Module Name & Name of the Module

EBA5005 Practice Module in Graduate Certificate in Specialized Predictive Modelling & Forecasting

Title of Proposal

**Automated Classification of Fine Needle Aspiration Biopsy for Breast Cancer Diagnosis**

File name of the Document

EBA5005\_Group2\_V2

Date of Proposal

10 September 2020

Team Name

Team 1MKBF

Team Members

Goh Pek Han Sam (E0384309)

Lee Suan Ching (E0384993)

Norvin Chandra (E0384966)

Tan Zhi Yang (E0384308)

1. Introduction

Breast cancer is one of the most common cancers among women worldwide, representing about 25 percent of all cancers in women worldwide, making it a significant public health problem today.

For diagnosis of breast cancer, patients will first go through screening test and to ascertain the condition, biopsy of breast mass will have to be conducted on suspected patients. Early diagnosis can allow patients to receive timely treatment and thus, improve their chance of survival significantly. Furthermore, accurate classification of benign tumours can prevent patients from undergoing unnecessary treatments.

1. Industry Overview

According to publication by Gleneagles Hospital Singapore, almost 1 in 16 women in Singapore will be diagnosed with breast cancer in their lifetime. If detected early, success rate of treatment is higher.

Currently breast cancer is detected through the following means (Gleneagles Hospital, n.d.):

1. Clinical examination
2. Mammogram
3. Ultrasound scan
4. MRI

Mammogram is commonly used as compared to other methods due to its accuracy and cost for initial detection of breast cancer (Healthline, 2017). MRI is more expensive, while clinical examination is less accurate and ultrasound scan is of lower specificity than mammogram (Conway Medial Center, n.d.).

When presence of cancer is suspected, biopsy has to be performed to confirm the malignancy of the tumour and the common biopsy techniques are (Gleneagles Hospital, n.d.):

1. Fine Needle Aspiration (FNA)
2. Core needle or tru-cut biopsy
3. Excision biopsy

FNA is a common procedure done by primary care providers and it is the technique of least invasive amongst the biopsy techniques. However, this method has low yield with sensitivity for malignancy at 64% for one aspiration sample. The sensitivity improves to 91% for patients with 3 samples taken which implies the need for multiple FNA procedures for 1 patient (Casaubon et al., 2020). This leads to repetitive stress and increased risk of complication from FNA procedure.

As such, there is a need to improve the accuracy of FNA to reduce the repetitive test and ineffective use of resources. With that, the classification of patients into malignant or benign groups, using FNA data is the subject of this project.

1. Business Problem & Objectives

We are a team of data scientists from Health Tech engaged by Cancer Research Centre to develop automated predictive modelling tool to improve breast cancer diagnosis using Fine Needle Aspiration (FNA).

Main Business Problem

* To develop a cost-effective predictive model to help classify the FNA data to improve the sensitivity of a single sample
* The model developed will then be subject to further study, clinical trial and fine tuning before actual implementation.
* The eventual model assisted FNA (enhanced FNA) will be compared against existing diagnostic techniques to determine optimal market positioning before roll-out. Study will be conducted using Conjoint Analysis.

1. Project Design

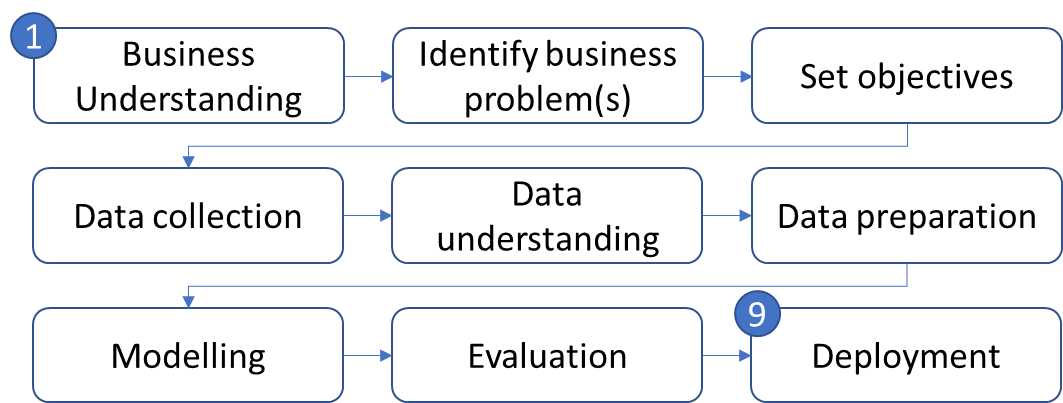


Figure 1: Proposed road map to address the business problems

Proposed road map as follows:

1. Business understanding
2. Identify business problem(s)
3. Set objectives
4. Data collection
5. Data understanding
6. Data preparation
7. Modelling
8. Evaluation
9. Deployment
10. Scope of work

a. Classification of fine needle aspiration biopsy for breast cancer diagnosis

Data set – Breast Cancer.csv

Source of data – <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>

Type of data – csv file

Potential data preparation

* Data cleaning – dealing with missing values, outliers and resolve inconsistencies
* Data exploration
* Data preparation
* dimensionality reduction – principal component analysis, factor analysis, linear discriminant analysis, feature engineering, feature selection
* sampling to deal with imbalanced datasets
* splitting data into training and evaluation sets

Proposed models & techniques

1. Logistics regression
2. Decision treeSupport vector machine
3. K Nearest Neighbours
4. Neural network

Evaluation methodologies

* Confusion Matrix – accuracy, sensitivity etc.
* Gain and Lift charts
* ROC chart

b. Choice Based Conjoint analysis for Product Development

* Identify product attributes and levels
* Design the survey
* Collect data
* Results analysis – Partworth functions and Attribute trade-offs

1. Key Deliverables

* Automate classification of patients into malignant or benign groups based on features computed from digitized images of FNA of breast mass
* Interpret model results
* Compare results between various classification methods and with existing results by other people
* Conjoint analysis - design a survey to determine optimal market positioning for enhanced FNA
* Conjoint analysis - Analyse results
* Discuss the work done and its implications
* Identify the limitations and how to address them in future

1. Effort Estimates and Timeline

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| S/N | Activity | Week1 | Week2 | Week3 | Week4 | Week5 | Week6 |
| (21-Sep  to 27-Sep) | (28-Sep  to 04-Oct) | (05-Oct  to 11-Oct) | (12-Oct  to 18-Oct) | (19-Oct  to 25-Oct) | (26-Oct  to 31-Oct) |
| 1 | Business Understanding |  | + identify business problems and set objectives | | | | |
| 2 | Data Understanding |  | |  |  |  |  |
| 3 | Data Preparation |  |  |  |  |  |  |
| 4 | Modelling |  |  | Each member to work on one method | |  |  |
| 5 | Evaluation |  |  |  |  |  |  |
| 6 | Attributes identification |  |  |  |  |  |  |
| 7 | Design of survey |  |  |  |  |  |  |
| 8 | Survey data collection |  |  |  |  |  |  |
| 9 | Conjoint evaluation |  |  |  |  |  |  |
| 9 | Presentation |  |  |  |  |  |  |

Figure 2: Project Gantt Chart

1. References

Casaubon, J. T., Sandra, T. H. & Regan, J. P. (2020). Fine Needle Aspiration of Breast Masses. Retrieved from <https://www.ncbi.nlm.nih.gov/books/NBK470268/>

Conway Medial Center. (n.d.). Is an Ultrasound Better Than a Mammogram?. Retrieved from <https://www.conwaymedicalcenter.com/news/ultrasound-vs-mammogram-faqs#:~:text=Mammogram%3A%20The%20Key%20Differences.,converts%20them%20to%20an%20image>.

Gleneagles Hospital. (n.d.). Breast Cancer. Retrieved from <https://www.gleneagles.com.sg/facilities-services/centre-excellence/cancer-care/breast-cancer>

Healthline. (2017). What Mammogram Alternatives Are Available and Do They Work? Retrieved from <https://www.healthline.com/health/womens-health/mammogram-alternatives>

UCI Machine Learning. (2016). Breast Cancer Wisconsin (Diagnostic) Data Set.

Retrieved from <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>

1. Appendix – Metadata

|  |  |  |
| --- | --- | --- |
| Variables | Description | Data type |
| Id | Patient ID | Numeric |
| Diagnosis | State whether the tumor is malignant or non-malignant | Categorical |
| radius\_mean | mean of distances from center to points on the perimeter | Numeric |
| texture\_mean | Mean texture of breast tissue | Numeric |
| perimeter\_mean | Mean perimeter of breast tissue | Numeric |
| area\_mean | Mean area of breast | Numeric |
| smoothness\_mean | Mean smoothness of breast | Numeric |
| compactness\_mean | Mean compactness of breast tissue | Numeric |
| concavity\_mean | Mean concavity of breast | Numeric |
| concave points\_mean | Mean concavity points of breast | Numeric |
| symmetry\_mean | Mean symmetry of breast | Numeric |
| fractal\_dimension\_mean | Mean fractional dimension of breast | Numeric |
| radius\_se | Radius of breast standard error | Numeric |
| texture\_se | Texture of Breast standard error | Numeric |
| perimeter\_se | Perimeter of Breast standard error | Numeric |
| area\_se | Area of Breast standard error | Numeric |
| smoothness\_se | Smoothness of Breast standard error | Numeric |
| compactness\_se | Compactness of Breast standard error | Numeric |
| concavity\_se | Concavity of Breast standard error | Numeric |
| concave points\_se | Concavity points of Breast standard error | Numeric |
| symmetry\_se | Symmetry of Breast standard error | Numeric |
| fractal\_dimension\_se | Fractional Dimension of Breast standard error | Numeric |
| radius\_worst | Radius of breast mean of three largest values | Numeric |
| texture\_worst | Texture of Breast mean of three largest values | Numeric |
| perimeter\_worst | Perimeter of Breast mean of three largest values | Numeric |
| area\_worst | Area of Breast mean of three largest values | Numeric |
| smoothness\_worst | Smoothness of Breast mean of three largest values | Numeric |
| compactness\_worst | Compactness of Breast mean of three largest values | Numeric |
| concavity\_worst | Concavity of Breast mean of three largest values | Numeric |
| concave points\_worst | Concavity points mean of three largest values | Numeric |
| symmetry\_worst | Symmetry of Breast mean of three largest values | Numeric |
| fractal\_dimension\_worst | Fractional Dimension of Breast mean of three largest values | Numeric |