Exploring the Breast Cancer Wisconsin Diagnostic Dataset

# Introduction

Breast cancer is the most common cancer in women, accounting for 15% of all cancer diagnoses in the UK. Improved treatments mean 76% of diagnosed patients now survive, however, improvements are needed at an early stage to help prevent such a high incidence rate, and further increase the level of survival. Although trained judgement is still required, machine learning is a tool that can aid in early diagnosis by helping pick up early signs at an accelerated pace. Here, the Breast Cancer Wisconsin dataset is studied, leading to the application of machine learning to predict the outcome of diagnosis.

# Methods

The Breast Cancer Wisconsin Dataset was loaded from the Scikit Learn website. All code developed to analyse the dataset was written in Python, using the browser based Jupyter Notebook platform. Please see associated README file for the code used.

# Results

The dataset required no cleaning due to no missing datapoints. The true diagnosis of the dataset was first studied (Fig. 1), giving an overall understanding of the data.

|  |
| --- |
| Chart, bar chart  Description automatically generated |
| Figure 1. Total number of malignant or benign diagnoses from the dataset. |

The dataset was then split into two, an individual dataset for cases diagnosed as either malignant or benign. This allowed comparison of correlation between variables, allowing detection of those more correlated in malignant cases (Fig. 2). Further analysis was then carried out, exploring interesting relationships between variables that showed higher correlation in malignant cases over benign cases (eg Mean radius vs mean concave points).

A logistic regression model was then formed. A logistic model is used to predict the outcome of a certain event, in this case diagnosis of a malignant or benign tumour. The model was trained on 75% of the data, allowing 25% to be predicted. The outcomes of the model are visualised as a confusion matrix (Fig. 3).

|  |  |
| --- | --- |
| Treemap chart  Description automatically generated with low confidence |  |
| Figure 2. Heatmap comparing correlation results from malignant or benign variables.  Correlation values of >0.75 were discarded due to high correlation of points such as mean radius and mean perimeter – by definition, this relationship is high. | |

|  |
| --- |
| Chart  Description automatically generated |
| Figure 3. Confusion matrix showing model predictions against true values. |

As seen in the matrix, the model was not 100% accurate, mislabelling 6 malignant cases (1) as benign. The accuracy of the model was 95.1%. Other analysis and models were explored and can be seen within the associated README file.

# Conclusion

The use of machine learning is still developing, with many useful applications in society. However, the application to something so delicate as exact cancer diagnosis is still far away.

Interesting relationships can be visualised, encouraging further study as to why certain factors are linked, potentially leading to new findings that help us understand cancer. Unfortunately, as seen with the results here, it is rare to find a model that is 100% accurate, a vital feature for cancer diagnosis. As machine learning develops, with increased data input, models will, inevitably, become more reliable to the extent that they can be relied upon.

References: Cancer Research UK; Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. **Github** = <https://github.com/Sga46>