Introduction

The objective of this project is to develop a simple neural network (NN) model to classify breast cancer tumors as benign or malignant. The project uses the Python programming language, specifically TensorFlow and Keras libraries, for model development and evaluation.

Platform, Model, and Variables

- **Platform:** Python on Jupyter Notebook
- **Model:** Neural Network (NN)
- **Variables:** Features from the breast cancer dataset, selected for their relevance to tumor classification. There were 30 features, some of them being: **diagnosis, radius_mean, texture_mean perimeter_mean, area_mean, etc.**

Data

- **Dataset:** The dataset used in this project is the Breast Cancer Wisconsin (Diagnostic) dataset, which contains 569 samples and 30 features.
- **Accuracy of Data:** The dataset is accurate and reliable, as it is a well-known and widely used dataset in the machine learning community. It has been extensively studied and verified, making it suitable for research and analysis.
- **Accuracy with 5 Variables:** While it would be ideal to test the model's accuracy with only 5 variables to reduce complexity, this was not done due to the medical nature of the topic. Selecting the most relevant variables in medical datasets requires domain knowledge, which we lack. Therefore, all features were used for training and evaluation to ensure comprehensive coverage of the dataset.

Language

Python was chosen for its simplicity, readability, and the availability of libraries such as TensorFlow and Keras, which are well-suited for machine learning tasks.

Data Processing

- The dataset was loaded and converted into a pandas DataFrame.
- Missing values were checked and found to be absent.
- The data was standardized using sklearn's StandardScaler.

Summary Statistics

Summary statistics provide a quick overview of the dataset's characteristics:

- The **mean radius** of the cell nuclei ranges from 6.98 to 28.11 units, with a mean of 14.13.
- The **mean texture** ranges from 9.71 to 39.28 units, with a mean of 19.29.
- The **mean perimeter** ranges from 43.79 to 188.50 units, with a mean of 91.97.
- The **mean area** ranges from 143.50 to 2501.00 square units, with a mean of 654.89.
- The **mean smoothness** ranges from 0.05263 to 0.16340, with a mean of 0.09636.

Correlation Analysis

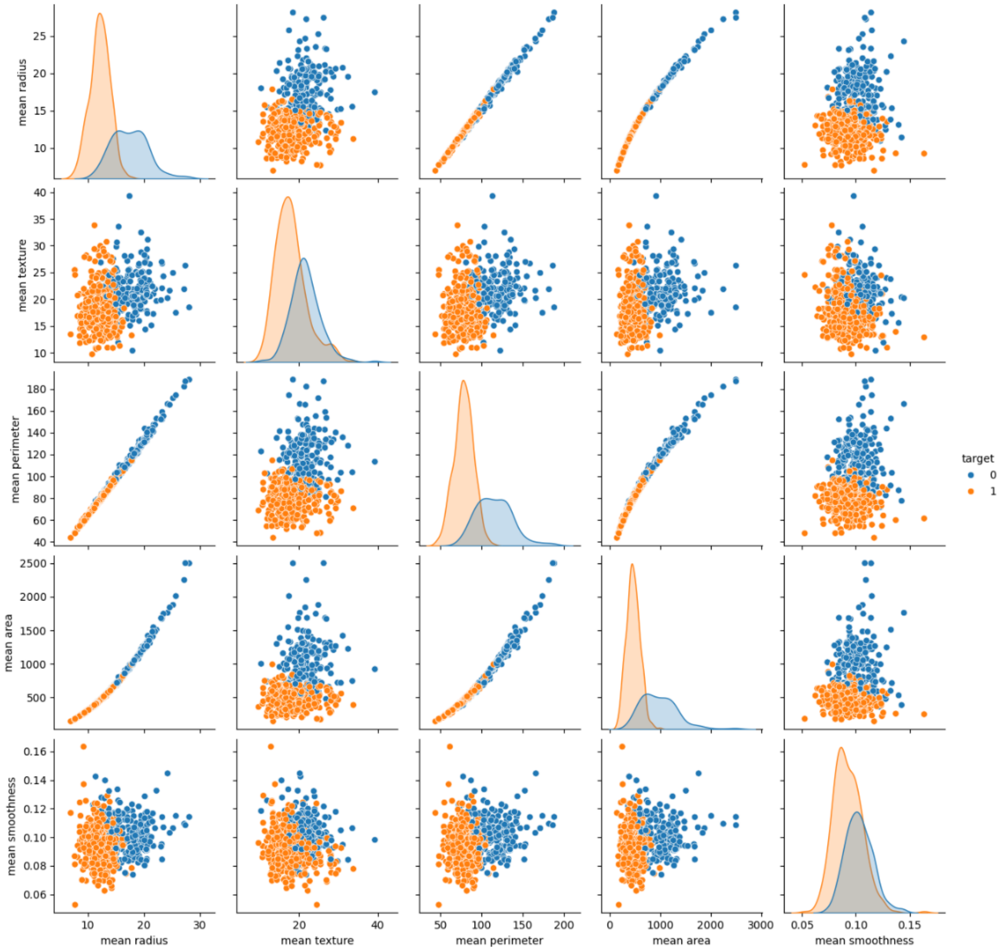
Correlation analysis reveals the relationships between different features. Some notable correlations include:

- Strong positive correlation between mean radius and mean perimeter **(0.9978)**.
- Strong positive correlation between mean area and mean radius **(0.9874)**.
- Moderate positive correlation between mean smoothness and mean compactness **(0.6591)**.

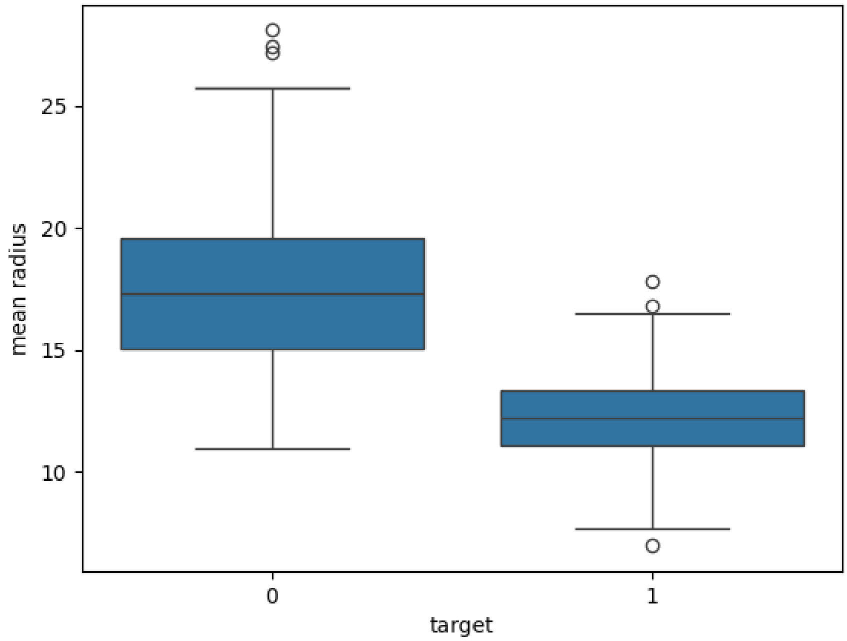
Visual Analysis

Visualizations provide a deeper understanding of the dataset:

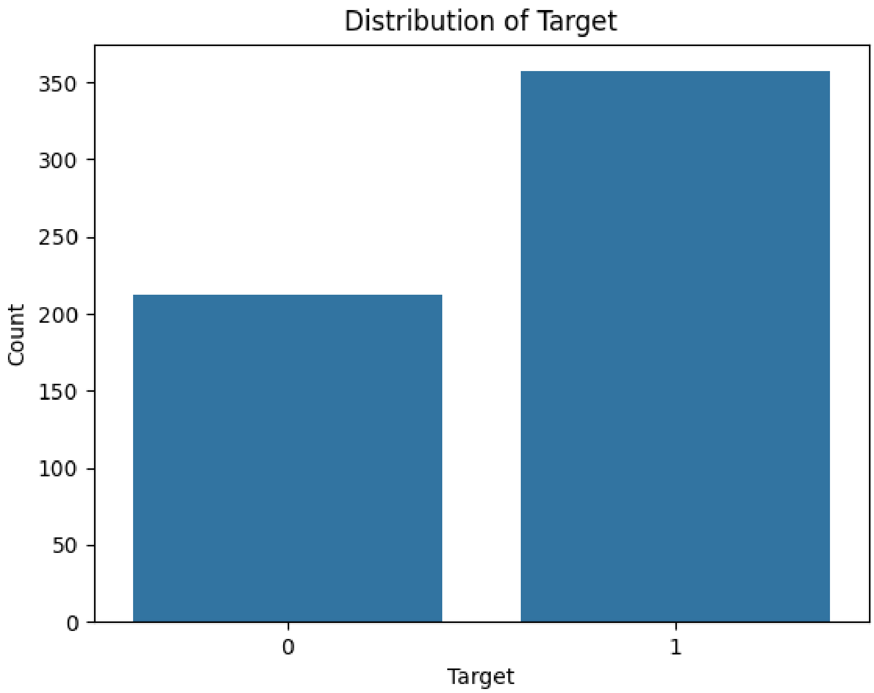
- **Pairplot:** The pairplot shows relationships between pairs of features. It indicates that certain features, such as mean radius and mean area, can help distinguish between benign and malignant masses.



- **Boxplot:** The boxplot illustrates the distribution of the mean radius for benign and malignant masses. It shows a clear difference in the mean radius between the two classes, indicating its potential as a predictive feature.



- **Bar Plot:** The bar plot visualizes the distribution of benign and malignant masses in the dataset. It shows that the dataset is slightly imbalanced, but not to a degree that would significantly affect the accuracy of the model.



Proposed Prediction Model

Based on the analysis, we propose building a prediction model using logistic regression. Logistic regression is suitable for binary classification tasks like this, where the goal is to predict whether a breast mass is benign or malignant. By leveraging the insights gained from the analysis, we can train a logistic regression model to make accurate predictions.

Model

- The neural network model consists of an input layer, a hidden layer with 20 neurons and ReLU activation, and an output layer with 2 neurons and sigmoid activation.
- The model was compiled using the Adam optimizer and sparse categorical crossentropy loss function.

Training

The model was trained on the training dataset with a validation split of 0.1 and 10 epochs.

Evaluation

The model achieved an accuracy of 94.73% on the test dataset.

Predictive System

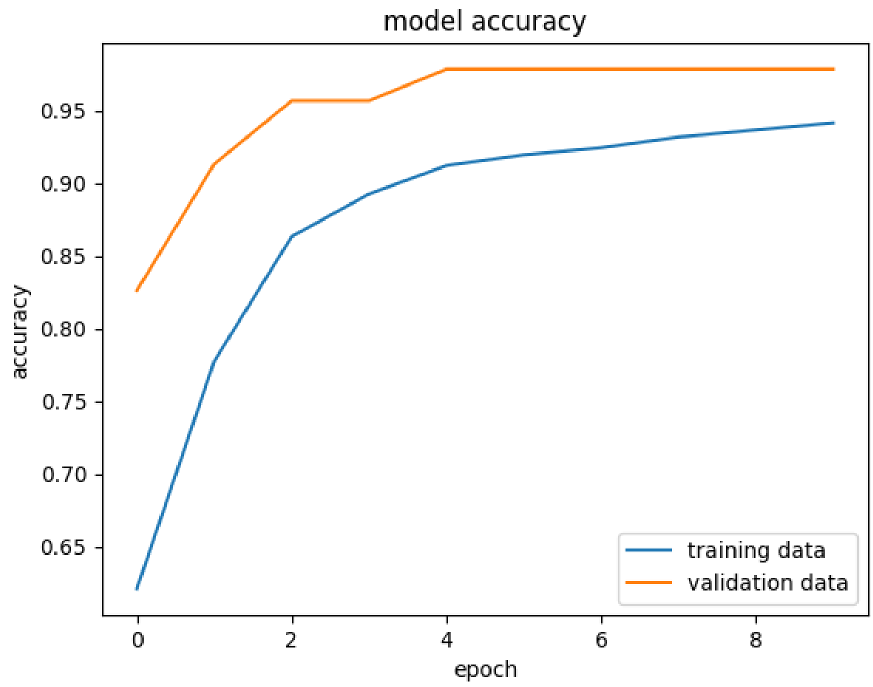
- A predictive system was developed to classify tumors based on user input.
- The system takes input data, standardizes it, and predicts the class label using the trained model.

Math and Logic Behind the Model

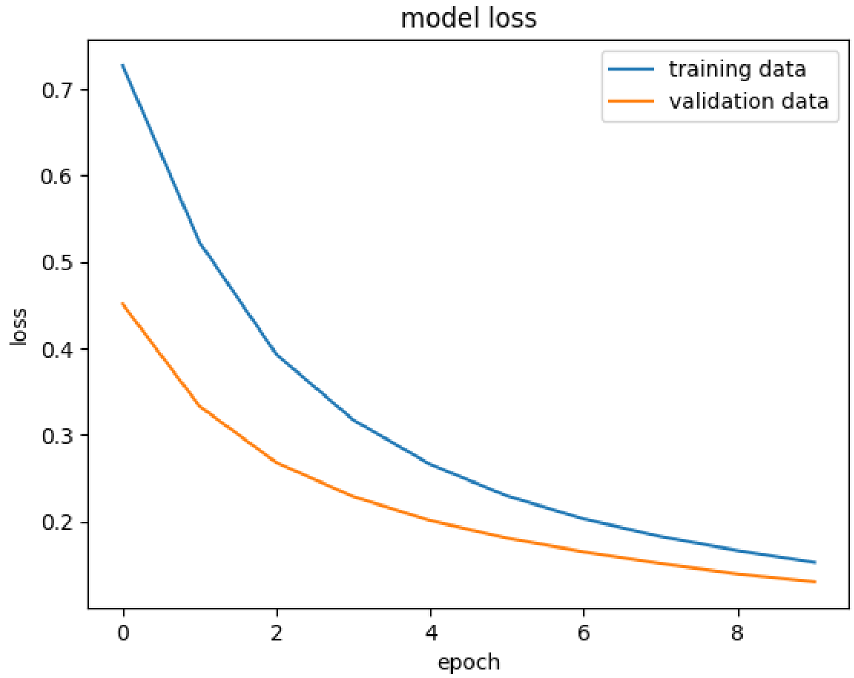
- **Logistic Regression:** While the model architecture is a neural network, the final layer with the sigmoid activation function effectively performs logistic regression. Logistic regression is suitable for this scenario because it outputs probabilities that a tumor is benign or malignant, which can be easily interpreted.
- **Decision Boundary:** The decision boundary separating the two classes is learned by the model during training. The weights and biases of the neural network are adjusted to minimize the loss function, effectively finding the optimal decision boundary in the feature space.

Visual Representation of Model Performance

Accuracy Plot:



Loss Plot:



Conclusion

In conclusion, this project has provided valuable insights into the process of developing a simple neural network model for breast cancer classification. While the model itself may not be directly applicable in a medical setting, the analysis process has been highly educational.

Through data analysis, we have gained a deeper understanding of classifiers and how mathematical techniques can be used to identify patterns within datasets. The use of summary statistics, correlation analysis, and visualizations has been particularly illuminating.

One of the most interesting findings was the clear separation of data points in the visual representations, indicating that certain features can effectively distinguish between benign and malignant tumors. This highlights the potential for using machine learning models to aid in medical diagnosis, where such visual patterns may be crucial for accurate classification.

Overall, this project has not only demonstrated the application of neural networks in classification tasks but also underscored the importance of thorough data analysis and feature selection in developing effective machine learning models.