UNIVERSITY OF MALAYA

WQD 7001 Principles of Data Science

Session 2023/2024 Semester 1 - Group Assignment Part 2

Improving Breast Cancer Diagnosis:

Data-Driven Insights and Predictive Analysis

Group 7 - 5A's

Name & Matric No.	Role
WONG KAI THUNG (17101556)	Leader
GUNTUPU SOWJANYA (22085697)	Maker
XIAO YUE (22110867)	Oracle
WANG PENGXIANG (22105619)	Detective
TAN FONG YI (22108436)	Presenter

Lecturer

Dr. Rohana Mahmud

Table of Contents

Project Background

Breast cancer is the most prevalent cancer in the world with more than 4 million US women with a history of invasive breast cancer being alive on January 1, 2022.[1] According to the American Cancer Society, it is the second leading cause of cancer death in women. (First is lung cancer due to its low mortality rates.) The chance of a woman dying from breast cancer in the United States is about 1 in 39 (about 2.5%). On the other hand, there's only about 1% of all breast cancer cases occur in men.[2]

The development of breast cancer arises from the abnormal growth of cells in breast tissue, often identified as a tumour. A tumour can be benign(not cancerous) or malignant (cancerous). [3] To determine if a tumour is malignant or benign, a breast biopsy is required to remove tissue or fluid from the suspicious breast area. Fine-needle aspiration(FNA) is one of the common types of biopsies that remove fluid from the breast lump. The fluid samples are then examined under a microscope by a pathologist, looking for abnormal or cancerous cells. [4] This process is known as tumour grading.

Pathologists utilise tumour grading to categorise malignant breast cancer tumours according to the severity of mutations and the probability of spreading. Examination of breast cancer cells is performed under a microscope to assess factors such as the histologic grade (degree of resemblance to healthy cells), nuclear grade (shape and size of tumour cells' nuclei), and the rate of cell division and multiplication. This systematic evaluation aids in understanding the characteristics and potential aggressiveness of the cancer. [5]

However, since the Microscopic examination of FNA results are highly operator dependent [6], the tumour grading result is prone to bias and inaccuracy. Its reproducibility has been the subject of debate for decades and the inter- and intra-observer variation has been extensively reported. [7]

Over the past few years, machine learning has emerged as a transformative force in the healthcare industry, changing the way how medical data is conventionally analysed and interpreted. By utilising advanced algorithms and computational techniques, machine learning has been instrumental in predicting various healthcare outcomes. One prominent application is

in predictive analytics, where machine learning models are employed to forecast diseases such as diabetes[8], cancer[9], cardiovascular disease[10], mental health disorder[11] and etc.

Therefore, by harnessing the power of machine learning, we hope to improve the accuracy of breast cancer diagnosis. These algorithms can assist healthcare professionals in making more accurate and timely assessments of breast cancer risk, distinguishing between the benign and malignant tumours. This, in turn, will enable patients to receive treatment earlier due to the more timely and accurate diagnosis result.

Project Objective

1) **To Identify Features Associated with Breast Cancer Diagnosis**

   Conduct an in-depth analysis to find features that may be relevant and associated with breast cancer diagnosis. Prioritize the feature with a correlation or the ability to recognize benign and malignant tumours.

2) **To Develop a Predictive Model for Binary Classification of Breast Tumours**

   Build a predictive model using classification algorithms such as logistic regression, decision trees and random forests to accurately classify breast tumours as either malignant or benign based on the features provided in the dataset.

3) **To Compare and Evaluate Different Models for Breast Cancer Classification**

   Assess and compare the performance of diverse classification models. The evaluation will focus on metrics such as accuracy, precision, recall, and F1 score to determine the most effective model for accurately classifying breast tumours using the dataset.

Data Modelling

The aim of this project is to address the pressing issue of inconsistent tumour grading in breast cancer diagnoses by leveraging machine learning. Consequently, we intend to develop supervised machine learning models designed to predict breast cancer diagnoses by utilising identified features. The implementation of these models will be carried out using the Python programming language, along with key libraries and packages such as NumPy, Pandas, OS, Matplotlib, Seaborn, Plotly Express, SciPy, and Scikit-learn.

Machine Learning Algorithms

The dataset selected for constructing the machine learning models originates from The UCI Machine Learning Repository and is entitled "Breast Cancer Wisconsin (Diagnostic)" [12]. Comprising 569 instances of breast tumour diagnoses, the dataset encompasses 32 features, encompassing patient identifiers, diagnostic outcomes, and attributes related to the cell nuclei characteristics identified within digitised images of fine needle aspirates (FNAs) of breast masses.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	rad
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	

5 rows × 32 columns

Figure 1 : Structure of the dataset

The target variable for the machine learning model is the breast tumour diagnosis. This variable is categorical and binary, comprising two distinct categories: benign (non-cancerous) and malignant (cancerous). Our approach involves the construction of classification algorithms within the framework of supervised machine learning to accurately predict the breast tumour diagnosis, classifying instances into either benign or malignant categories.

In pursuit of the development of a robust and high-performance predictive model, we will systematically train and evaluate three distinct classification algorithms. The selected algorithms, listed below, are designed to effectively predict breast cancer diagnoses.

1. **Logistic Regression**: a supervised machine learning algorithm specifically designed for binary classification tasks. It's used to predict the probability of an instance belonging to a specific category. In the context of this dataset, the two potential predicted outcomes are represented as benign (denoted as 0, corresponding to negative instances) and malignant (denoted as 1, corresponding to positive instances). The algorithm estimates the likelihood of a breast tumour falling into either category based on its input features.

2. **Decision Tree**: a supervised learning algorithm used for both classification and regression tasks. It is a flowchart-like tree structure where each internal node denotes the feature, branches denote the rules, and the leaf nodes denote the result of the algorithm. The algorithm makes decisions by recursively splitting the data based on the feature that provides the most information about the target variable. At each node, a decision is made by evaluating the values of a particular feature, leading to further nodes or branches until a final decision is reached at a leaf node.

3. **Random Forest**: a specific type of ensemble machine learning algorithm designed for both classification and regression tasks. It operates by combining the outcomes of numerous decision trees to produce a singular and more robust result. This is accomplished through the construction of an ensemble of decision trees during the training phase of the algorithm. Each decision tree is trained on a distinct subset of the dataset, and their collective predictions are used to make a final decision, often through a voting mechanism in classification tasks.

Test Harness

For the purpose of training and evaluating the performance of the trained models, the dataset will undergo a random split, allocating 70% of instances (398 instances) to the training data and the remaining 30% (171 instances) to the testing data. This will ensure that the training data is exclusively employed for model training, while the testing data serves as an independent set for assessing the accuracy and generalizability of the models.

Evaluation

The efficacy of the machine learning classification models will undergo evaluation based on the following performance metrics:

1. **Accuracy:** A measure of the overall correctness of the model, calculated as the ratio of correctly predicted instances to the total instances.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{ False Negative} + \text{True Negative}}$$

2. **Precision and Recall:** Precision and recall are complementary metrics. Precision quantifies the accuracy of positive predictions, indicating the proportion of correctly predicted positive instances among all instances predicted as positive. Recall, on the other hand, assesses the model's ability to capture all positive instances, representing the proportion of correctly predicted positive instances among all actual positive instances.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad , \quad \text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

3. **F1-Score:** The F1-Score is the harmonic mean of precision and recall, providing a balanced assessment of the model's performance. It is particularly useful when there is an uneven distribution between positive and negative instances.

$$\text{F1 Score} = \frac{2 \text{ x } (\text{Recall x Precision})}{\text{Recall} + \text{Precision}}$$

These performance metrics collectively serve to offer a comprehensive evaluation of the classification models, ensuring a nuanced understanding of their strengths and limitations. The model deemed most effective based on these metrics will undergo further development into data products for deployment to the target user demographic.

Machine learning interpretability

Upon evaluating and selecting the preferred model for predicting breast cancer diagnoses, our subsequent step involves interpreting and explaining the model's predictions. This will be achieved by determining the feature importance scores through the utilisation of SHAP values (SHapley Additive exPlanations). Employing a game-theoretic approach, SHAP values measure the contribution of each "player" (feature) to the final outcome.

In the context of machine learning, SHAP values allocate an importance value to each feature, representing its contribution to the model's output. This interpretability tool enhances our understanding of the model's decision-making process, shedding light on the influence and significance of individual features in shaping the predictions related to breast cancer diagnoses.

Data Interpretation

The predicted outcomes of the models trained with Logistic Regression, Decision Tree, and Random Forest algorithms are visually represented by the Confusion Matrix in Figures 2, 3, and 4, respectively. The Confusion Matrix serves as a comprehensive display of the model's performance by illustrating the counts of true positive, true negative, false positive, and false negative predictions in comparison to the actual outcomes of the testing dataset.

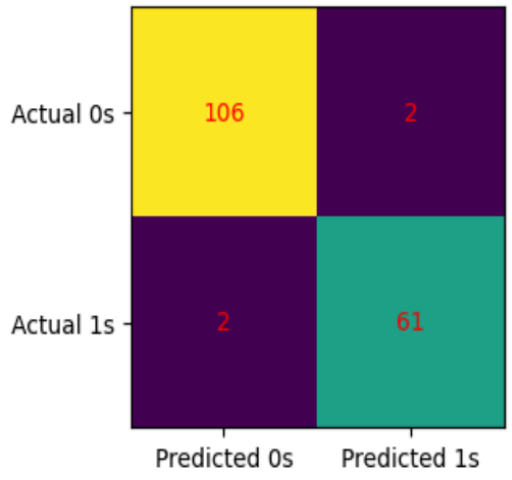

Figure 2 : Confusion matrix of Logistic Regression model

The Logistic Regression model demonstrates accurate predictions in 106 instances of malignant tumours (True Positives) and 61 instances of benign tumours (True Negatives). However, there are 2 instances where the model inaccurately predicts malignant tumours (False Positives) and 2 instances where it inaccurately predicts benign tumours (False Negatives). These details are indicative of the model's performance in correctly identifying and classifying instances of malignant and benign tumours.

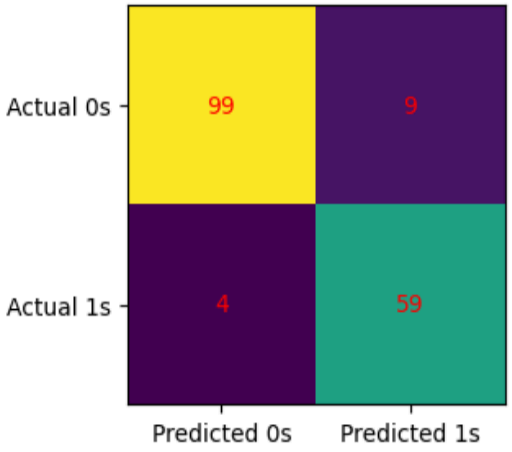

Figure 3 : Confusion matrix of Decision Tree classifier model

The Decision Tree classifier model exhibits a relatively less robust performance when compared to the Logistic Regression model. It accurately predicts 99 instances of malignant tumours (True Positives) and 59 instances of benign tumours (True Negatives). However, there are 9 instances where the model makes inaccurate predictions of malignant tumours (False Positives) and 4 instances where it inaccurately predicts benign tumours (False Negatives).

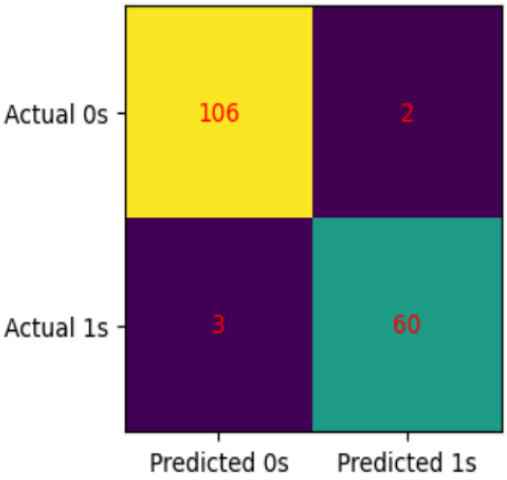

Figure 4 : Confusion matrix of Random Forest classifier model

The Random Forest classifier model outperforms the Decision Tree classifier model but exhibits a less robust performance compared to the Logistic Regression model. Similar to the Logistic Regression model, the Random Forest model accurately predicts 106 instances of malignant tumours (True Positives), but it achieves only 60 accurate predictions of benign tumours (True Negatives). There are 2 instances where the model makes inaccurate predictions for malignant tumours (False Positives) and 3 instances where it inaccurately predicts benign tumours (False Negatives).

Table 1 presents the performance metrics of the trained predictive models. Logistic Regression model exhibits a high accuracy of 97.6%, surpassing the accuracy of both the Random Forest model (97%) and the Decision Tree model (92.39%), indicating that it correctly predicts breast cancer diagnoses for a large proportion of instances. The F1-score is highest for the Logistic Regression model (0.968), suggesting a good balance between precision and recall. Precision and recall metrics for the Logistic Regression model are both 0.968, indicating a high proportion of correctly identified positive instances (precision) and capturing a large proportion of actual positive instances (recall). Random Forest exhibits a precision of 0.968 and recall of 0.952, while the Decision Tree model has lower precision (0.868) and higher recall (0.937). This concludes that the Logistic Regression model consistently outperforms the Random Forest and Decision Tree models across all metrics, making it the preferred model for predicting breast cancer diagnoses in this scenario.

Table 1: Performance metrics of the three trained predictive models

Models	Accuracy	F1-score	Precision	Recall	Balanced Accuracy
Logistic Regression	0.976608	0.968254	0.968254	0.968254	0.974868
Random Forest	0.970760	0.960000	0.967742	0.952381	0.966931
Decision Tree	0.923977	0.900763	0.867647	0.936508	0.926587

Figure 5 presents a summary plot illustrating the average SHAP values of each feature on the Logistic Regression model outputs. The outcomes of the analysis highlight that certain features, including "Perimeter Mean", "Area Mean," "Area Worst", "Area Standard Error", "Radius Mean", "Texture Worst", "Perimeter Worst", and "Texture Mean" of cell nuclei, play pivotal roles in influencing the output of breast cancer diagnosis. The prominence of these features, as indicated by their substantial average SHAP values, underscores their significant contributions to the model's predictions. These findings contribute valuable insights into the specific characteristics of cell nuclei that strongly impact the Logistic Regression model's ability to distinguish between benign and malignant breast cancer cases.

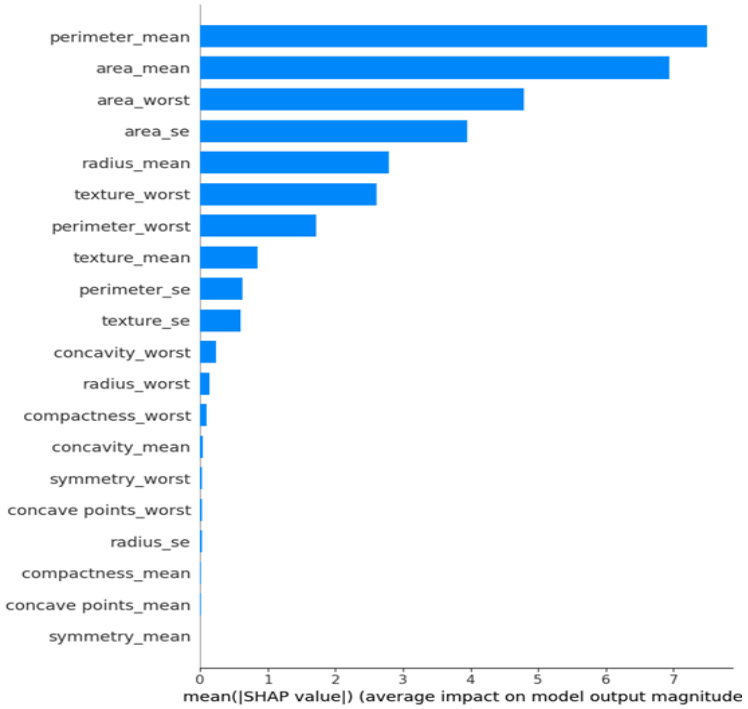

Figure 5 : Average SHAP Values: Logistic Regression Model on Breast Cancer Diagnosis

On the other hand, Figure 6 provides a comprehensive summary plot illustrating the SHAP values and the directional impact of each feature on the Logistic Regression model outputs. The plot discerns how individual features contribute either positively or negatively to the likelihood of a breast tumour being classified as malignant or benign. From the observations in the plot, it can be inferred that higher values of "Perimeter Mean," "Area Worst," "Area Standard Error," "Texture Worst," and "Perimeter Worst" correspond to an increased probability of the breast tumour being classified as malignant. Conversely, lower values of these features are associated with a higher likelihood of classification as benign.

Additionally, the plot suggests that lower values of "Area Mean," "Radius Mean," and "Texture Mean" contribute to an elevated chance of the breast tumour being classified as malignant, while higher values of these features are indicative of a higher likelihood of benign classification. These insights underscore the significance of specific feature values in influencing the Logistic Regression model's predictions, offering valuable information for understanding the factors that contribute to the categorization of breast tumours as either malignant or benign.

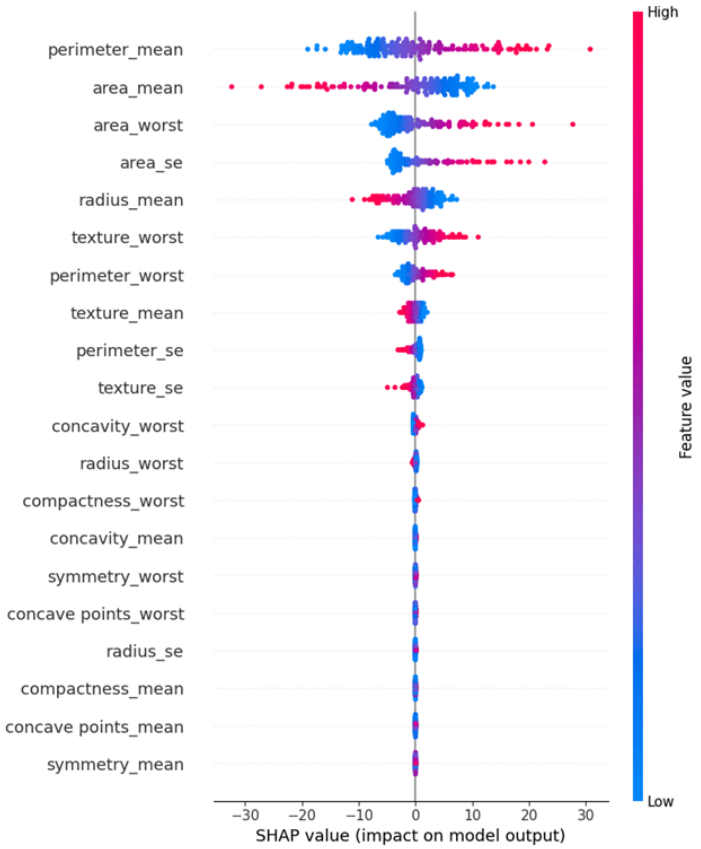

Figure 6 : SHAP Values Impact: Logistic Regression Model on Breast Cancer Diagnosis

Plan for Reproducible Research

An essential component of our project is a well-designed strategy that ensures the legitimacy and dependability of findings through the facilitation of independent verification and validation. Moreover, it promotes openness and cooperation, enabling others to expand upon and faithfully reproduce the findings.Our efforts toward reproducible research are listed below:

1. Documentation and Reporting:

Thoroughly documented and reported every stage of our project. We conducted an in-depth analysis of the methodologies employed, the machine learning models utilised, and the creation of our breast cancer prediction application. Our report provided a comprehensive overview of the project, including its objectives, the valuable insights obtained from data analysis, and the impressive performance of our predictive models. In addition, we discussed potential future improvements to our application, offering a comprehensive perspective on our project's development and its significance in the field of breast cancer diagnosis. We have in-depth documentation of the below steps in this report:

- Data Preprocessing and Exploration
- Exploratory Data Analysis (EDA)
- Model Development and Validation
- Performance Evaluation
- Application Development
- Ethical Considerations:

2. Data Source and Details:

We used the "Breast Cancer Wisconsin (Diagnostic)" dataset from the UCI Machine Learning Repository.

Direct Link: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

This dataset contains 569 instances with 32 features each, obtained from digitised images of fine needle aspirates of breast masses. The features centre around the attributes of the cell nuclei in the images. Its main purpose is to categorise tumours as either malignant or benign, and it is worth noting that it does not contain any missing values.

3. Software and Tools Specification:

Programming Language: Python was used for data analysis and model development.

Packages and Libraries: Different Python libraries were utilised, including NumPy, Pandas, OS, Matplotlib, Seaborn, Plotly Express, SciPy, and Scikit-learn.

Machine Learning Algorithms: The project utilised logistic regression, decision trees, and random forests for predictive modelling.

Code Availability: The code for our breast cancer diagnosis project is available on GitHub at the following repository:

https://github.com/SowjanyaGSVL/Improving-Breast-Cancer-Diagnosis-Data-Driven-Insights-and-Predictive-Analysis

4. Enhancing Collaborative Reproducibility:

Our main goal is to improve collaborative reproducibility through a strong emphasis on transparency and comprehensive documentation of our methods, data, and tools. By following this practice, students and others can easily replicate our study, thus verifying and expanding upon our findings. We keep thorough records of our coding, algorithmic approaches, and analytical processes. Sharing our data and code on open-source platforms like “GitHub” is an important strategy that encourages community involvement and promotes further research. These steps are essential for promoting scientific integrity and driving progress in the field through collaborative efforts.

Deployment of Data Product

The ultimate goal of this project is deployment, in which we have developed an EXE application, a data product that leverages the power of linear regression to predict the likelihood of breast cancer malignancy based on cell nucleus information. This application can assist healthcare professionals and individuals in making informed decisions about potential malignancy.

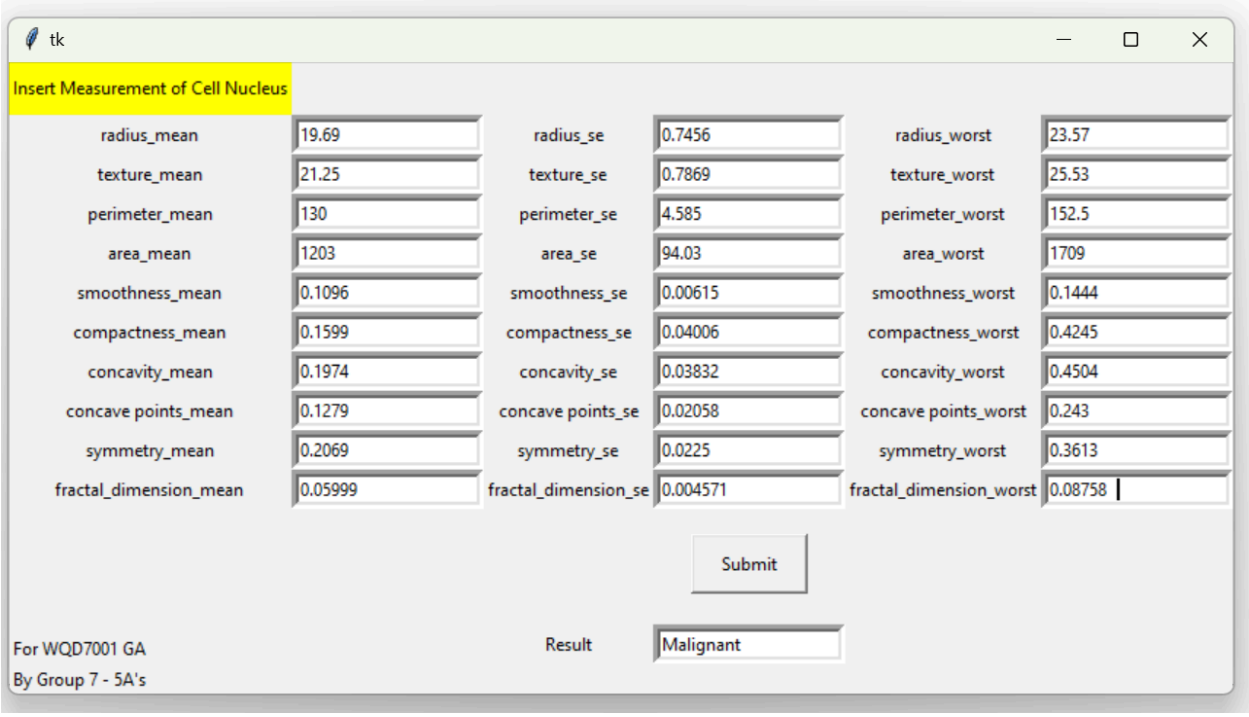

Figure 7: An EXE application that predicts the likelihood of breast cancer malignancy

Development Process

The application was developed with the Python programming language and converted into a standalone executable using PyInstaller, with the graphical user interface (GUI) built using the tkinter library. The user interface allows individuals to input specific cell nucleus measurements conveniently. These measurements, such as radius, texture, perimeter, area, and other relevant features, serve as input parameters for the linear regression model.

The heart of the application lies in its use of the logistic regression model from the scikit-learn library. The model has been trained on a comprehensive dataset to learn the patterns associated with malignant and benign breast tumours. Leveraging the insights gained from the training process, the tool can make predictions when new cell nucleus information is provided.

User Experience

The user experience is designed to be intuitive, with a clear layout featuring input fields for each cell nucleus measurement. Upon entering the required data, users can click the "Submit" button to trigger the linear regression model's predictions. The results are then displayed in the GUI, indicating whether the provided cell nucleus information suggests a malignant or benign diagnosis.

Key Features

i) **Accuracy**: The predictive model's accuracy is a key feature of this data product. The choice of linear regression, after rigorous experimentation, ensures reliable predictions based on the provided cell nucleus information.

ii) **Usability**: The GUI provides a user-friendly experience, making it accessible to healthcare professionals and individuals alike. The straightforward design allows users to input data easily and interpret the results.

iii) **Real-time Prediction**: The application provides real-time predictions, offering a quick assessment of the likelihood of malignancy based on the entered information.

Future Improvement

In pursuit of continuous improvement, we propose a few features that could be implemented on our application in the future. From refining machine learning capabilities to enhancing user-centric features, we aim to elevate the tool's effectiveness, making it a reliable resource for early breast cancer detection and diagnosis.

i) **Integration with Electronic Health Records (EHR)**:
By integrating the application with electronic health record systems, we could streamline the process for healthcare professionals, allowing them to directly input relevant patient data from existing records, fostering better collaboration and data continuity in the medical field.

ii) **Integration with External Databases**:

Explore the possibility of integrating the application with external databases or research studies to continually update and improve the model with the latest advancements in breast cancer research.

iii) **Feedback Mechanism**:

Introduce a feedback loop that allows users to provide feedback on the accuracy of predictions. This user feedback can be used to continually improve and refine the model over time.

iv) **Confidence Intervals**:

Include confidence intervals for predictions to quantify the uncertainty associated with each prediction. This additional information can be valuable, especially in cases where the model might be less certain about a particular diagnosis.

v) **Educational Resources**:

Include educational resources within the application to help users understand how the model arrived at a particular prediction.

vi) **Multimodal Input Support**:

Expand input options to support different data modalities. For instance, allow users to input images of cell nuclei in addition to numerical values. Integrating image processing capabilities could open avenues for incorporating more advanced machine learning models, such as convolutional neural networks, for improved diagnostic accuracy.

Insights and Conclusion

Breast cancer prediction using machine learning, mainly the Logistic Regression model, is a leading healthcare effort. This project has addressed the long-standing challenges related to the diagnosis of breast cancer, moving away from traditional and somewhat subjective methods towards a more objective, data-driven approach. The remarkable accuracy of the Logistic Regression model in distinguishing between benign and malignant tumours is a testament to the immense impact of machine learning on advancing medical diagnostics. It is essential to note the model's exceptional accuracy and reliability, as they are crucial for detecting cancer early and greatly impacting the success of treatment. This technology has the potential to revolutionise medical care by allowing for faster and more precise diagnoses. As a result, patients can receive timely interventions, leading to better treatment outcomes and potentially saving lives.

With the development of a user-friendly application that incorporates the Logistic Regression model, this project has taken a significant step towards practical application. Now, both healthcare professionals and individuals can easily access this advanced technology. This tool equips users with the knowledge to make better decisions, promoting proactive health management. This project showcases the immense potential of machine learning applications in different areas of healthcare, extending beyond its direct impact on breast cancer diagnosis. It highlights the increasing importance of technological advancements in medical diagnostics and paves the way for more advanced, precise, and patient-centred healthcare. The remarkable achievements of this initiative in the field of breast cancer diagnosis serve as a guiding light, paving the way for future advancements in healthcare technology. It showcases the immense possibilities that arise from the fusion of advanced computational methods and medical expertise, leading to improved health outcomes and a transformative impact on medical diagnostics and treatment.

In conclusion, the project effectively showcases the capacity of machine learning models, specifically Logistic Regression, to improve the precision and dependability of breast cancer diagnosis. This progress sets the stage for the integration of these models into clinical environments, ultimately leading to improved patient outcomes.

Reference

[1] Miller, K. D., Nogueira, L., Devasia, T., Mariotto, A. B., Yabroff, K. R., Jemal, A., Kramer, J., & Siegel, R. L. (2022). Cancer treatment and survivorship statistics, 2022. *CA: A Cancer Journal for Clinicians*, *72*(5). https://doi.org/10.3322/caac.21731

[2] American Cancer Society. (2022). *Breast Cancer Facts & Figures 2022-2024*. Atlanta, Georgia.

[3] Patel, A. (2020). Benign vs. Malignant Tumors. In *JAMA Oncology* (Vol. 6, Issue 9). https://doi.org/10.1001/jamaoncol.2020.2592

[4] “Breast Biopsy: Procedure Types, What to Expect & Results Guide.” *National Breast Cancer Foundation*, June 2023, www.nationalbreastcancer.org/breast-cancer-biopsy/.

[5] Bloom, H. J., & Richardson, W. W. (1957). Histological grading and prognosis in breast cancer a study of 1409 cases of which 359 have been followed for 15 years. *British Journal of Cancer*, *11*(3). https://doi.org/10.1038/bjc.1957.43

[6] Giard, R. W. M., & Hermans, J. (1992). The value of aspiration cytologic examination of the breast a statistical review of the medical literature. *Cancer*, *69*(8). https://doi.org/10.1002/1097-0142(19920415)69:8<2104::AID-CNCR2820690816>3.0.CO;2-O

[7] van Dooijeweert, C., van Diest, P. J., & Ellis, I. O. (2022). Grading of invasive breast carcinoma: the way forward. In *Virchows Archiv* (Vol. 480, Issue 1). https://doi.org/10.1007/s00428-021-03141-2

[8] Zafar, F., Raza, S., Khalid, M. U., & Tahir, M. A. (2019). Predictive analytics in healthcare for diabetes prediction. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3326172.3326213

[9] Choudhury, A., & Eksioglu, B. (2019). Using predictive analytics for cancer identification. *IISE Annual Conference and Expo 2019*.

[10] Muniasamy, A., Muniasamy, V., & Bhatnagar, R. (2021). Predictive analytics for cardiovascular disease diagnosis using machine learning techniques. *Advances in Intelligent Systems and Computing*, *1141*. https://doi.org/10.1007/978-981-15-3383-9_45

[11] Hahn, T., Nierenberg, A. A., & Whitfield-Gabrieli, S. (2017). Predictive analytics in mental health: Applications, guidelines, challenges and perspectives. In *Molecular Psychiatry* (Vol. 22, Issue 1). https://doi.org/10.1038/mp.2016.201

[12] Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B