**Project Plan**

## **Project Title:**

Feature Importance and Cluster Analysis for Predictive Accuracy on Wisconsin Breast Cancer data.

## **Research Question**

1. What are the most important features for classifying malignant tumours from benign tumours using Wisconsin breast cancer data?
2. Can unsupervised learning algorithms be used to detect clusters among the data that have an association with breast cancer?

## **Objectives**

* Use the Wisconsin breast cancer data to develop multiple supervised and unsupervised machine learning algorithms.
* Compare the performances of supervised ML algorithms among each other in tumour classification.
* Develop multiple unsupervised algorithms and check their associations with the known classifications (benign and malignant).
* To find out the most important features in the classification using ML algorithms.

## **Background and Summary**

Breast cancer is one of the most common types of diseases in women. Early diagnosis is one of the major factors that help in avoiding fatal outcomes. The diagnosis of breast cancer is traditionally done using histopathological analysis of the biopsied tissue. This process takes a lot of time, effort and requires a lot of specialized medical expertise. With the exponential growth in the performance of Machine Learning and Artificial Intelligence this process can be sped up to provide a lot of support to the medical practitioners in this field (Barth, 2024).

In this research a carefully curated dataset called “Wisconsin Breast Cancer” data is taken from UCI Machine Learning repository (Karanam, 2022). ML and AI can be used to analyse this data and understand which feature is affecting the predictions from these algorithms. Understanding the importance of each feature is very important as it can help to enhance model accuracy and provide further insights into the behaviour of the tumour itself and also about the tumour’s progression. Another approach is to use unsupervised learning techniques like clustering that have the capability to reveal the inherent structures in the data that may not be very apparent to a human viewing the data. By grouping patients together, it is possible to understand if a group of patients have a feature that all have same type of target variable classification.

In this way the application of ML in the diagnosis can help to understand various aspect of the disease and help to provide more insights into the behaviour of the tumour.

## **List of References**

Devi, R.D.H. and Deepika, P., 2015, December. Performance comparison of various clustering techniques for diagnosis of breast cancer. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-5).

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7435711>

Marne, S., Churi, S. and Marne, M., 2020, March. Predicting breast cancer using effective classification with decision tree and k means clustering technique. In *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)* (pp. 39-42). IEEE.

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9167544>

Radha, R. and Rajendiran, P., 2014, February. Using K-means clustering technique to study of breast cancer. In *2014 World Congress on Computing and Communication Technologies* (pp. 211-214). IEEE.

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6755142>

**2: Task List and Project Time Line**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Task** | **8-Jun** | **25-Jun** | **10-Jul** | **24-Jul** | **08-Aug** | **22-Aug** | **29-Aug** |
| **Data collection and project plan** |  |  |  |  |  |  |  |
| **Literature review** |  |  |  |  |  |  |  |
| **Data pre-processing** |  |  |  |  |  |  |  |
| **Supervised machine learning** |  |  |  |  |  |  |  |
| **Unsupervised machine learning** |  |  |  |  |  |  |  |
| **Comparison and results** |  |  |  |  |  |  |  |
| **Project documentation and reporting** |  |  |  |  |  |  |  |

As seen in the above Gantt chart the research is divided into 6 parts where each part is dedicated to a major part in the research.

* The first part for data collection and project plan is already completed.
* The next part is the Literature review where other research papers that have conducted similar research will be explored and a comprehensive document reporting them will be generated.
* The data pre-processing part will take care of eliminating inconsistencies like outliers and more. This will make the data suitable for the model development part.
* The next sections will cover the model development cycle from training, testing to comparison of both supervised and unsupervised algorithms.

**3: Data Management Plan**

## **Summary of Dataset**

As mentioned earlier the dataset used in this research is the Wisconsin Breast Cancer data. This dataset has information regarding the breast tissue of various patients with tumours. Overall, this dataset has around 569 observations and 32 features with various attributes, represent multiple attributes of the tumour in patients. This dataset contains a target variable that has two classes – benign and malignant. Some of the key features in the dataset include perimeter, smoothness, and concavity. The dataset has no missing values.

## **Data collection**

The data is collected from UCI Machine Learning Repository. This repository is owned by University of California which maintains the repository for public use on research and analysis of hundreds of open-source datasets.

## **Document control**

To control the updates of this research project GitHub and Git version control is used.

GitHub: <https://github.com/SusanVincent>

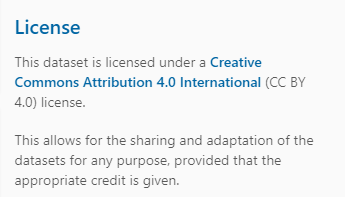
## **Ethical requirements**

1.     Does the data meet GDPR requirements? - Yes

2.     Does the project conform to UH ethical policies? – Yes

3.     Do you have permission to use the data for your proposed research project? - Yes

4.     Are you assured that the data was collected ethical (i.e. by the original people who gathered/collected/ collated/made the data)? – Yes



**References:**

Barth, S. (2024) Machine learning in healthcare: Guide to applications & benefits, ForeSee Medical. Available at: <https://www.foreseemed.com/blog/machine-learning-in-healthcare>

Breast cancer Wisconsin (diagnostic) dataset

Available at: <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>

Karanam, S. (2022) Exploratory data analysis — breast cancer Wisconsin (diagnostic) dataset, Medium.

Available at: <https://medium.com/@shashmikaranam/exploratory-data-analysis-breast-cancer-wisconsin-diagnostic-dataset-6a3be9525cd>