Milestone 3: Preliminary Analysis

Aditya Sumbaraju

Bellevue University

DSC 630 Predictive Analytics

Professor Fadi Alsaleem

Oct 01, 2021

Abstract

A crucial stage in data analysis is exploratory data analysis (EDA), which enables researchers and data analysts to uncover patterns in large quantities of data. Using medical data analysis, doctors and healthcare researchers may better understand their patients via text and visual data analysis. In this project, I will be exploring the feature reduction properties of independent component analysis on the breast cancer prediction model. The actual data with 30 features and reduced one feature is used to evaluate the accuracy of the classifiers such as support vector machine (SVM), k-nearest neighbor (k-NN), and logistic regression. These classifiers are evaluated in terms of specificity, sensitivity, accuracy, F-score that helps to categorize tumors as benign and malignant. The result will be a comparison of the proposed classification using the initial feature set is also tested on different validation using 10-fold cross-validations and partitioning (20%–40%) methods.

Introduction

Treatment, rehabilitation, and medication for breast cancer depend on cutting-edge technology such as predictive analytics, data science, machine learning, and deep learning. These advances have enabled researchers to make more accurate and timely judgments. Breast cancer comes in a variety of forms and affects people worldwide. According to a 2018 World Health Organization study, 2.09 million new instances of breast cancer were anticipated, with 6.27,000 deaths. Around 70% of cancer fatalities take place in low- and middle-income countries. It is critical to discover prognostic variables and associated treatment predictive components for the disease since the clinical and molecular elements of the illness are inextricably linked (Chawla et al., 2021). Predictive indicators will assist in patient decision-making about breast cancer therapy, particularly for individuals with early-stage disease and tumors that have not metastasized, invaded, or multiplied in regional lymph nodes. Patients will live longer and have fewer diseases as a consequence of this therapy. Investigating the statistical and descriptive features via experimentation and data analysis is a technique for gaining new knowledge. The display enables more detailed and rapid data analysis. By analyzing prognostic factors from breast cancer data, such as menopausal status, age, lymph node status, and tumor size, this EDA dataset may be utilized to identify new features associated with patient survival and nonrecurrence of breast cancer. After the researcher has completed the machine learning steps, the data should look like this:

Figure 1.1 shows the general flow of EDA.

According to cancer.org, breast cancer (BC) is common cancer among women worldwide. 1 in 8 chance that a woman can develop breast cancer, representing most new cancer cases and cancer-related deaths, making it a significant public health problem in today's society.

The early diagnosis of BC can significantly improve the prognosis and chance of survival, promoting timely clinical treatment. The accurate classification of benign tumors can prevent patients from undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classifying patients into benign or malignant groups are subject to extensive research. Data mining and machine learning (ML) is widely recognized as the methodology of choice in breast cancer pattern classification and forecast modeling because of its unique advantages in critical features detection from complex BC datasets.

Data Mining and Classification methods are effective ways to classify data, especially in clinical analysis; these methods are widely used in diagnosis and research to make decisions.

My solution to this problem involves building a model that accurately classifies tumors as Benign or Malignant based on the tumor shape and geometry.

Method

I would be observing for categorical data and preprocess the features for deriving better accuracy of the Model.

In this particular use case, I would be deriving

- EDA on the identified features from the Dataset
- Feature decomposition using Principal Component Analysis(PCA)
- Build classifiers using Linear Regression, Random Forest, and Neural Networks
- Deriving the classification report for better understanding of accuracy and scores of Model. Perform cross-validation to measure each methods classification accuracy
- Check for roc_aur_acore and the F1 score of Model.
- Evaluate and finalize a better Model for the use case based on a 10-Fold cross-validation accuracy score.

Data collection

Dataset classifies the Benign and Malignant cells using the description of the cells in the form of columnar attributes. This data was donated by researchers of the University of Wisconsin and included the measurements from digitized images of fine-needle aspirate of a breast mass.

The data sets are provided with 569 examples of cancer biopsies, each with 32 features. One feature is an identification number, another is the cancer diagnosis, and 30 are numeric-valued laboratory measurements.

The diagnosis is coded as

- ✓ "M" to indicate malignant
- ✓ "B" to distinguish benign.

The other 30 numeric measurements comprise the mean, standard error, and worst (i.e., most significant) value for ten different characteristics of the digitized cell nuclei.

Data dictionary:

column_name	Description
id	ID number
diagnosis	The diagnosis of breast tissues (M = malignant, B = benign)
radius_mean	mean of distances from the center to points on the perimeter
texture_mean	standard deviation of gray-scale values
perimeter_mean	mean size of the core tumor
area_mean	area of the tumor
smoothness_mean	mean of local variation in radius lengths
compactness_mean	mean of perimeter^2 / area - 1.0
concavity_mean	mean of the severity of concave portions of the contour
concave_points_mean	mean for the number of concave portions of the contour
symmetry_mean	
fractal_dimension_mean	mean for "coastline approximation" - 1
radius_se	standard error for the mean of distances from the center to points on the perimeter
texture_se	standard error for the standard deviation of gray-scale values
perimeter_se	
area_se	
smoothness_se	standard error for local variation in radius lengths
compactness_se	standard error for perimeter^2 / area - 1.0
concavity_se	standard error for the severity of concave portions of the contour
concave_points_se	standard error for the number of concave portions of the contour

fractal_dimension_se	standard error for "coastline approximation" - 1
symmetry_se	
texture_worst	largest mean or "worst" value for the standard deviation of gray-scale values
radius_worst	largest mean or "worst" value for the mean of distances from the center to points on the perimeter
perimeter_worst	
area_worst	
smoothness_worst	largest mean or "worst" value for local variation in radius lengths
compactness_worst	largest mean or "worst" value for perimeter^2 / area - 1.0
concavity_worst	"worst" or largest mean value for the severity of concave portions of the contour
concave_points_worst	largest mean or "worst" value for the number of concave portions of the contour
symmetry_worst	
fractal_dimension_worst	"worst" or largest mean value for "coastline approximation" – 1

Missing attribute values: none

Exploratory data analysis and Results:

By using exploratory data analysis, physicians may more accurately predict whether or not a patient's cancer will return and how long they will survive (EDA). Python is used in EDA research to analyze and compute statistics and graphically display the results. Since the Dataset is derived from actual patient records of breast cancer patients, missing values cannot be imputed.

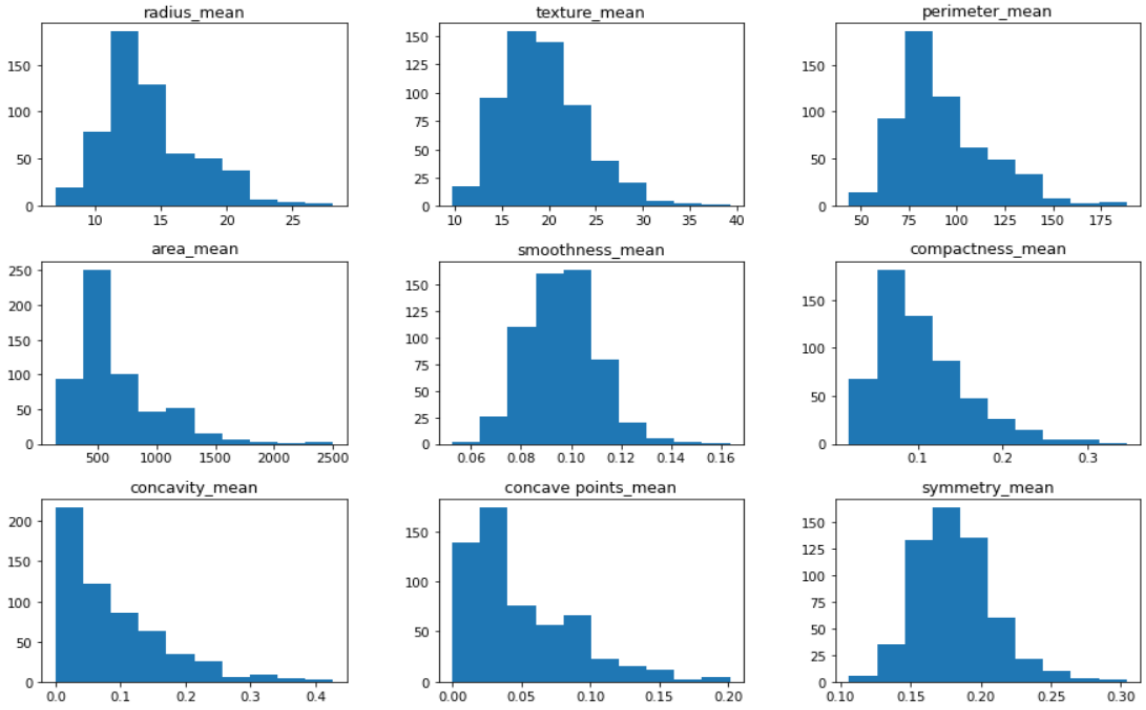
I have used descriptive statistics to derive mean, SD, and correlation. After grouping on diagnosis, it has been observed that 357 observations indicating the absence of cancer cells and 212 show absence of cancer cell

Out[28]:

diagnosis	Number of observations per diagnosis
B	357
M	212

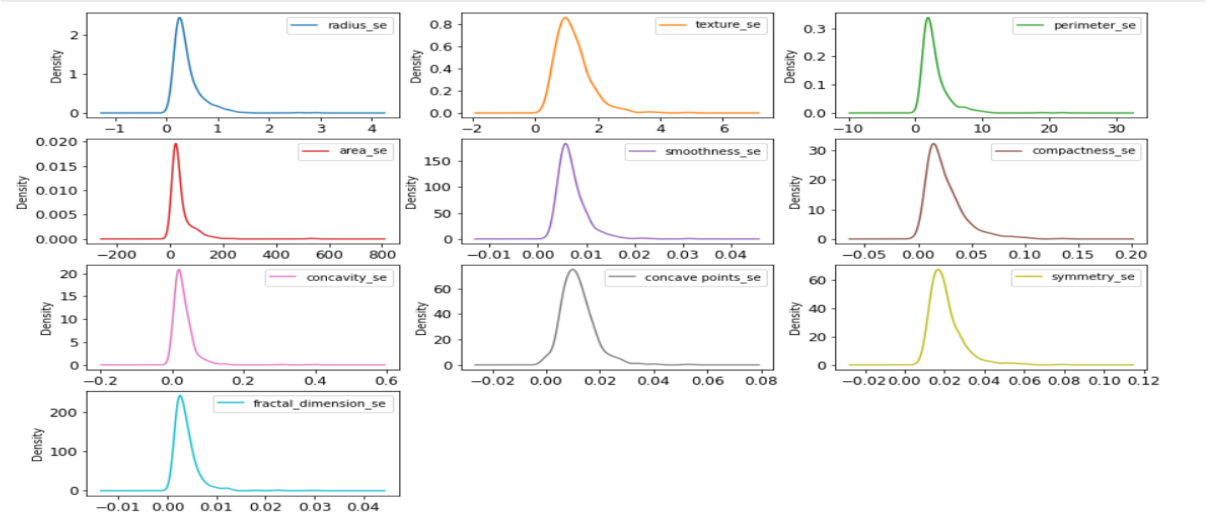
Histogram Plots on "_Mean" suffix variables:

Milestone 3: Preliminary Analysis



The above plot shows that the attributes concavity_mean and concave_point_mean may have an exponential distribution ( ). On the other hand, the texture_mean, smoothness_mean, and symmetry_mean features may have a Gaussian or nearly Gaussian distribution.

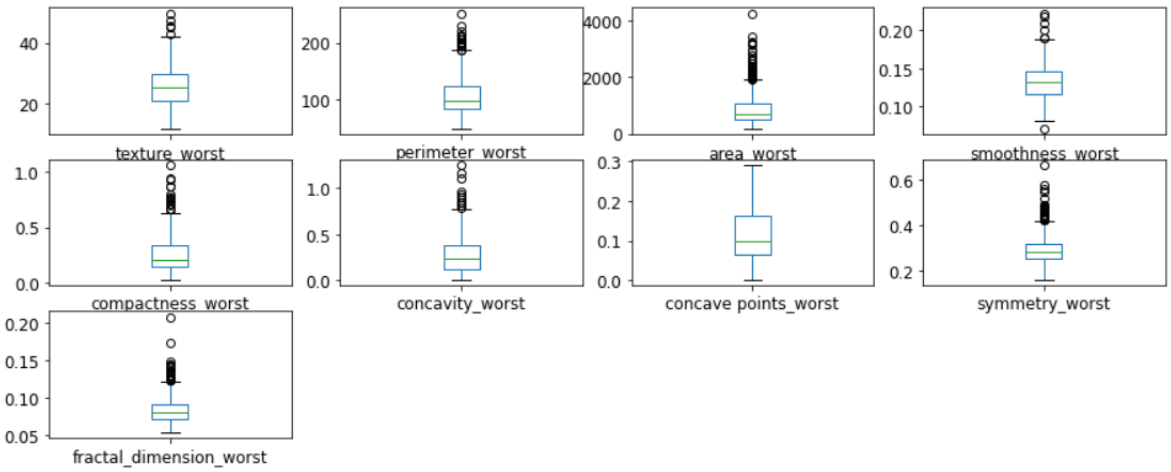
Density plots on "_se" suffix variables:



Milestone 3: Preliminary Analysis

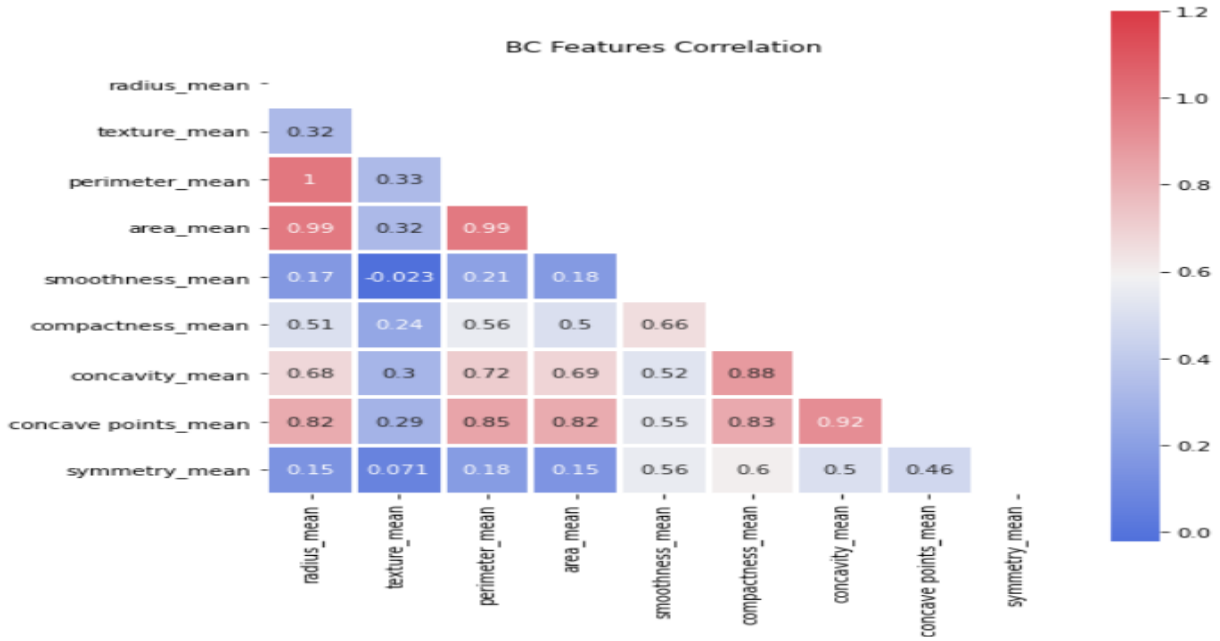
The attributes radius, perimeter, area, concavity may have an exponential distribution ( ). We can also see that the texture, smooth, and symmetry attributes may have a Gaussian or nearly Gaussian distribution.

BoxPlots on "_worst" suffix variables:



The attributes **concavity** and **concavity_point** may have an exponential distribution ( )

Correlation matrix



Conclusion

- In any of the histograms, there are no noticeable significant outliers that warrant further cleanup.
- The area_mean of the tissue nucleus has a strong positive correlation with mean values of radius and parameter.
- Likewise, we see some solid negative correlation between symmetry mean with radius, texture, parameter, and the area mean values.
- Mean values of cell radius, perimeter, area, compactness, concavity, and concave points can be used to classify cancer. Larger values of these parameters tend to show a correlation with malignant tumors.
- Mean values of texture, smoothness, symmetry or factual dimension do not show a preference for one diagnosis over the other.

Code:

Git Hub Link:

https://github.com/adityasumbaraju/DSC-630/blob/main/project/DSC630_ProjectMilestone3_DataPreperation_PreliminaryAnalysis.ipynb