CS 2120 Assignment #4 – Machine Learning Fever
Name: Michael Tran
Student number: 251033279

Introduction

Data set: Mammographic Mass data
source: http://archive.ics.uci.edu/ml/datasets/mammographic+mass
filename: mammographic_mass.csv

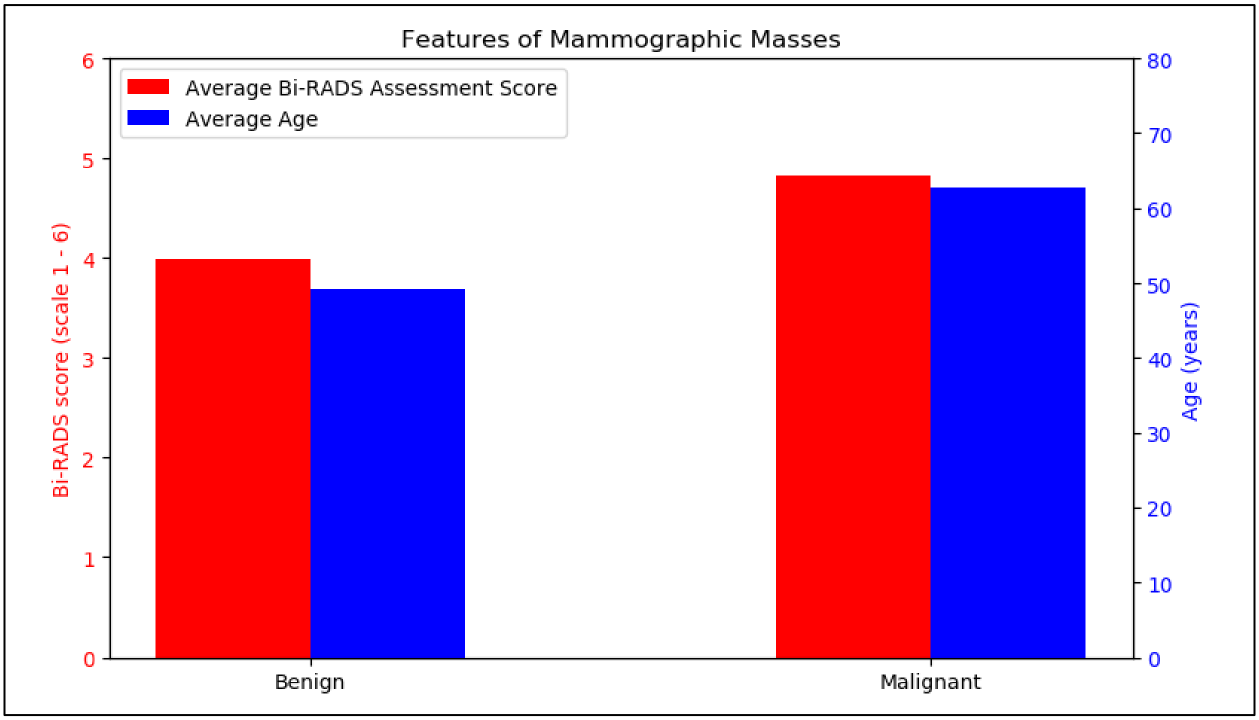
In this assignment, I worked with breast imaging data in the form of a .csv file (comma separated values). I wanted to work with this dataset because it is relevant to the medical health field, and the module I am interested in applying to next year is Medical Health Informatics in the BMSc Program. I also wanted to get some practice working with biological data and analyzing machine learning data relevant to real issues in our lives; in this case, analysing mammographic mass data and using machine learning to predict the severity of the mass i.e. benign or malignant.

Visualisation 1

Input:

```
mammographic_masses_data = load_data()
viz1(mammographic_masses_data)
```

Output:



Visualization Interpretation 1

From the mammographic_masses.csv data, two features were specifically chosen to be plotted in this double bar graph – the Bi-RADS assessment score and the age of the patients. These features were chosen because the values were ordinal values, meaning they could be represented by a value on a scale, and the average could be taken. The other features were nominal, meaning that the numbers were discrete and represented certain characteristics.

For some background, a Bi-RADS assessment score is essentially a way that radiologists can categorize mammogram results. It must be an integer ranging from 0 to 6, with lower numbers representing healthy results and greater numbers representing higher abnormality and malignancy. The values and interpretations are summarized in Table 1. In the graph, although the values are discrete, the scale will be taken as a measure of severity of the tumour, with greater values being more severe.

Table 1. Bi-RADS scoring taken from American Cancer Society (https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/understanding-your-mammogram-report.html)

Bi-RADS score	Interpretation
0	Incomplete - Additional imaging evaluation and/or comparison to prior mammograms is needed.
1	Negative
2	Benign (non-cancerous) finding
3	Probably benign finding – Follow-up in a short time frame is suggested
4	Suspicious abnormality – Biopsy should be considered
5	Highly suggestive of malignancy – Appropriate action should be taken
6	Known biopsy-proven malignancy – Appropriate action should be taken

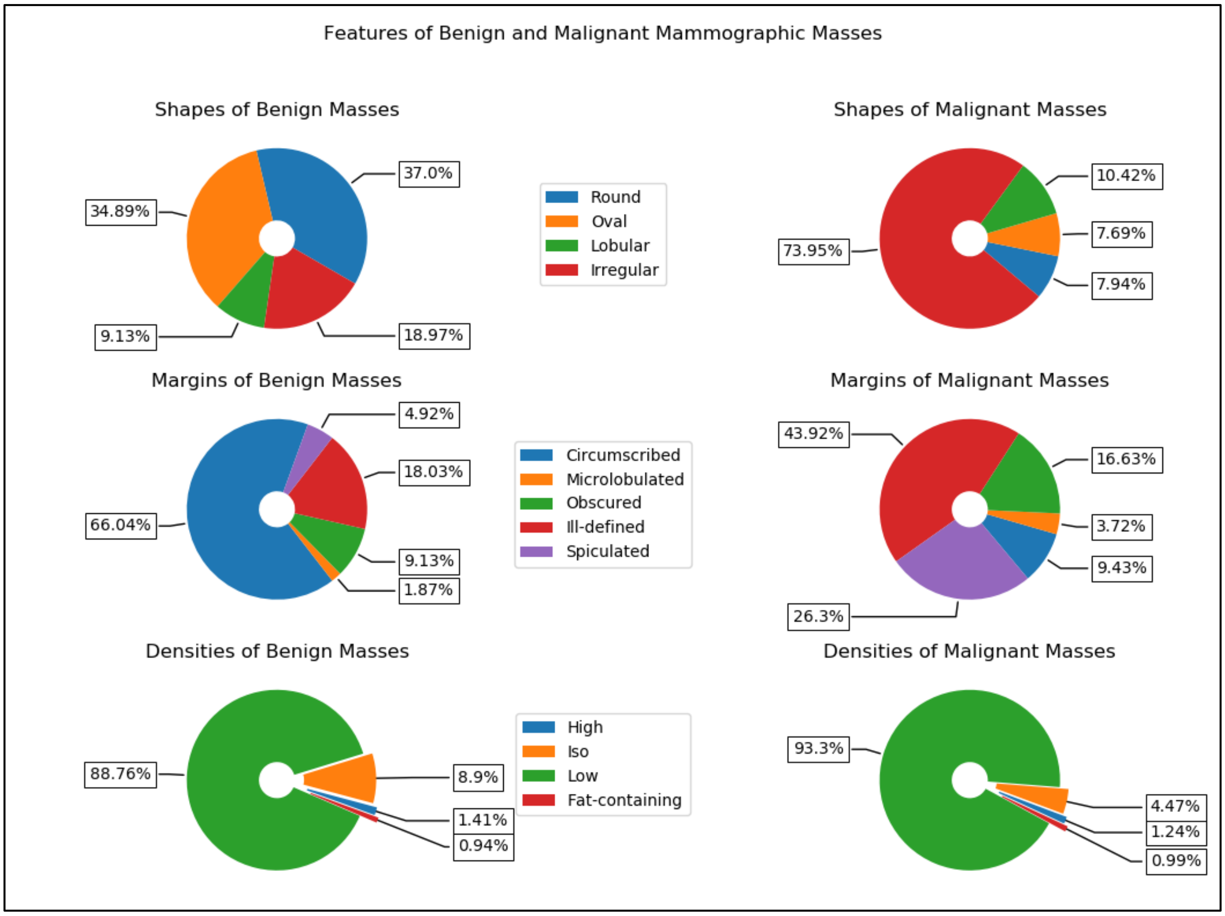
Overall, an immediate trend can be seen where the values of both Bi-RADS score and age is greater in malignant tumours than in benign tumours. This is what is expected of the values, as breast cancer is known to emerge in older women and higher Bi-RADS scores are representative of malignancy. It should be noted that the dataset contains patients who are believed to be at risk of breast cancer, and thus the high average Bi-RADS score and age for patients with benign mammographic masses is much higher than one containing complete data for a population including healthy women.

Visualization 2

Input:

```
mammographic_masses_data = load_data()
viz2(mammographic_masses_data)
```

Output:



Visualization Interpretation 2

In this series of pie charts, data was taken from the mammographic_masses.csv for the shape, margin, and densities of both the benign and malignant masses. These values in the data are nominal and thus a pie chart seemed to be the most appropriate way to compare the distribution of different characteristics in the mammographic masses. On the left, the subplots represent data from known benign masses, and malignant is presented on the right. In the middle column of the visualization is a legend for the colours represented in the pie charts to the left and right.

Comparing the two pie charts for shape, the benign masses are made up of primarily round and oval shaped masses, whereas the majority of the malignant ones are irregular. In the margin pie charts, benign masses show mostly circumscribed margins while malignant masses are more ill-defined and spiculated. These results line up with scientific literature, as these features are commonly looked for when diagnosing breast cancer and malignant tumours. Examples of what these mass traits might look like are found in Figure 1.

Mammographic mass density refers to the density of the tumour or mass in relation to the density of the surrounding breast tissue. Examples of this feature can be seen in Figure 2. The density pie charts seem to show little difference between benign and malignant masses. The only noticeable difference is that benign tumours contain a slightly larger ratio of iso density masses. This seems to suggest that benign masses are more likely to be of similar density to surrounding breast tissue, or that malignant masses are more likely to be of lower density. Overall, it can be deduced from the pie charts that in general, low density masses are more common, and rarity increases to iso and high density. Interestingly, the other end of the scale i.e. fat-containing masses is the least common, even less so than high density masses. I believe that these pie charts reflect more the distribution of density across a whole population and do not show a significant trend between benign and malignant mammographic masses.

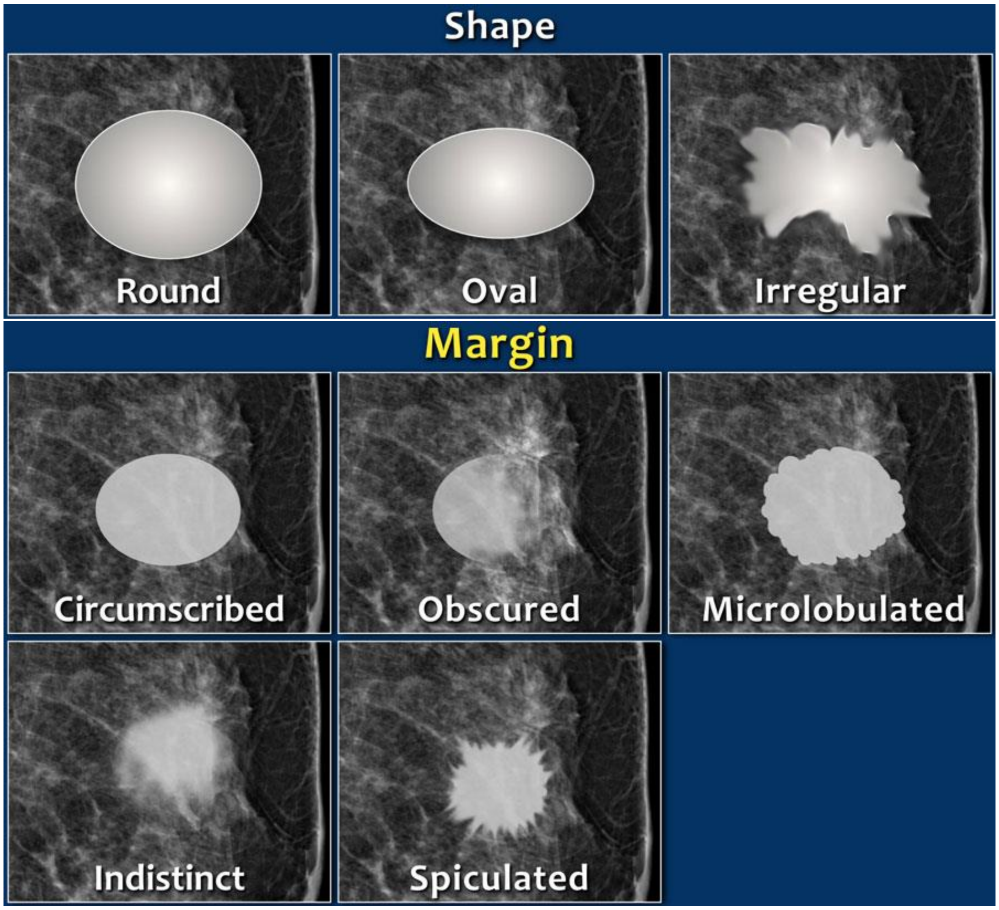


Figure 1. Example pictures of mammographic mass shape and margin. (https://radiologyassistant.nl/breast/bi-rads-for-mammography-and-ultrasound-2013)

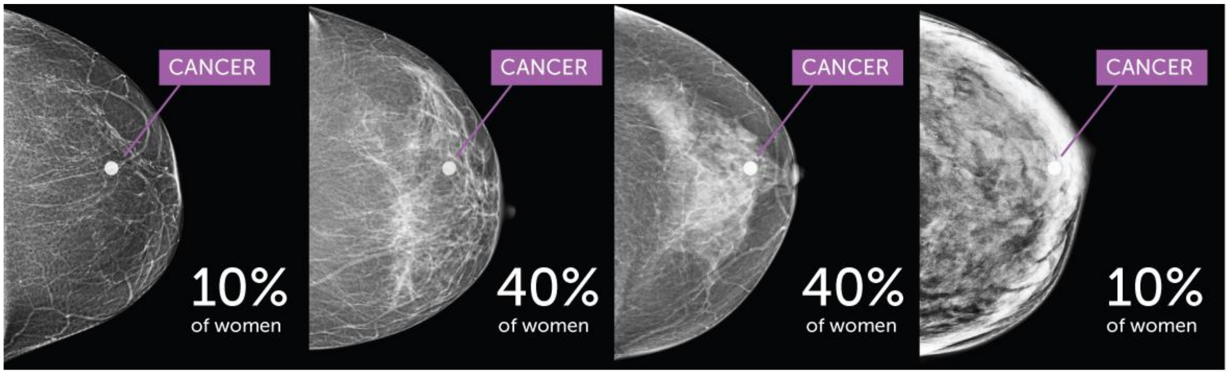


Figure 2. Pictures showing (left to right) fat containing, low, iso, and high mammographic mass density. (http://www.bccancer.bc.ca/screening/breast/breast-health/breast-density)

Machine Learning 1

Input:

```
mammographic_masses_data = load_data()
learn1(mammographic_masses_data)
```

Output:

```
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[1] = [0] [False]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[1] = [0] [False]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[1] = [0] [False]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[1] = [0] [False]
[1] = [0] [False]
[0] = [0] [ True]
[0] = [0] [ True]
[1] = [0] [False]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[0] = [0] [ True]
[1] = [0] [False]
[0] = [0] [ True]
[1] = [1] [ True]
[0] = [1] [False]
[1] = [1] [ True]
[0] = [1] [False]
[0] = [1] [False]
[1] = [1] [ True]
[1] = [1] [ True]
[1] = [1] [ True]
[1] = [1] [ True]
[0] = [1] [False]
[1] = [1] [ True]
[1] = [1] [ True]
[1] = [1] [ True]
[0] = [1] [False]
[0] = [1] [False]
[0] = [1] [False]
[1] = [1] [ True]
[1] = [1] [ True]
[0] = [1] [False]
[1] = [1] [ True]
[1] = [1] [ True]
[1] = [1] [ True]
[0] = [1] [False]
[0] = [1] [False]
[1] = [1] [ True]
[0] = [1] [False]
[0] = [1] [False]
[1] = [1] [ True]
[1] = [1] [ True]
[0] = [1] [False]
[0] = [1] [False]
[1] = [1] [ True]
[1] = [1] [ True]
[0] = [1] [False]
[1] = [1] [ True]
[1] = [1] [ True]
[1] = [1] [ True]
[0] = [1] [False]
[0] = [1] [False]
[1] = [1] [ True]
Test Result: 71.08% accuracy.
```

Machine Learning Interpretation 1

The dataset I used is one that supports the classification machine learning model. Classification is where data fitting is applied to certain features and compared to corresponding labels, after which it uses the information in order to predict the label of an unknown data point given the same feature values. This is useful in applications such as detection of certain traits in data, or when taken to the next level, image recognition. The machine learning model used in this function is `sklearn.svm.svc`, or the support vector classifier under support vector machines. A visualization of the classification trend in this model is found in Figure 3. In this dataset, the features include Bi-RADS assessment score, age of patient, shape of mass, margin of mass, and density of mass. The label corresponding to these values is binomial: either benign or malignant.

The original dataset is first split into a training group and testing group at a ratio of 9:1. The training group is fitted into the svc model while the testing group is set aside to test the accuracy of the machine learning model. For each set of features in the testing group, the svc model attempts to predict the label as being either 0 (benign) or 1 (malignant). This predicted value is compared to the actual label associated with the features, and either true (correct prediction) or false (incorrect prediction) is printed. The process is repeated for every single feature set in the testing group and the final accuracy percentage is calculated as the number of correct predictions out of the total number of predictions made.

Machine Learning 2

Input:

```
mammographic_masses_data = load_data()
learn2(mammographic_masses_data)
```

Output:

```
Balanced Accuracy Score: 0.7695
```

Machine Learning Interpretation 2

This machine learning function is very similar to the first function. One primary difference is the use of a different machine learning model, this one being `sklearn.neural_network.MLPClassifier`, or multi-layer perceptron classifier. This is a supervised machine learning model that uses an artificial neural network to fit data, with a perceptron algorithm used as a binary classifier to predict using linear functions. A visualization of this machine learning approach can be seen in Figure 3.

Like the first machine learning function, the dataset is split into training and testing groups of ratios 9:1. Using the training group to fit data to the model, a new function, `compute_bac`, is called in order to judge the accuracy of the predictions made for the testing group. The accuracy score is taken from `sklearn.metrics.balanced_accuracy_score`, which takes two input lists, one for real values and one for predicted values. It returns a balanced accuracy score value between 0 and 1 that is calculated as the average of recall per label. A balanced accuracy score is useful for creating normalized accuracy measurements for binary classification (i.e. this dataset) that takes imbalance into account. The mammographic dataset is imbalanced because it contains an unequal number of each label (less malignant masses than benign masses).

The outputted balanced accuracy score is 0.7695—equivalent to 76.95%. When `judge_accuracy` is called, it returns a similar score to the `compute_bac` function for the same machine learning model used. Since the latter function takes into account more factors when computing the accuracy, it can be deduced that this score is a better estimation of accuracy. Overall, the machine learning models have a fair accuracy score of 70% or greater, and the applications of such a classifier can be very useful in predicting breast cancer in patients. When taken further, malignant tumours can be identified with greater certainty and thus treatment options can be taken earlier and increase survival rates. These applications are highly convenient in decreasing the high cancer mortality rates and improving medical diagnosis.

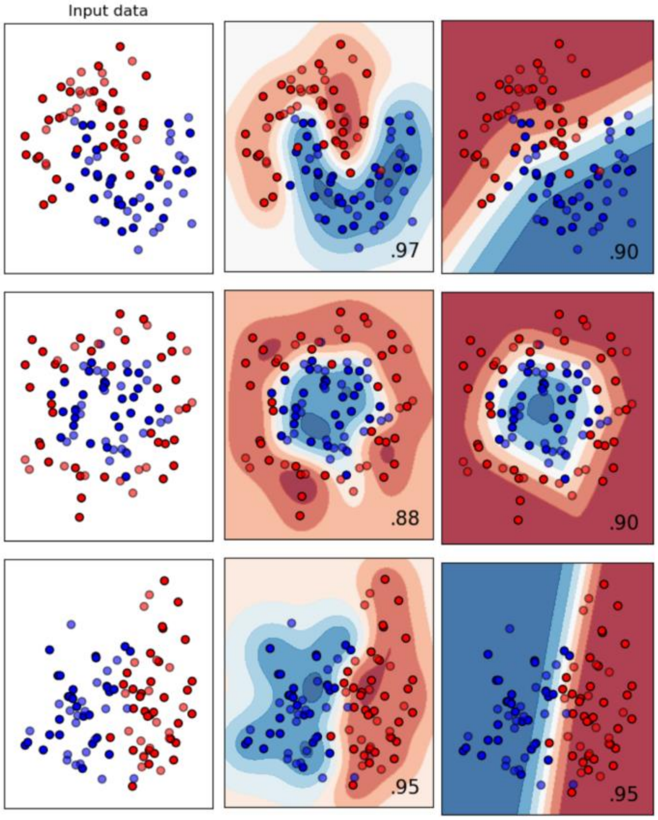


Figure 3. Example visualizations of machine learning classifiers. The plots show training points in solid colors and testing points semi-transparent. The lower right shows the classification accuracy on the test set. Input fitted data is pictured in the left column, the middle column shows the `sklearn.svm.svc` classifier accuracy, and the right column shows the `sklearn.neural_network.MLPClassifier` classifier accuracy. (https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html#sphx-glr-auto-examples-classification-plot-classifier-comparison-py)