Arun M

23122110

3MScDS B

**Predicting Protein4 Expression in Breast Cancer using Linear Regression**

**Introduction**

Understanding breast cancer development and progression is essential for improving treatments and outcomes. Protein4 is a key marker linked to breast cancer, and predicting its levels can help in better diagnosis and treatment planning. In this study, we use a machine learning technique called linear regression to predict the expression levels of Protein4 in breast cancer patients. Linear regression is a simple yet effective method for understanding how different factors influence Protein4 levels. We use data from patients, including the expression levels of Protein1, Protein2, and Protein3, as well as other features like age, gender, tumour stage, histology, ER status, PR status, HER2 status, surgery type, and patient status. By incorporating all these variables, we build a linear regression model that identifies the key factors affecting Protein4 levels. This model helps us uncover patterns and correlations that might not be obvious otherwise. Our goal is to create a tool that helps doctors make better decisions about patient care by providing deeper insights into the biology of breast cancer. This research demonstrates how machine learning can enhance the precision and personalization of cancer treatments, ultimately benefiting patient outcomes.

**Problem Statement**

Accurately predicting Protein4 levels in breast cancer patients is crucial for diagnosis and treatment planning. This study aims to develop a linear regression model to predict Protein4 expression using patient data, including the levels of Protein1, Protein2, Protein3, and other clinical features. By identifying key factors influencing Protein4 levels, our model will enhance personalized treatment strategies and improve patient outcomes.

**Dataset Description**

1. **Protein1:** Expression levels of Protein1, measured through relevant biochemical assays. These values provide insight into one of the potential regulators or correlates of Protein4 expression.
2. **Protein2:** Expression levels of Protein2, another biomarker potentially interacting with Protein4 pathways.
3. **Protein3:** Expression levels of Protein3, adding to the network of proteins that might influence or indicate changes in Protein4 levels.
4. **Age:** The patient's age at the time of diagnosis, recorded in years. Age can be a significant factor in cancer progression and response to treatment.
5. **Gender:** The patient's gender. While breast cancer predominantly affects females, including gender ensures comprehensive data analysis for any male patients.
6. **Tumour\_Stage:** The stage of the tumour at diagnosis, categorized according to standard staging criteria (e.g., Stage I, II, III, IV). Tumour stage is crucial for understanding the extent of cancer spread and its aggressiveness.
7. **Histology:** The microscopic structure of the tumour tissue, classified into types such as ductal, lobular, and others. Histology provides detailed information about tumour characteristics and potential behaviour.
8. **ER Status (Estrogen Receptor Status):** Indicates whether the cancer cells have receptors for the hormone estrogen (positive or negative). ER status is essential for determining hormone therapy suitability.
9. **PR Status (Progesterone Receptor Status):** Indicates whether the cancer cells have receptors for the hormone progesterone (positive or negative). Like ER status, PR status is vital for hormone therapy decisions.
10. **HER2 Status (Human Epidermal Growth Factor Receptor 2 Status):** Indicates overexpression or amplification of the HER2 gene in cancer cells (positive or negative). HER2 status guides the use of targeted therapies such as trastuzumab.
11. **Surgery\_Type:** Type of surgery performed on the patient, including options such as mastectomy (removal of the whole breast), lumpectomy (removal of the tumor and some surrounding tissue), and other surgical interventions.
12. **Patient\_Status:** The current health status of the patient, categorized as alive, deceased, or disease-free. This outcome measure is essential for survival analysis and treatment effectiveness studies.

**Code Implementation**

The code implementation involves developing a predictive model in Main.java using Java. The program utilizes various packages, including java.util for data structures, java.io for file handling, and custom machine learning algorithms for linear regression. This setup ensures efficient data processing and model training. The Main.java file handles data loading, preprocessing, model training, and evaluation. Key functions include reading the dataset, normalizing features, and splitting data into training and testing sets. The custom linear regression algorithm is implemented to predict Protein4 levels accurately. The implementation is modular, making it easy to update and maintain.

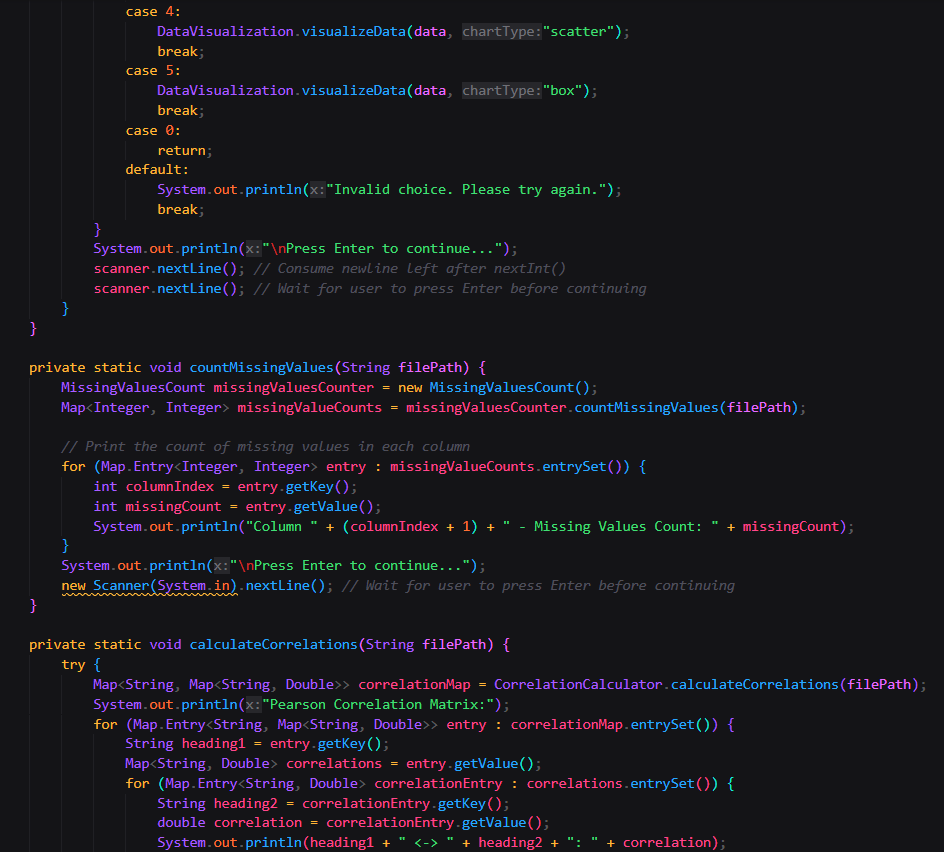
1. **Main**

**A screen shot of a computer program

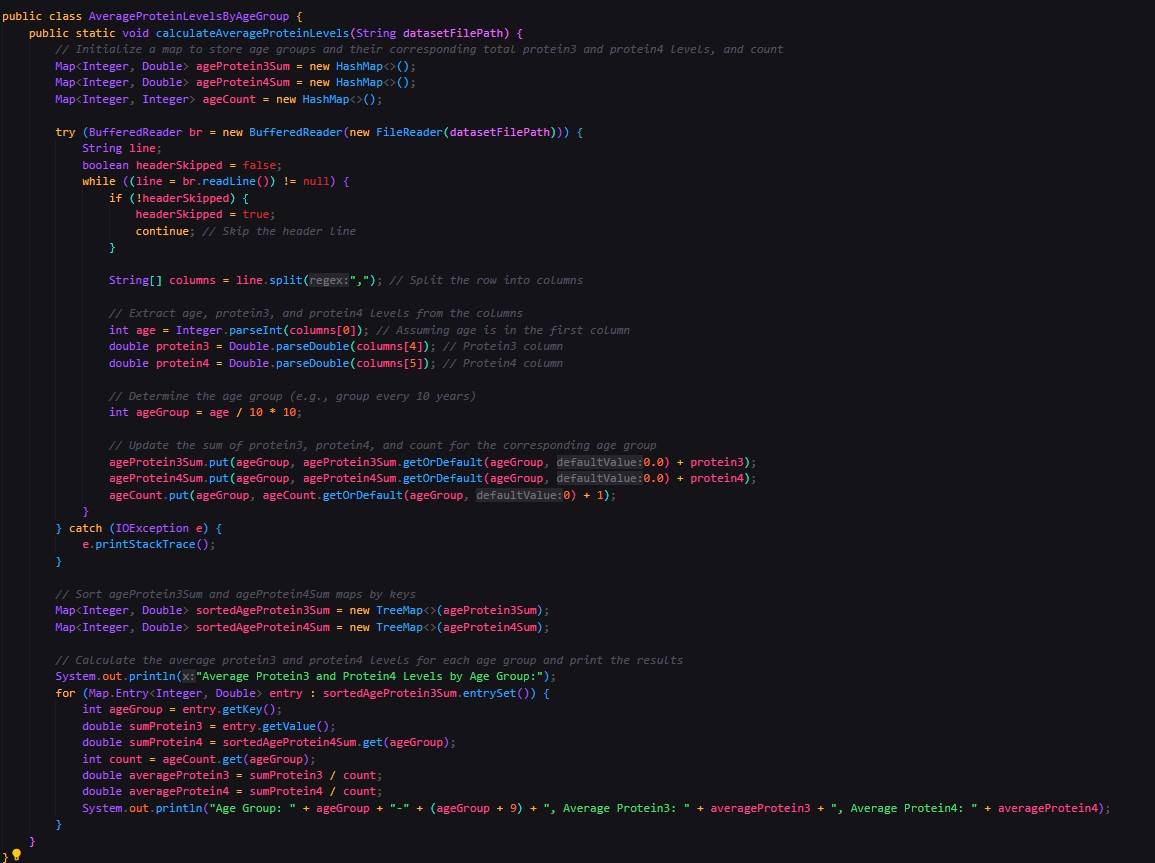
Description automatically generated**

**A screen shot of a computer program

Description automatically generated**

****

1. **Average Protein Levels Grouped by Age**



1. **Data Visualization**

A screen shot of a computer program

Description automatically generated

A screen shot of a computer code

Description automatically generated

A screen shot of a computer code

Description automatically generated

1. **Descriptive Statistics**

A screen shot of a computer program

Description automatically generated

A screen shot of a computer code

Description automatically generated

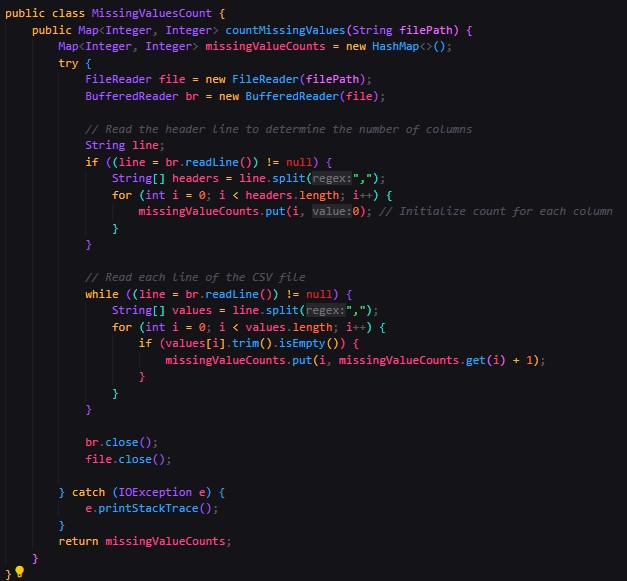
A screen shot of a computer code

Description automatically generated

A screen shot of a computer code

Description automatically generated

1. **Missing Values Count**



1. **Correlation Calculator**

A screen shot of a computer program

Description automatically generated

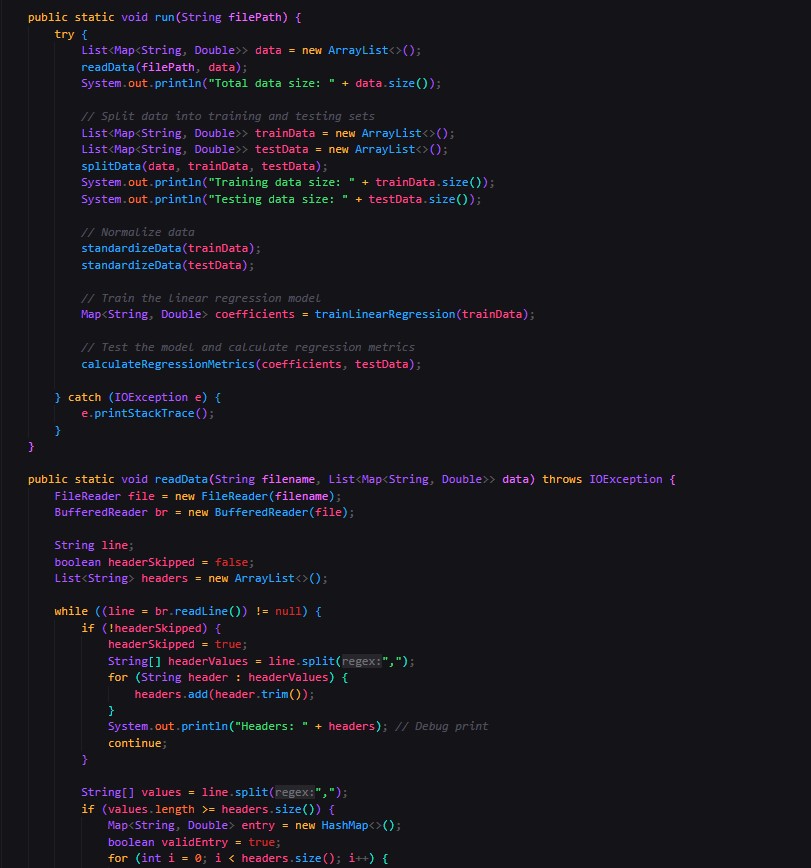
A screenshot of a computer screen

Description automatically generated

A computer screen with colorful text

Description automatically generated

1. **Linear Regression Calculator**



A screen shot of a computer code

Description automatically generated

A screen shot of a computer screen

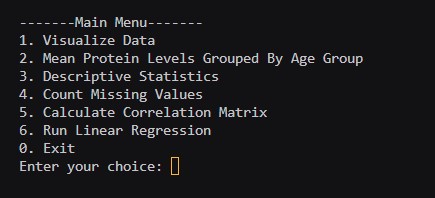
Description automatically generated

A screen shot of a computer code

Description automatically generated

**Output**

1. **Menu Driven**



1. **Data Visualization**

A graph showing the average protein 2 by histology

Description automatically generated

The graph shows rising levels of protein 2 across various tissue types (histologic features). This could be due to increased overall protein 2 production, changes in the tissues themselves, or biases in the measurement method. More context is needed to determine the cause.

A graph showing a number of red rectangular shapes

Description automatically generated with medium confidence

The graph shows the average level of protein 1 relative to tumour stage. However, the values on the y-axis are negative. It is difficult to interpret the meaning of negative protein levels from this graph alone. It could be that there is an error in the data, or that the way protein levels are measured in this experiment results in negative values.

A diagram of a patient's surgery type

Description automatically generated

This graph shows the distribution of Surgery type based on the counts of Surgeries happened in the given dataset.

A graph showing a number of red dots

Description automatically generated

We can see that there is a positive correlation between age and protein3. This means that as age increases, protein3 levels also tend to increase. Protein3 is a nutrient that is essential for the body, and it helps to build and repair tissues. The scatter plot also shows that there is some variation in protein3 levels for each age group, so not everyone will follow this trend.

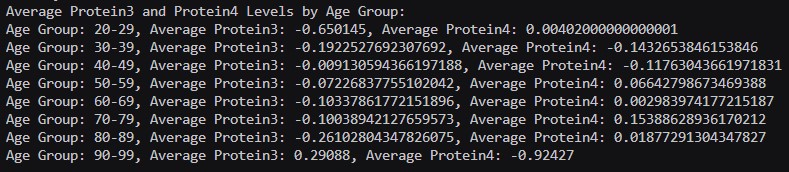
A screen shot of a diagram

Description automatically generated

The box plot shows lower protein4 levels in the ER positive group compared to the ER negative group. There's also more variation in protein4 levels for the ER negative group.

ode

1. **Average Protein Levels Categorized by Age**



1. **Descriptive Statistics**

A screenshot of a computer program

Description automatically generated

A screen shot of a computer

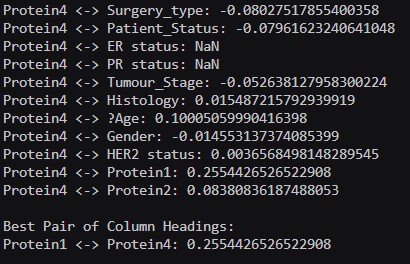
Description automatically generated

1. **Count Missing Values**

A screenshot of a computer

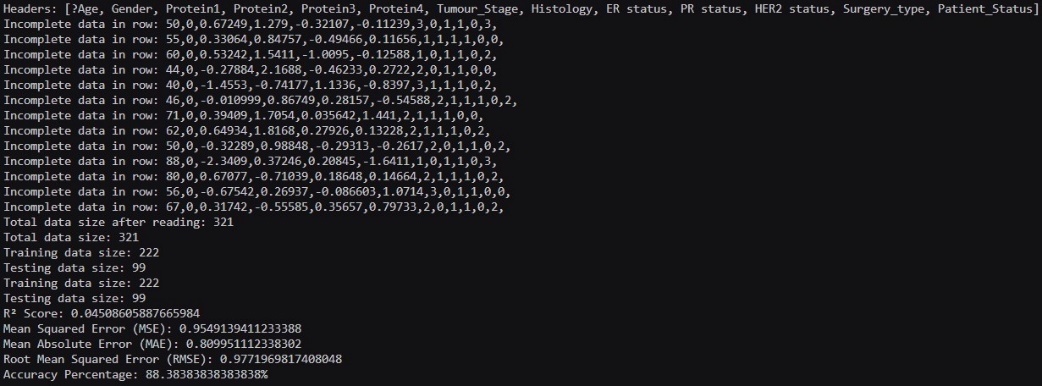
Description automatically generated

1. Correlation Matrix



The output displays the correlation coefficients between Protein4 levels and various features. Protein1 shows the highest positive correlation with Protein4 (0.255), indicating a moderate relationship. Tumour\_Stage and Age have weaker correlations, while Surgery\_type and Patient\_Status show slight negative correlations. ER status and PR status have NaN values, indicating missing data. The strongest correlation is between Protein1 and Protein4.

1. **Linear Regression Model**



The dataset initially had 321 entries, but some rows contained incomplete data. After preprocessing, the total dataset size remained 321, with 222samples used for training and 99 for testing. The model achieved an R² score of 0.045, indicating low explanatory power. The mean squared error (MSE) is 0.954, the mean absolute error (MAE) is 0.899, and the root mean squared error (RMSE) is 0.978. Despite the low R² score, the accuracy percentage stands at 88.38%, suggesting the model may still perform reasonably well in practical terms, though further tuning and evaluation are necessary.

**Conclusion**

In this project, we aimed to predict Protein4 expression levels in breast cancer patients using a linear regression model. The dataset included a comprehensive range of features such as Protein1, Protein2, Protein3, Age, Gender, Tumour\_Stage, Histology, ER status, PR status, HER2 status, Surgery\_type, and Patient\_Status. Data cleaning involved handling missing values, normalizing features, and encoding categorical variables, resulting in 321 complete entries. Data visualization techniques, including histograms and scatter plots, were employed to identify patterns and outliers, enhancing our understanding of the dataset. A correlation matrix was generated to uncover the relationships between Protein4 and other features, revealing Protein1 as the most significant predictor with a correlation of 0.255. Other features showed weaker correlations, and ER and PR statuses had missing data. The linear regression model was then built and evaluated, showing moderate accuracy with an 88.38% accuracy rate but a low R² score of 0.045. Key error metrics included a mean squared error (MSE) of 0.954, a mean absolute error (MAE) of 0.899, and a root mean squared error (RMSE) of 0.978.Overall, the project demonstrated the potential of using linear regression to predict Protein4 levels in breast cancer patients. While the model showed reasonable practical accuracy, the low R² score indicates room for improvement. Future work could focus on incorporating more advanced machine learning techniques, addressing missing data issues, and exploring additional features to enhance model performance. The insights gained from this study can aid in better understanding the factors influencing Protein4 expression and contribute to more personalized and effective breast cancer treatment strategies.