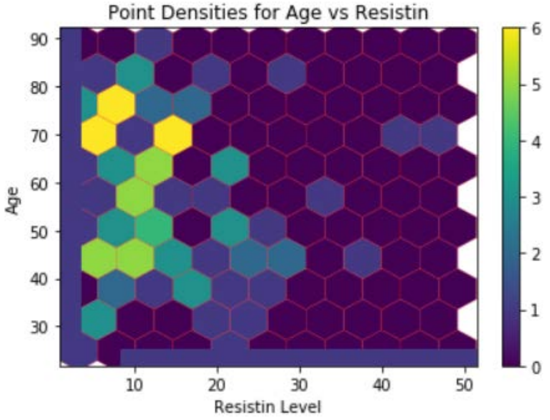
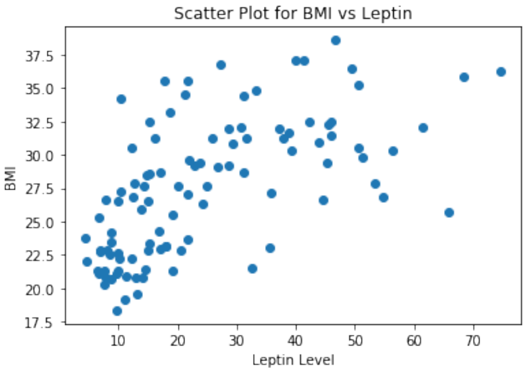


Figure 3.5: Body Mass Index (BMI) vs Leptin Scatter Plot



It was predicted that high leptin levels would correspond to high BMIs as leptin increases with fat stores. Figure 3.5 above shows that this hypothesis could be true, but no clear relationship, linear or otherwise, can be derived from this data. Thus, the hypothesis must be rejected.

It was predicted that healthy patients and patients with breast cancer would have the same distribution based on glucose. Figure 3.6 below indicates that there is a clear difference in the variances of the two groups, as well a slight difference in the median. The hypothesis is rejected.

Figure 3.6: Insulin Boxplot Grouped by Classification

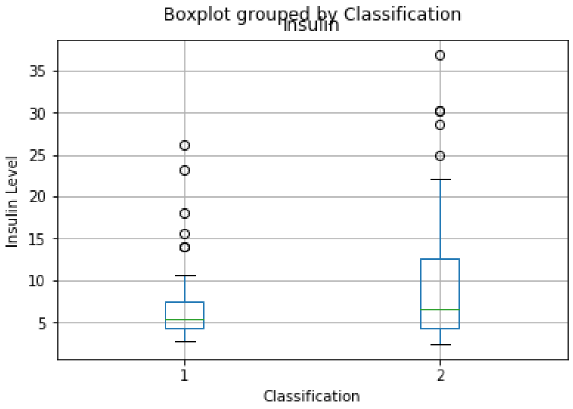
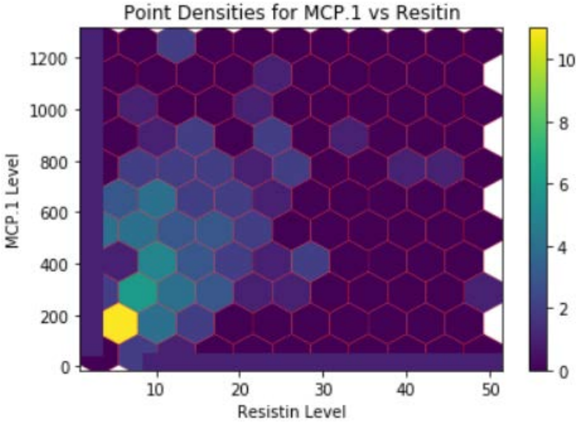


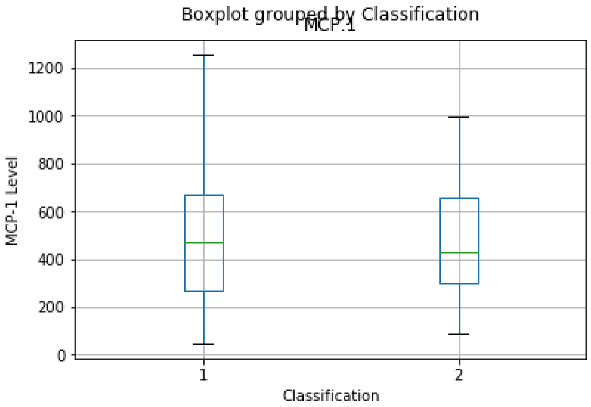
Figure 3.7: MCP.1 vs Resistin Point Density



It was predicted that there would be a linear relationship between MCP-1 and resistin as they are both involved in various immune system functions. In reference to Figure 3.7, low values for both variables coincided with each other quite significantly. The other datapoints mostly draw out a linear relationship. The hypothesis is accepted.

It was predicted that healthy patients and patients with breast cancer would have the same distribution based on MCP-1. Whereas figure 3.8 shows that there is a difference in the median values between the two groups and their overall ranges are dissimilar, their inter-quartile ranges are almost exactly the same, leading to a failure to reject the hypothesis. The hypothesis is accepted.

Figure 3.8: MCP.1 Boxplot Grouped by Classification

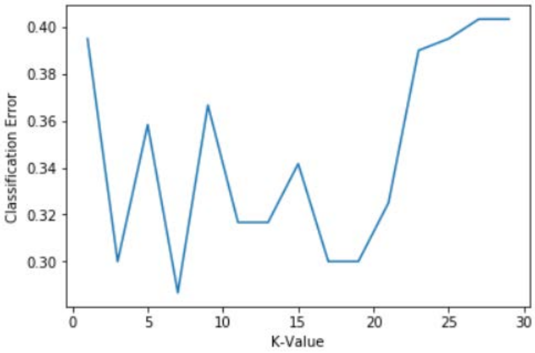


K-Nearest Neighbours Classifier

The approach taken to determine the best parameters for KNeighborsClassifier(n_neighbors, weights, metric, p) for each train/test split is through the use of hyperparameter tuning algorithms. From the sklearn package, RandomizedSearchCV determines the best values of these parameters by trying out a fixed number of parameter settings from a specified distribution, *n_iter*. With a predefined set of parameters whose names coincide with the parameters used for the selected model; cross-validation is used to evaluate possible combinations from the values of the parameters. The number of cross-validation and iterations can be increased to prevent overfitting and increase the accuracy at the expense of computational time. However, there are diminishing returns if these values are set too high. Once the algorithm had time to be fitted with the training data, the best parameter values can be extracted to be used for our KNN model.

To complement the result of RandomizedSearchCV, a method was developed to plot a list of k-values along with their classification error rate. Each k-value from the list is cross-validated and the score from the cross-validation is used to calculate the classification error rate of that k-value. The k-value with the lowest classification error is visualised as seen in (Figure 4.0) below. The optimal k-value should be like the one found by RandomizedSearchCV.

Figure 4.0: K-Value vs Classification Error



Decision Tree Classifier

The same RandomizedSearchCV approach was taken for the DecisionTreeClassifier(criterion, max_depth, min_samples_split, mine_sames_leaf, max_features, max_leaf_nodes). A predefined set of parameters appropriate for the decision tree classifier was used for RandomizedSearchCV to determine the best values for those parameters and the score (mean accuracy) it will give.

In additional to RandomizedSearchCV, another technique is employed to tune the parameters called GridSearchCV. Instead of searching a random subset of the provided parameter values, GridSearch will try every combination possible within it's given constricts. Because of this, GridSearch is computationally more expensive than RandomizedSearch. The higher scoring algorithm out of the two will be used for the decision tree classifier.

Results

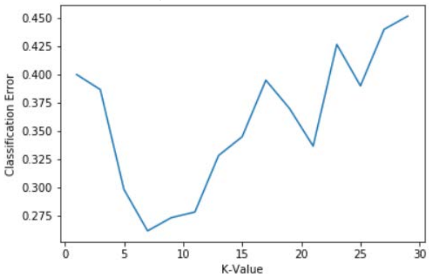
K-Nearest Neighbours Classifier

1st Train-test split (50% train, 50% test):

Confusion Matrix

21	**3**
13	14

Optimal K-value



Classification Report

	Precision	**Recall**	**F1-Score**	**Support**
1	.62	.88	.72	24
2	.82	.52	.64	27
Micro Avg	.69	.69	.69	51
Macro Ave	.72	.70	.68	51
Weighted Avg	.73	.69	.68	51

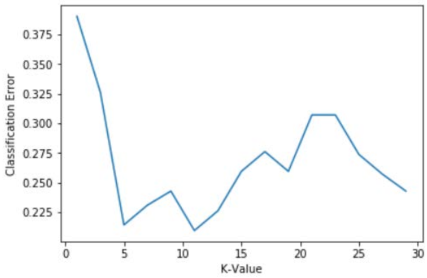
Classification Error Rate
16/51 = 0.3137

2nd Train-test split (60% train, 40% test):

Confusion Matrix

14	**5**
5	17

Optimal K-value



Classification Report

	Precision	Recall	F1-Score	Support
1	.74	.74	.74	19
2	.77	.77	.77	22
Micro Avg	.76	.76	.76	41
Macro Ave	.75	.75	.75	41
Weighted Avg	.76	.76	.76	41

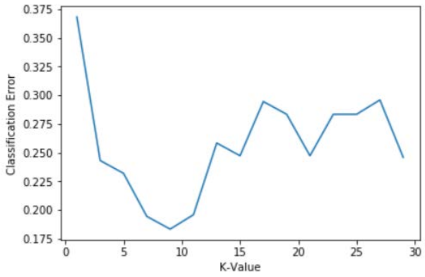
Classification Error Rate
10/41 = 0.24390

3rd Train-test split (80% train, 20% test):

Confusion Matrix

8	**1**
6	6

Optimal K-value



Classification Report

	Precision	Recall	F1-Score	Support
1	.57	.89	.70	9
2	.86	.50	.63	12
Micro Avg	.67	.67	.67	21
Macro Ave	.71	.69	.66	21
Weighted Avg	.73	.67	.66	21

Classification Error Rate
7/21 = 0.3333

Decision Tree Classifier

1st Train-test split (50% train, 50% test):

Confusion Matrix

21	**8**
6	16

Classification Error Rate
14/51 = 0.27451

Classification Report

	Precision	Recall	F1-Score	Support
1	.78	.72	.75	29
2	.67	.73	.70	22
Micro Avg	.73	.73	.73	51
Macro Ave	.72	.73	.72	51
Weighted Avg	.73	.73	.73	51

2nd Train-test split (60% train, 40% test):

Confusion Matrix

11	**8**
9	13

Classification Error Rate
17/41 = 0.41463

Classification Report

	Precision	Recall	F1-Score	Support
1	.55	.58	.56	19
2	.62	.59	.60	22
Micro Avg	.59	.59	.59	41
Macro Ave	.58	.58	.58	41
Weighted Avg	.59	.59	.59	41

3rd Train-test split (80% train, 20% test):

Confusion Matrix

7	**5**
1	8

Classification Error Rate
6/21 = 0.28571

Classification Report

	Precision	Recall	F1-Score	Support
1	.88	.58	.70	12
2	.62	.89	.73	9
Micro Avg	.71	.71	.71	21
Macro Ave	.75	.74	.71	21
Weighted Avg	.76	.71	.71	21

Discussion

K-Nearest Neighbours Classifier

The lowest classification error rate, from the 60%-40% train test set was less than 0.25, indicating that less than 1 in 4 observations was predicted incorrectly. The weighted average f1 score for each of the splits was vastly different, ranging from 0.66 to 0.76 (weighting precision and recall equally), solidifying the assumption that this model achieves around a 75% level of accuracy at best.

The precision for predicting a healthy patient fluctuates greatly between the train-test splits and is generally quite a low score, whereas the precision for predicting a breast cancer sufferer remained above 0.76. The recall for predicting a healthy patient did not drop lower than 0.74, while the recall for breast cancer patients was below 0.6 in two of three splits. This indicates that the model is more likely to misdiagnose a cancerous patient as healthy than it is to misdiagnose a healthy patient as one with breast cancer.

It would generally be expected that the values for precision, recall, f1 score and classification error rate became more desirable as the size of the test set increases, however, the third train-test split bucks that trend, possibly due to an anomaly in the data that was not picked up in the other two splits. More data would likely make this model more valid and lessen the effect of said anomaly.

Decision Tree Classifier

There is clearly an anomaly in the form of the second train-test split. The classification error rate is a very significant 0.13 more than the next highest value. The weighted average f1 score is also 0.12 lower than the next lowest. This indicates that the data in the train set was not a good representation of the whole set, and outliers may have been present. However, the same problems are not present in the other two train-test splits, so the 60-40 train-test split will be excluded from subsequent discussions.

Somewhat counterintuitively, the classification error rate and the weighted average f1 score became less desirable as the size of the training set increased. This indicates that somehow, the training set made up of 50% of the data was a better representation of the whole set than the training set made up of 80% of the data. This may be due to a group of observations that showed a different trend than the rest and were present in only the training set of the 3rd train-test split, while they showed up in both sets in the 1st train-test split.

The precision of the model when predicting healthy patients was quite high for both considered splits, as was the recall when predicting patients with cancer. On the other hand, the recall when predicting healthy patients and the precision when predicting cancerous patients were considerably lower, indicating that the model is again more likely to misdiagnose a healthy patient as one with breast cancer than it is to misdiagnose a cancerous patient as healthy.

Conclusion

There is most certainly a relationship between the presence of breast cancer in a person and the variables present in the explored dataset.

Based on the classification error rates and the weighted f1 scores, the k-nearest neighbours classifier produced a more accurate model than the decision tree classifier. The best train-test split of all was the 60%-40% k-nearest neighbours split.

The models differed in whether they were more likely to misdiagnose a healthy patient as one with breast cancer or to misdiagnose a cancerous patient as healthy. Both of these misdiagnoses are unacceptable, but the better of the two may be misdiagnosing a healthy patient as cancerous, meaning the decision-tree classifier may be a better fit for the industry. However, it would be more beneficial to create a model with a much lower rate of misdiagnosis altogether, ideally 0%.

It is recommended that the current model be improved by adding more data to the current model, adding a few more meaningful predictors into the dataset or finding a different modelling technique that achieves greater accuracy.

References

Dataset: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

Cancer Research UK. (2019). *Cancer cells*. [online] Available at: https://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-starts/cancer-cells [Accessed 18 May 2019].

Training.seer.cancer.gov. (2019). *Introduction to Breast Cancer | SEER Training*. [online] Available at: https://training.seer.cancer.gov/breast/intro/ [Accessed 18 May 2019].