Exploratory Data Analysis

Ali Torabi

1. Introduction

Machine learning is changing our lives swiftly. While a simple task like image recognition seems so trivial for humans, there is a lot of work to do in the machine learning pipeline, such as data cleaning, feature engineering, finding which models best fit for the specific problem, and hyperparameters tuning, among many other tasks. A lot of these tasks are still not automated and need an expert to do some of these trials and errors. In this project, I will be using a sample dataset named the Breast Cancer dataset to do some engineering tasks before diving into creating and training the model. It compromise of sets of steps to take in order to make ourself familiarize ourselves with data and finding what problems it has and which preprocessing tasks are suitable to do.

2. Dataset Description

The dataset chosen for this experiment is the Breast Cancer dataset [1]. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. [1]. This data set includes 569 instances and 30 features. The target label (Diagnosis) is binary which it takes 0 (No cancer) and 1 (cancer) in which it is suitable for classification. Each of these algorithms would be to detect the quality of wine ranges from poor to excellent. As mentioned in the description of the dataset itself, it has no missing value. All other features are continuous values and entirely related to the field of medical science.

3. Experimental Setup

Analyzing raw data before diving into machine learning is a critical step. There are some steps before going to the modeling:

1 – Understanding the Data Structure: Determine the size of the data, what and how many features it has. The type of features and label data makes it really simple to find what machine learning algorithm to choose from.

2 – Initial Data Exploration: Using basic statistics like mean, median, mode, and standard deviation to understand the distribution of data.

3 – Data Quality Assessment: Finding missing values and outliers and selecting an appropriate method to handling those anomalies in the data. Even, It worth to check for any inconsistency or inaccuracies in the data like negative values where only positive values are allowed. We can then apply data cleaning and preprocessing, accordingly. For example, using some sort of imputation to fill missing values with mean or median. Apply Normalization to scaling every feature to a similar scale.

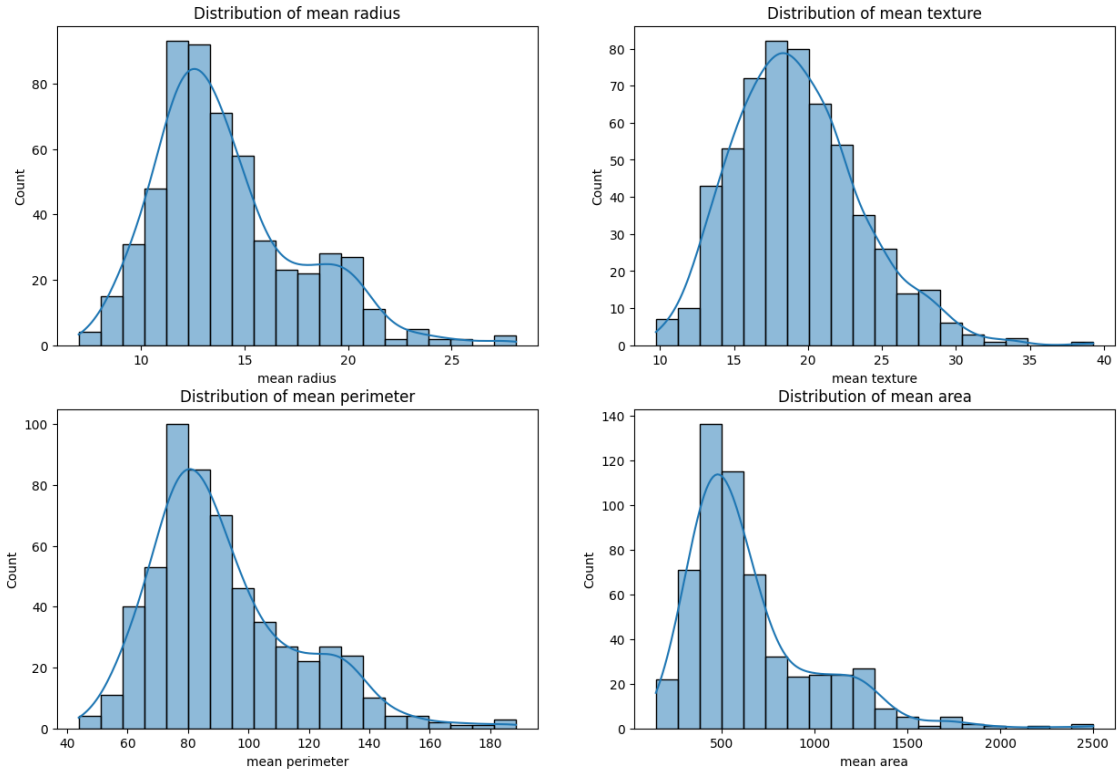
4 – Data Visualization: Use some plots to visualize distributions of numerical data. Also, we can provide correlation analysis to create scatter plots to understand the relationships between variables.

5 – Feature Engineering: You can create new features based on understanding of the data. Also, it is encouraging to reduce features if there are too many features using some techniques like PCA.

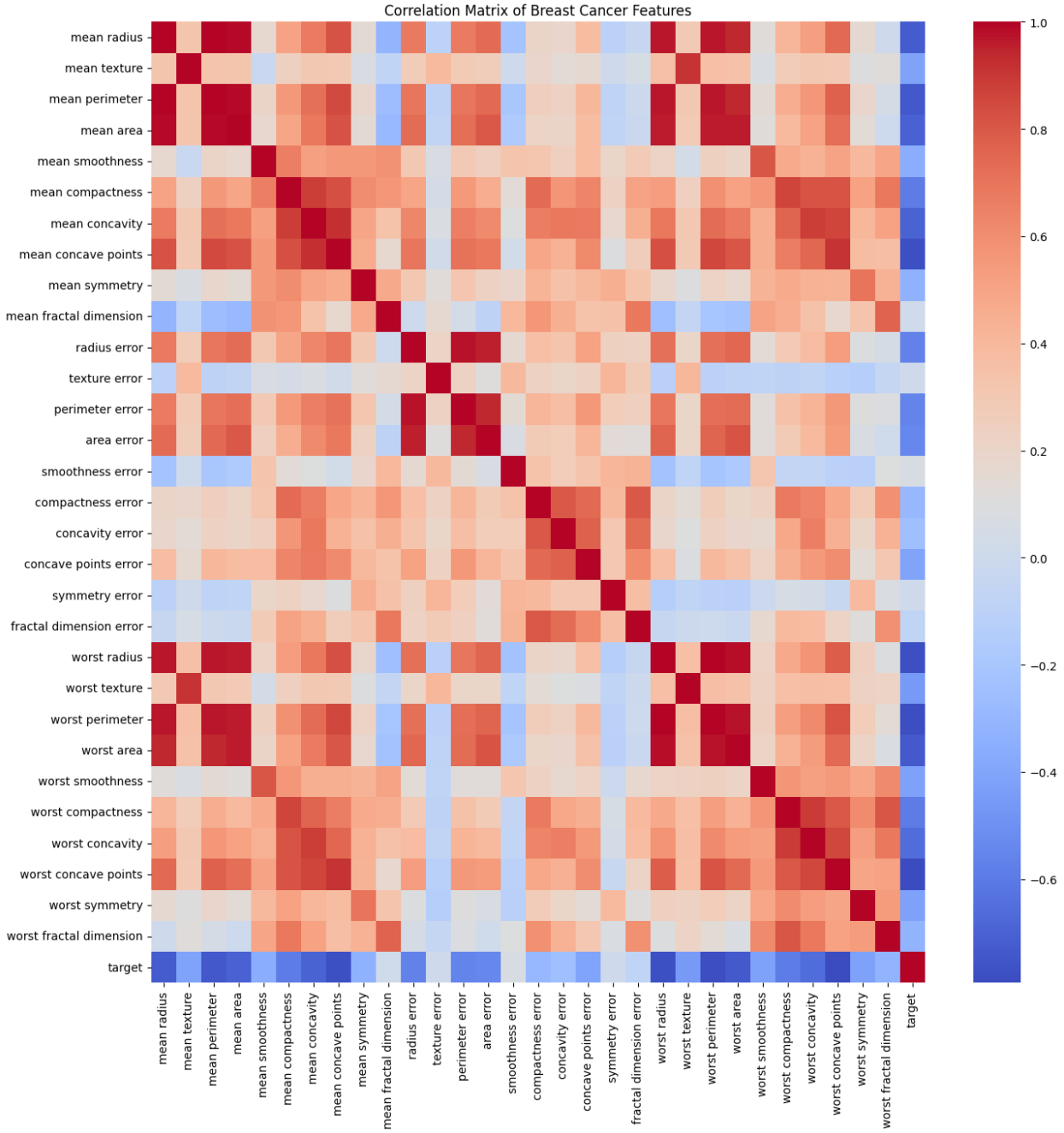
At first, for some of the features we can plot histograms to see their distributions. These subsets of features are mean radius, mean texture, mean perimeter and mean area. The interpretation could be as follows:

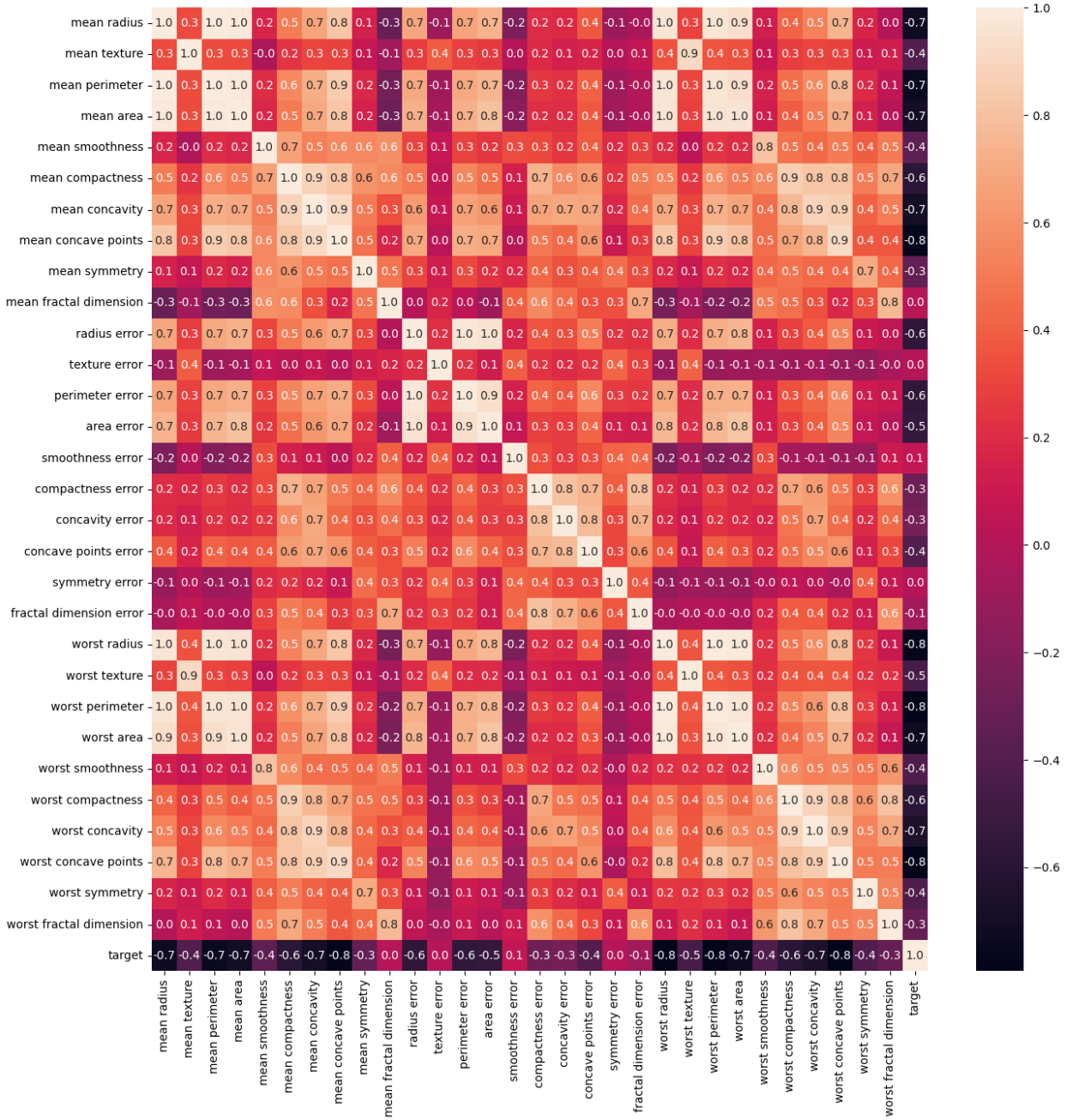
The height of each bar shows the frequency (count) of values within that bin of the histogram.
In terms of shape, if the histogram is approximately mirror-imaged across the center (symmetric), it means it is Normal or Gaussian distribution just like the distribution of mean texture. For some distributions like mean texture, we have skewed distribution of data, which means Tail is longer on one end. For this particular example, the tail is longer on right and he mean is typically greater than the median.

In terms of central tendency, the position of the peak indicates information about the mode, which is the count of most frequent value. It also provides some insight on mean and median, too.
We can investigate the range of data by looking at the width of the data in histogram that shows the spread of data. It also shows that the variance or standard deviation in such a way that wider distributions indicating higher variance. Isolated bars on the far ends of the histogram might indicate outliers. Gaps in the histogram may suggest that certain values are not present in the dataset. For example, in mean area, it seems like for values between 35 to 40, there is no data. We can also pinpoint outliers by using Z-Score function which we should provide what value counts as outliers. For example, for mean radius, the outliers can be detected by Z-Score where values are more than 3, which is the data values reside in indices: [82, 180, 212, 352, 461].



Another work that could be done for understanding data and their relationships, is by using the correlation analysis. This is part of Feature engineering to find which features have more impact on the result. A heatmap will be used to visualize the correlation between different features. This helps in identifying highly correlated features which might impact the model performance. Strongly correlated features appear in warmer colors (red), and weakly correlated features in cooler colors (blue). Even we can use another plot to provide correlation by showing correlation values (1 most correlated to -1, less correlated).


Correlation Matrix of Breast Cancer Features
mean radius
mean texture
mean perimeter
mean area
mean smoothness
mean compactness
mean concavity
mean concave points
mean symmetry
mean fractal dimension
radius error
texture error
perimeter error
area error
smoothness error
compactness error
concavity error
concave points error
symmetry error
fractal dimension error
worst radius
worst texture
worst perimeter
worst area
worst smoothness
worst compactness
worst concavity
worst concave points
worst symmetry
worst fractal dimension
target
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6



As part of the data cleaning and preprocessing, we can use a method like imputation for handling missing values, but because we don't have missing values in this particular dataset, there's no need to do that. Depending on the chosen machine learning model, normalization or standardization might be necessary to ensure that all features contribute equally to the prediction. All of these can be pack into pipeline preprocessing in machine learning. In the end, by exploring dataset and each

features, we can come up with the idea that how to work with those data, which machine learning algorithm is more fitted for this particular data, among many other insights to use.

References:

[1] https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

[2] https://medium.com/@ndleah/eda-data-preprocessing-feature-engineering-we-are-different-d2a5fa09f527