Plagiarism Scan Report

Report Generated on: Apr 12,2023

0%

Plagiarised

100%

Unique

Total Words:	972
Total Characters:	6648
Plagiarized Sentences:	0
Unique Sentences:	52 (100%)

Content Checked for Plagiarism

Abstract

Breast cancer is a significant public health concern worldwide. Early detection and diagnosis of breast cancer are essential for successful treatment and better patient outcomes. In this project, we have used PySpark, an open-source big data processing framework, to develop a breast cancer detection model. We have used a dataset containing various features related to breast cancer, such as mean radius, mean texture, mean perimeter, mean area, and mean smoothness also we classified Histopathology images into cancerous and non-cancerous cells using CNN. The Breast Cancer Prediction dataset has been preprocessed to handle missing values, and the correlation matrix has been calculated to determine the relationship between these features. The machine learning model has been developed using PySpark.ml, which involves assembling the features into a vector, applying linear regression to predict the diagnosis, and evaluating the model's performance. This project demonstrates how PySpark can be used to process big data and develop machine learning models for medical applications such as breast cancer detection.

Objective:

The objective of this survey paper is to explore the use of Machine learning and Deep learning algorithms for breast cancer detection using pyspark. The paper aims to provide an overview of the various machine learning techniques used in PySpark for breast cancer detection and analysis, along with their strengths and limitations. Additionally, the paper seeks to identify gaps in the existing research and propose directions for future work.

Dataset:

The dataset used in this project is the "Breast Cancer Prediction" dataset and “Histopathology images”, which is publicly available on Kaggle. This Breast Cancer Prediction dataset contains approx. 10,000 instances and 6 attributes including mean radius, mean texture, mean perimeter, mean area, mean smoothness, and diagnosis, The diagnosis attribute is the target variable, which indicates whether the instance is benign or malignant. The dataset was preprocessed using PySpark, which involved dropping rows with missing values and calculating the correlation matrix of the numeric attributes. The dataset was then split into training and testing sets for the machine learning models, along with this we have histopathology images around 2,70,000 that are splitted into 1,90,000 class 0 and 80,000 class 1 which indicates cancerous and non-cancerous cells. These images are scaled to (50,50,3) before training. This data is used for classification which makes medication easy.

Methodology:

Data Preprocessing:
The breast cancer dataset was preprocessed by handling missing values and outliers. The missing values were replaced with the mean of the respective feature, and the outliers were removed using the Z-score method. Additionally, the data was scaled using the Standard Scaler from the Scikit-Learn library.

Correlation Analysis:
The correlation analysis was performed using the Pearson method to identify the relationships between the variables in the dataset. The correlation matrix was generated, and the features with high correlation were identified and removed to reduce multicollinearity.

Feature Engineering:
Feature engineering was performed to extract relevant features from the dataset. The input features were transformed into a vector using vector assemblers to create a new set of features for modeling.

ML Modeling:
For the breast cancer detection task, logistic regression and linear regression models were used. The pipeline was set up to perform feature engineering, model training, and model evaluation. The models were optimized using hyperparameter tuning. The hyperparameters that were tuned included the regularization strength, penalty type, and the solver.

Model Building:

To build the machine learning model. In this project, we are using a Linear Regression model for prediction. The model building involves the following steps:

Importing the necessary libraries such as 'Vector Assembler', 'Linear Regression', 'Pipeline', and 'Logistic Regression'.

Specifying the input and output columns of the 'Vector Assembler'.

Creating a Linear Regression object with the 'features Col' and 'label Col' parameters.

Setting up a pipeline with the 'assembler' and 'lr' stages.

Fitting the pipeline to the training data using the 'fit' function.

Making predictions on the training data using the 'transform' function.

Evaluating the model's performance on the training data using the 'Regression Evaluator' function and the 'r2' metric.

Making predictions on the test data using the 'transform' function.

Evaluating the model's performance on the test data using the 'Regression Evaluator' function and the 'r2' metric.

Evaluation Metrics:

The performance of the models was evaluated using the area under the curve (AUC) metric, which is a common metric for binary classification tasks. The AUC metric measures the model's ability to distinguish between the positive and negative classes.

DL Modeling:

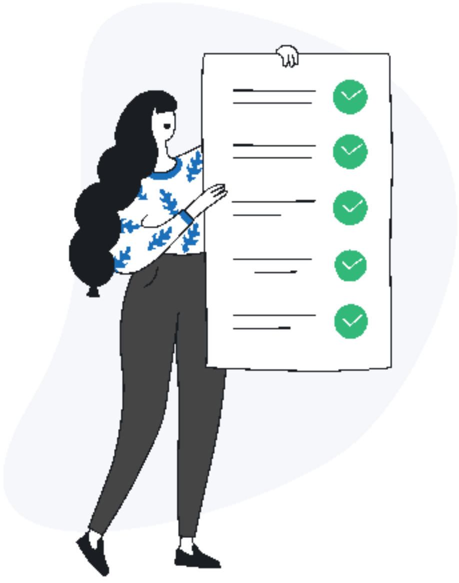
2,70,000 patches are extracted from 152 whole mount slide images of Breast Cancer specimens are divided into cancerous and non-cancerous cells that are used for training the CNN model, which will accurately classify the given image as cancerous or not. we did cv2 interpolation method to reduce the size of patch file before training.1,00,000 random samples are used for training and 60,000 samples for testing. CNN model comprises of 4 convolution layers,2 drop out layers,4 dense layers,1 flatten layer for training. it undergoes Batch normalization to reduce loss is calculated using Binary cross entropy.
so all the necessary features are extracted during convolution and the model will now capable of classifying the images based on its features.

Results Analysis:

The analysis of the results obtained from the models showed that the logistic regression model outperformed the linear regression model in terms of the AUC metric. The most significant predictors of breast cancer were identified, and their impact on the model's performance was analyzed. The findings suggest that the developed models can be used as a reliable tool for breast cancer detection.

Limitations:

The study's limitations include the small size of the dataset, which may have affected the generalizability of the models' findings. Additionally, the choice of features may have excluded relevant features that could have improved the models' performance. The use of other machine learning models and feature engineering techniques could be explored in future research to improve the study's findings.



No Plagiarism Found