A Study Into The Applications of Machine Learning in The Diagnosis & The Prognostication of Breast Cancer

Department of Computing

Letterkenny Institute of Technology

Machine Learning

Ultan Kearns & Liam Millar

***Abstract*—Machine Learning has been used in a variety of healthcare fields and has proven to be very successful in regards to automating / helping with certain tasks ranging from identifying similarities in different strains of DNA and making inferences on new DNA strains to diagnosing various types of diseases and predicting the prognosis from sample data. Machine Learning is particularly useful when it comes to analyzing and comparing data, this data can then be used to train a model to notice certain similarities in malignant and benign tumours. This paper aims to analyze which methods would be best suited to identifying factors which contribute to a cancer being malignant or benign and to predict the prognosis of breast cancer based on those factors.**
keywords - Machine Learning, Breast Cancer, Prognosis, Diagnosis

I. Introduction

Machine Learning is a field which aims to train machines to think and learn like humans. This is done by feeding data into the machine, the machine then analyzes the data and begins to start noticing patterns within the data and starts to create a model, this model can be used to train the machine so that when new data is feed in the machine can easily detect certain characteristics about the data and classify it. Machine Learning is a subset of AI which is comprised of many fields, and has been proven to perform very well in predicting outcomes in a large variety of areas such as STEM fields, Economics, Social Sciences and various others. The ultimate goal of machine learning is to train a machine to perform a task without any human intervention and to learn to adapt to new data and new patterns which appear in the data. There has been an increased reliance on machine learning in the past few years and it has become ubiquitous in the modern world.

There are many ways which a machine may choose to learn we will explain the types of learning we used when trying this model in the upcoming sections.

A. Supervised Learning

Supervised learning is a type of machine learning that involves training a machine with labelled training data to make inferences about new data. The machine reads in the data and begins to notice certain patterns from it and uses the learned patterns to make inferences about new data. There are many algorithms which can be used in supervised learning, we will list a few we used below.

1) Linear Regression: This algorithm involves finding a line to best fit the data, this is achieved by modifying parameters to find the best fitting line for a given dataset. This method works by learning to predict a pattern within the data and drawing a

line through a series of data points so that the line fits is close to as many of the points as possible, in this way we can make predictions by seeing how closely a given point would fit our line.

2) Naive Bayes: Naive Bayes is a technique which makes features of an object have the same weight when trying to classify the data. This algorithm works well on small data sets. The model has some cons to it such as the zero frequency phenomenon which we will discuss in our Methodology section. An example of Naive Bayes would be to train an image classifier for identifying cars, using Naive Bayes the property of a cars wheels would have just as much impact as the number of doors of the car. The type of Naive Bayes we used in our notebook is called Gaussian Naive Bayes and assumes our data is in the shape of a Gaussian Curve, in other words normally distributed(think of a bell curve).

B. Unsupervised Learning

Unsupervised learning is a type of machine learning where the machine is not provided with any labelled data and makes inferences about the data based on patterns found in the training data, the most common types of unsupervised learning are reinforcement learning which ”rewards” a machine for performing in a certain way and ”punishes” it for acting in the ”wrong way” and clustering which groups data together based on similar features.

1) Random Forest: The random forest algorithm creates multiple decision trees based on different features in the dataset to determine whether a tumour is benign or malignant. It first creates multiple trees then combines them into the final tree using the average of all predictions, we use the final tree to make our predictions. We can also see the importance of each feature from it's GINI value.

2) K Nearest Neighbour: K Nearest Neighbour is an algorithm which measures the distance between data points by using a distance metric eg: Cosine distance, Euclidean distance, Jaccardian distance etc and groups data points which are near each other into clusters. In K Nearest Neighbour we make the assumption that data points within a certain distance must have some common properties or features, in this way we can separate the data into groups which share these features.

C. Breast Cancer

Breast cancer is a type of cancer which forms in the breast tissue, it has many signs and symptoms such as:

- New lump in the breast or underarm (armpit).
- Thickening or swelling of part of the breast.
- Irritation or dimpling of breast skin.
- Redness or flaky skin in the nipple area or the breast.
- Pulling in of the nipple or pain in the nipple area.
- Nipple discharge other than breast milk, including blood.
- Any change in the size or the shape of the breast.
- Pain in any area of the breast.

[1]

Breast cancer is also one of the most common occurring cancers in women and the average risk of a woman developing cancer in her lifetime is about 13% or 1/8 in the United States alone[11] and in 2020 there were 2.3 million women diagnosed with the disease of which 685,000 succumbed to the disease[10]. In Ireland 3,700 new cases of breast cancer are diagnosed each year[2]. Breast cancer treatment is has a very high survival rate if caught early, the survival rate for those in high income countries was 90% [10].

D. Libraries we used in this project

I have listed the libraries we have used in this project below:

- numpy referred to throughout our notebook as np[6] - the Numpy library was used to perform numerical operations on our dataset
- pandas referred to throughout our notebook as pd [7] - Pandas was used to create dataframes from our dataset
- sklearn referred to throughout our notebook as sk [8] - We used SKlearn to implement many of our models and to output statistics relating to the models.

- matplotlib [5] - We Used the matplotlib library to create visual interpretations of our data
- google.colab [3] - Google Colab was used to upload our data set and also as an editor so we could compile and run our Jupyter Notebook[4]

E. Aims of the project

The aim of this project is to train a model using machine learning to analyze data from both benign and malignant tumours to detect patterns in the data and to predict whether tumours will be malignant or benign given certain parameters with reasonable accuracy.

II. Methodology

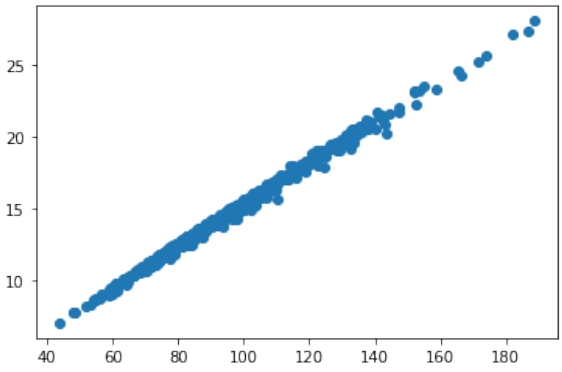
A. Creating testing and training sets

We created testing and training sets from the data set by splitting the data set using the Sklearn Library[8] and creating four dataframes two for our X & Y test sets and two more for our X & Y training sets. We then used the training sets to train a variety of models to make predictions about our test sets, we then measured the models performance based on how accurately it predicted data in the test set. We used a variety of metrics to compare our data when using different models and gauged how accurately they performed on our test data. When deciding how we should split the data we experimented with a number of values to decide on which split best suited our models, we determined this through trial and error rerunning our models with different splits to determine which split yielded the most accurate prediction.

B. Cleaning The Data

The dataset we used was the Breast Cancer Wisconsin (Diagnostic) [9]. This dataset consists of features computed from digitized images of fine needle aspirate (FNA) of breast mass. It contains 569 instances(rows) and 32 variables(cols). There are 2 categorical variables ID and Diagnosis(our target). There is also 30 continuous variables used to describe the features broken into 3 sets of ten measurements using different methods. There was no missing data in this set so no imputation methods needed to be used. I have included an image below to show a positive correlation between the Mean Radius and Mean Perimeter columns. We also changed the diagnosis columns from 'M' for malignant and 'B' for benign to 1 for malignant and 0 for benign, this made the data much easier to work with and helped us in performing various algorithms with it.

Fig. 1. Example of a positive linear correlation between the mean radius & perimeter, if the lines slope was the inverse of our current line that would be an example of high negative correlation and we would not have needed to remove these columns from our data set



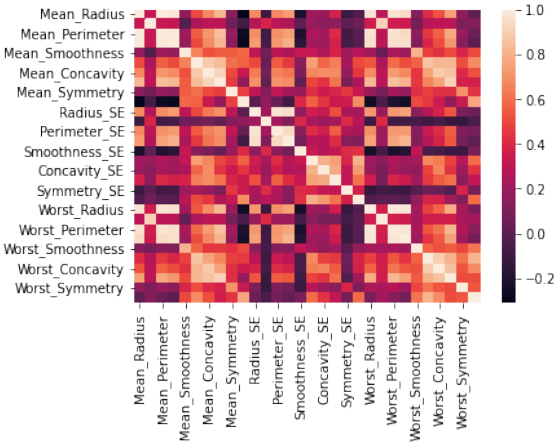
We were able to find this highly correlated data by using Pearson's correlation and I will include an image of the correlation in our dataset below

Fig. 2. Example of highly correlated variables in our data set as you can see the mean area corresponds highly with the data highlighted

	Diagnosis	Mean_Radius	Mean_Texture	Mea
Diagnosis	1.000000	0.732785	0.461971	
Mean_Radius	0.732785	1.000000	0.340956	
Mean_Texture	0.461971	0.340956	1.000000	
Mean_Perimeter	0.748496	0.997802	0.348142	
Mean_Area	0.734122	0.999602	0.344145	
Mean_Smoothness	0.371892	0.148510	0.024649	
Mean_Compactness	0.609288	0.497578	0.266499	
Mean_Concavity	0.733308	0.645728	0.342646	
Mean_Concave_Points	0.777877	0.759702	0.306891	
Mean_Symmetry	0.332567	0.120242	0.110130	
Mean_Fractal_Dimension	-0.025903	-0.349931	-0.059303	
Radius_SE	0.616912	0.550247	0.363621	
Texture_SE	0.019419	-0.144499	0.450720	
Perimeter_SE	0.630411	0.565520	0.386813	
Area_SE	0.714184	0.738077	0.395139	
Smoothness_SE	-0.052193	-0.326385	0.037048	
Compactness_SE	0.380666	0.264904	0.263591	
Concavity_SE	0.470338	0.364555	0.287188	
Concave_Points_SE	0.488717	0.410576	0.238610	
Symmetry_SE	-0.092303	-0.241376	0.008945	
Fractal_Dimension_SE	0.201492	-0.008411	0.147605	
Worst_Radius	0.787933	0.978604	0.366547	
Worst_Texture	0.476720	0.314911	0.909218	
Worst_Perimeter	0.796319	0.971555	0.375273	
Worst_Area	0.786902	0.978863	0.368335	

Additionally we produced a heat-map of the correlations.

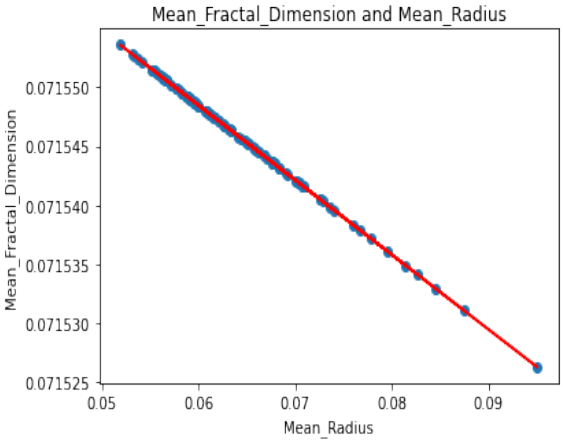
Fig. 3. The heat-map of correlations. This indicates a clear correlation between some of the columns and the complexity emphasises the need to reduce these columns where possible.



C. Algorithms

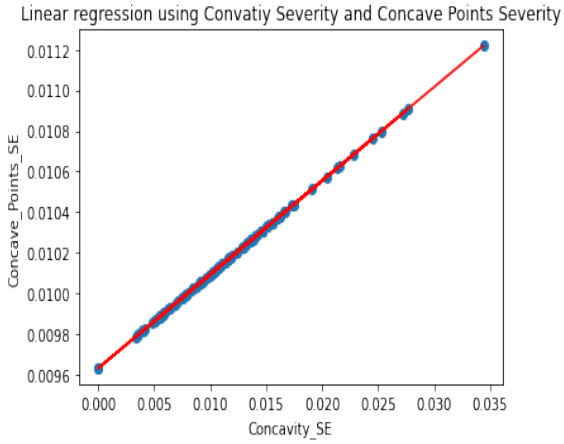
1) Linear Regression: In the notebook we used linear regression to predict values from our X test set based on features in the X training set. We decided to use Mean Radius and Mean Fractal Dimension as an example of our negatively skewed linear regression model. Our dependent variable was the Mean Radius and our independent variable was the Mean Fractal Dimension. We trained the model by first creating our X training set which contained the Mean Radius feature and then we created or X test set using the Mean Fractal Dimension we then reshaped the training and test data into a 2d array containing our feature values. We then fit a model to display our results as we can see in the image below the relation between Mean Fractal Dimension and Mean Radius is negatively correlated. We also observed that the correlation was negative and fairly small compared to the correlation value in the second model.

Fig. 4. Linear Regression Model of Mean Fractal and Mean Radius.



For our second linear regression model we decided to use two features which were highly correlated Concave Points Severity & Concavity Severity our independent variable was the Concavity Severity which we used to predict the dependent variable which was the Concave Points Severity. As you can see from the image below the data was highly correlated and thus gave us the inverse line of the above negatively skewed model.

Fig. 5. Linear Regression Model of Concave Points Severity and Concavity Severity.



2) Decision Trees: Decision Trees are a form of supervised learning that construct a flow chart from a set of training data and its results. This flow chart allows the prediction of future results based on a comparison to the features. However they can be prone to over fitting as all future decisions are based on the importance attributed to features of the training data. As part of our exploration of the data set we did a side by side comparison of decision trees based on information gain(entropy) and the gini index(gini) methods. From our results we discovered that while both methods provided quite accurate results 0.9005847953216374 or about 90.0% for information gain and 0.9298245614035088, or about 93.0% for the gini method. the scikit default gini proved to be the most accurate.

3) Random Forest: Random Forest is used to construct multiple decision trees from samples of the training data in order not to limit its scope and try to alleviate some of the issues with over fitting and precision that can occur with a single decision tree, this collection of trees produce separate predictions that are used as votes to form a majority decision on the final predicted result. We used the same data split to test using the random forests for both gini and entropy(information gain), once again accuracy was high with 0.9473684210526315 or 94.7% for gini and 0.935672514619883 0r 93.6% for entropy, these values varied a little for the random forest due to the randomness but they remain consistently higher than the respective decision tree model and even the entropy random tree models accuracy improved upon the gini singular decision tree.

4) Cross Validation: Because our dataset was reasonably small we ran some cross validation testing on the tree and random forest models

5) Naive Bayes: We implemented Naive Bayes in our notebook by using all features from our training set as well as the diagnosis values in the y training set to train a model to predict the diagnosis of breast cancer. We chose to use all features because in general the more features Naive Bayes has the more accurate the prediction. When first implementing this algorithm we encountered an issue where the model would predict the same values over and over again. We rectified this by scaling the data which ensured the model had a relatively even probability of predicting both malignant and benign.

6) K Nearest Neighbour: We implemented K Nearest Neighbour by using our X train set which contained our training values and our y train set which contained our diagnosis values, the model was then fitted using the X and y training sets. We chose a value of 55 for the neighbours we've passed in to the model as this yielded the best results. We then tested the model by comparing values in our test array to the predicted values of our model, the model was right on average around 520 out of 569 times, giving it a 91% accuracy rating. The models accuracy changes with the test / training set split the values mentioned above were achieved using a 50/50 split.

III. Challenges

A. Over fitting

Over fitting occurs when the model is performing very well on the training data but performs poorly on new data, this is due to the fact that the model has been trained to fit the training data so well that it cannot adapt to different data. To try and alleviate this we performed dimensionality reduction to find the features that are discriminative.

1) Principal Component Analysis(PCA): This is a form of unsupervised dimensionality-reduction and as our dataset contained a high number of features 30 we also explored the possibility of reducing these features by finding principle components that retain the essential parts of the data with the most variation while eliminating the sections that had little variation and less essential. A number of the features we saw were highly correlative as they were derived from each other such as radius, diameter and area of a circle. This implied that PCA was a good method of dimensionality reduction to explore. We standardised the dataset to remove the impact of diverse measurement scales.

B. Under fitting

When training the models we also had to worry about underfitting the data, underfitting occurs when the model does not predict well enough on either the training or test sets. Such a model is not practical and has no purpose or uses.

C. Naive Bayes

The Naive Bayes model suffered from an issue in the beginning where it only predicted one value for all values in our dataset. This may have been caused by a phenomenon known as zero frequency, which is where our data may not have had any malignant or benign labels, we rectified this by scaling the data and achieved fairly accurate of 158 / 171 or around 92%

IV. Applications

A. Diagnosis of Breast Cancer

The KNN model has a fairly accurate prediction of the diagnosis as does our first Naive Bayes model, as with all diagnostics tools there will need to be a doctor on hand also to analyze the information in case the AI generates false negatives or false positives. The models do provide however an insight into whether a tumour is benign or malignant.

B. Prognosis of Breast Cancer

The models can predict whether a tumour will be malignant or benign from features in our dataset this means that doctors can feed in the values of said features and produce a prognosis indicating the possibility of the tumour being benign or malignant. The models we included in this notebook can offer predictions on tumours being benign or malignant but these predictions should always be verified by a trained professional as some models can lead to false positives or much worse in our case, false negatives. Machine Learning is very useful in the medical field but even the best trained models can yield false predictions.

V. Conclusion

A. Accuracy of Models

1) Naive Bayes: We included two Naive Bayes models in our notebook, the first and third seems to predict poorly due to the data not being scaled although due to the random nature of our splits the performance varies. In our second alternative Naive Bayes model we scaled the data producing much better predictions. The issue we had with the first and third models is due to a phenomenon called Zero Frequency, from our probabilities we noticed one value was much closer to 0 than the other in the first model the probability for 1 is very high where as the probability of 0 is very low and vice versa for the third model. We decided to leave the first and third model in our notebook to demonstrate this behaviour.

2) KNN: The KNN model predicts fairly accurately when the training / test split is 0.5, below I will include the correctly predicted values out of the actual values.

Total values: 569 correctly predicted values: 520 which is approximately 0.91%

From the above results we can see that this technique works very well with our data set with around 9 / 10 predictions being accurate.

B. What could be done differently

Our dataset was quite small which would lead to some concerns about over fitting for the training data especially for the likes of decision tree. We took steps to alleviate some of the concerns about this. For the purpose of our exercise this was fine but ideally an expanded set of real world data points for training would ensure a more confident outcome.

CONTENTS

REFERENCES

[1] CDC. *What Are the Symptoms of Breast Cancer?* URL: https://www.cdc.gov/cancer/breast/basic_info/symptoms.htm.

[2] Breast Cancer Ireland. *Breast Cancer, Facts and Figures*. URL: https://www.breastcancerireland.com/education-awareness/facts-and-figures/.

[3] Multiple. *Google Colab*. URL: https://colab.research.google.com/.

[4] Multiple. *Jupyter*. URL: https://jupyter.org/.

[5] Multiple. *Matplotlib*. URL: https://matplotlib.org/.

[6] Multiple. *numpy*. URL: https://numpy.org/.

[7] Multiple. *Pandas Library*. URL: https://pandas.pydata.org/.

[8] Multiple. *Sklearn*. URL: https://scikit-learn.org/stable/index.html/.

[9] Wolberg & William Street & W. & Mangasarian & Olvi. *Breast Cancer Wisconsin (Diagnostic)*. UCI Machine Learning Repository. 1995. URL: https://archive-beta.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+diagnostic.

[10] World Health Organization. *Breast Cancer*. URL: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[11] American Cancer Society. *How Common Is Breast Cancer?* URL: https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html.

LIST OF FIGURES