FRONT SHEET
Individual Coursework

CANDIDATE NUMBER (C-NUMBER)	C2096237
MODULE NAME	Data and decision making
WORD COUNT	
SUBMISSION DATE	

DECLARATION

I certify that this assessment submission is entirely my work and I have fully referenced and correctly cited the work of others, where required. I also confirm the contents of my submission have not been generated by a third party, or through an Artificial Intelligence generative system*.

I have read the Student Discipline Regulations (Student Discipline Regulations) and understand any Assessment Related Offence/ Academic Misconduct may result penalties being applied.

By submitting this assessment submission, I am confirming that I am fit to sit according to the Assessment Regulations.

I declare that:

- This is my own unaided work.
  Yes ☒
  No ☐
- The word count stated by me is correct.
  Yes ☒
  No ☐
- I'm happy for my work to be retained on the Elite repository and made available to staff and future students**
  Yes ☒
  No ☐

*Please note that all the assignments are submitted to Turnitin.
**Please note personal information (such as names) will be deleted.

Instructions to candidates:

1. **Please complete this cover sheet by entering your Candidate Number, Module Name, Word Count, and Submission Date.**
2. **You must NOT use your NAME on this cover sheet or on any part of your coursework.**

Executive Summary

M. Ruvina S. Fernando
MSc (Hons) in Computing

Contents

Introduction

In the healthcare industry, data analytics plays a pivotal role in improving patient outcomes, enhancing operational efficiencies, and enabling informed decision-making. One area where data analytics has shown remarkable impact is in the early prediction and diagnosis of life-threatening conditions like breast cancer. Breast cancer remains a significant global health challenge due to its prevalence and the potential severity of its complications if not detected early.

This assignment focuses on leveraging machine learning: a leading data analytics tool, to predict the likelihood of breast cancer and classify tumors as malignant or benign. Datasets such as the Wisconsin Breast Cancer Dataset containing critical clinical features like clump thickness, uniformity of cell size, mitoses, and other cytological parameters as well as the biopsy dimensions.

Three machine learning algorithms: Logistic Regression, K-Nearest Neighbors, and Decision Trees will be employed to analyze the input features and predict the malignancy of breast masses. These predictive models are invaluable for assisting healthcare providers in diagnosing breast cancer earlier, leading to timely interventions, improved patient outcomes, and reduced treatment costs.

By integrating machine learning into breast cancer prediction, the process of tumor classification can be automated, increasing diagnostic accuracy and scalability, particularly in settings with limited medical resources. This approach not only enhances the efficiency of breast cancer screening but also supports the global goal of reducing cancer-related mortality through early detection and precise diagnosis.

M. Ruvina S. Fernando
MSc (Hons) in Computing

1. Business Decision Supported by the Solution

This solution supports a business decision to improve early detection and classification of biopsied breast masses or tissue samples as either malignant or benign based on their dimensions and related clinical data. By leveraging machine learning models, healthcare providers can make more accurate and faster diagnoses, ensuring timely and appropriate treatments for patients.

The ability to classify biopsied masses accurately helps avoid unnecessary treatments for benign cases while prioritizing critical care for malignant ones. Early detection of malignancies enables more effective treatment options, reducing the severity of disease progression and the overall cost of cancer management. For benign cases, it minimizes patient anxiety, invasive procedures, and unnecessary healthcare expenses.

Integrating machine learning into the classification of biopsied breast masses aligns with the broader goal of improving diagnostic accuracy and operational efficiency in the healthcare industry. It allows healthcare providers to focus resources on high-risk patients while ensuring that those with benign conditions are spared from overtreatment. This approach supports better patient outcomes, reduces the burden on healthcare systems, and promotes cost-effective, scalable, and patient-centric cancer care.

2. Critical Evaluation of Data Sources

i. Dataset for training the breast cancer detection model from tissue samples collected.

This Breast Cancer Wisconsin dataset is sourced from tissue samples collected for diagnostic purposes, with attributes derived from microscopic examination. The features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass [1] and stored in electronic servers.

Although this dataset is from an earlier period, the diagnostic criteria for breast cancer remain largely consistent with current practices, although treatments have advanced significantly. This dataset was chosen for its inclusion of critical features essential for breast cancer identification and its substantial sample size, which provides a solid foundation for training models.

This dataset consists 699 samples each of 10 features and the class (either 4 for Malignant/cancerous or 2 for benign/non-cancerous). A brief description of the feature attributes is given below. Note that each attribute as a value between 1-10.

Feature	Description
Sample code number	Unique identifier for each tissue sample.
Clump Thickness	Refers to the thickness or density of cell clusters. In benign tumors, cell clusters tend to be smaller, more uniform, and well-defined and clumps are generally less dense and exhibit more orderly arrangements. In malignant tumors, cells form irregular, larger, and denser clusters and clumps may show overlapping or disorganized arrangements, indicating uncontrolled cell growth. [1]
Uniformity of Cell Size	Refers to the consistency or variability in the size of cells within a tissue sample In benign tumors, cells tend to be uniform in size, appearing similar to one another. in malignant tumors, cells show significant variability in size, a condition known as anisocytosis. [2]
Uniformity of Cell Shape	Refers to the consistency or variability in the shape of cells within a tissue sample In benign tumors, cells generally have uniform and regular shapes. In malignant tumors, cells often display irregular and varied shapes, a condition known as pleomorphism. [2]

Marginal Adhesion	Refers to the attachment or binding of tumor cells to the surrounding tissue or to each other at the edges (margins) of a tumor. In benign tumors they typically show well-defined and cohesive margins Cancerous tumors tend to exhibit poor or irregular marginal adhesion. [3]
Single Epithelial Cell Size	Refers to the size of individual epithelial cells in a tissue sample. Epithelial cells line the surfaces of organs and structures in the body In benign tumors tend to be uniform in size and shape. Malignant tumors show significant variability in size, which is referred to as anisocytosis. [4]
Bare Nuclei	presence of nuclei that are not surrounded by a cytoplasm In benign breast tumors, the presence of bare nuclei is rare Bare nuclei are more commonly found in malignant tumors [2]
Normal Nucleoli	refers to chromatin (the material that makes up chromosomes) that appears uniform and non-course In benign tumors, chromatin is typically smooth, evenly distributed In malignant tumors, chromatin tends to become irregular [5]
Mitoses	Refers to the frequency and appearance of mitoses in a tissue sample. In benign breast tumors, mitoses are typically rare and not as frequent. In malignant breast tumors, mitoses are often abundant and occur at a higher rate [6]

ii. Dataset for training the breast cancer detection model from biopsied masses

This Breast Cancer Wisconsin dataset is sourced from biopsied breast mass samples collected for diagnostic purposes, with attributes derived from microscopic examination. The features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass [1] and stored in electronic servers. Same as the previous dataset, although this dataset is from an earlier period, the diagnostic criteria for breast cancer remain largely consistent with current practices, although treatments have advanced significantly.

This dataset includes total of 569 malignant (cancer) and benign (not cancer) breast masses data and the features in this dataset correspond to properties of cell nuclei, such as size, shape and regularity. The mean, standard error, and worst value of each of 10 nuclear parameters is reported for a total of 30 features. [8]. So, not as in the previous dataset, this dataset includes only the certain dimensions on the biopsied breast masses so that no need of detailed explanation of what the property is.