Worksheet 3 – Data Exploration and Preparation

Theory

Review the videos from week 3. Answer the following questions based on those lectures

1. What is meant by data exploration and why is it important?

Answer: Data exploration is concerned with understanding our dataset, to gain deep insights into its content and suitability for using in a machine learning application.

Recall that a dataset can come from a variety of sources. With such variety comes the risk that some instances may suffer from data quality issues such as missing values or invalid feature values through. When quality issues affect a significant proportion of the dataset, then this will inevitably impact the quality or even the viability of any machine learning model we train with this dataset.

1. What is the process of data cleaning and preparation? When and why do we carry out this activity?

Answer: The process of data cleaning and preparation is to get our data ready for machine learning training and testing.

We’ll do this before attempting to train the models. And the reason why we do this is that the quality and viability of the model is directly affected by the quality of the dataset.

1. What does covariance measure and why do we use it? What is an alternative measure and why might we use that instead?

Answer: Covariance is a function of the sample means of the variables in question and is defined as the weighted sum of the differences between the observed values for the feature pair and their respective population mean. It measures the degree of correlation between different variables.

A better approach would be to use a normalized measure such as correlation. Because the covariance measure only makes sense if the features being compared are measured in the same units.

1. What is a scatter plot matrix and what does it tell you about the features in a dataset?

Answer: A scatter plot matrix is the total set of pairwise plots of our dataset presented as one graph. It’s easy to see when two features are positively or negatively correlated though apparent linear relationships in the graphs. If we see any such suspected correlations, this would warrant further investigate and analysis of the impacted features such as calculating formal measures of this relationship.

1. What is the problem with including correlated features in a dataset when training a model?

Answer: Strongly correlated features can impact on many algorithms’ performance. Having two features contributing to the same information and learning process is both redundant and potentially error prone, especially if they are negatively correlated.

1. What is meant by dimensionality reduction?

Answer: Features which do not contribute any useful information or would actually impact on the performance of our training or resultant model should be removed altogether.

1. Why might it be recommended to normalize a continuous feature?

Answer: Because having continuous features which are in different ranges and different measurement units can adversely impact the ability of many learning algorithms to infer useful information and relationships across these variables.

1. What is the purpose of binning? Explain how it works with an example.

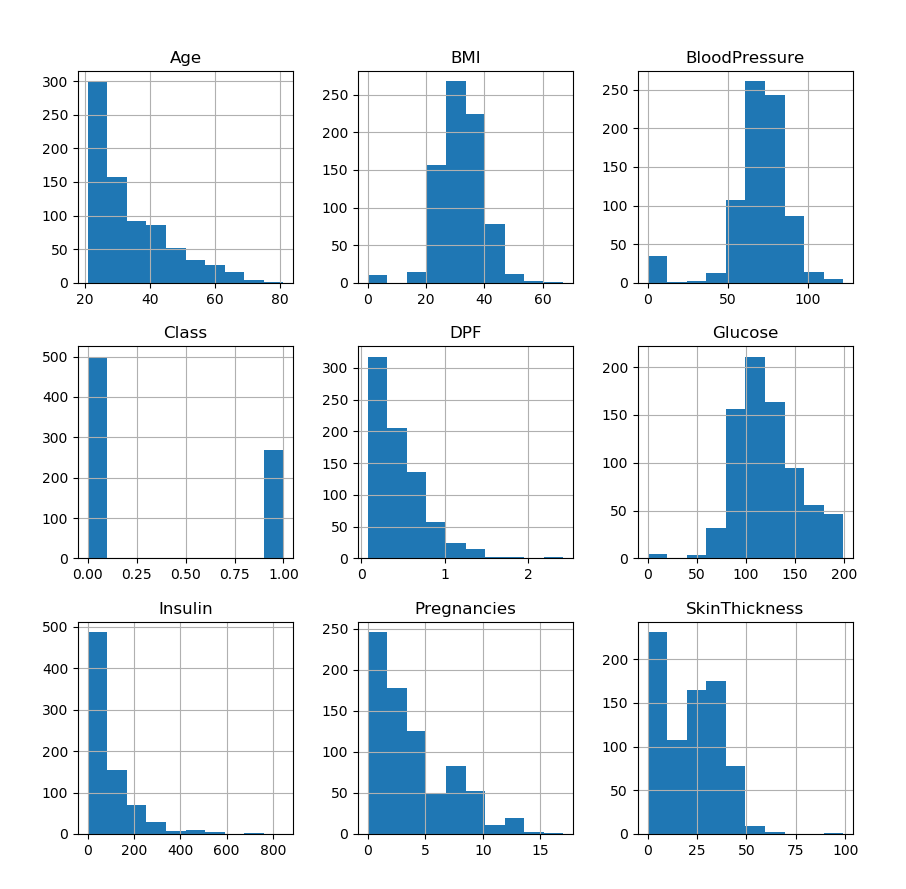
Answer: A large cardinality for a feature may not be useful for some learning approaches, for example decision trees. One way to address this is to use a technique called binning. There are two kinds of binning commonly used. These are equal-width binning and equal-frequency binning.

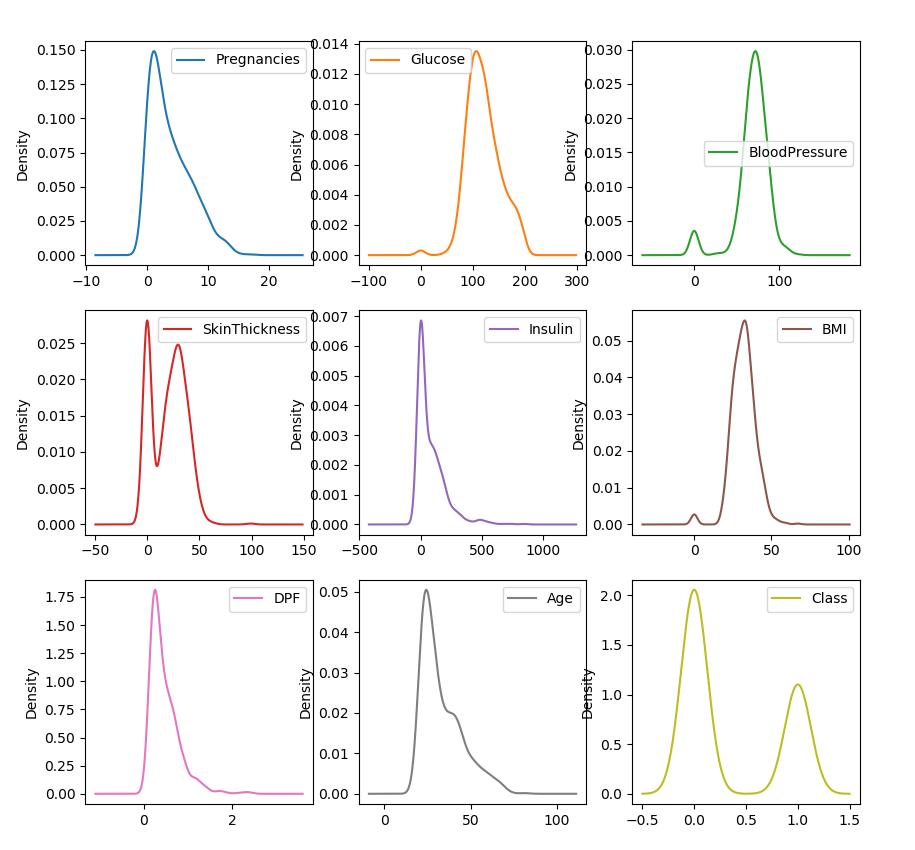
Practice

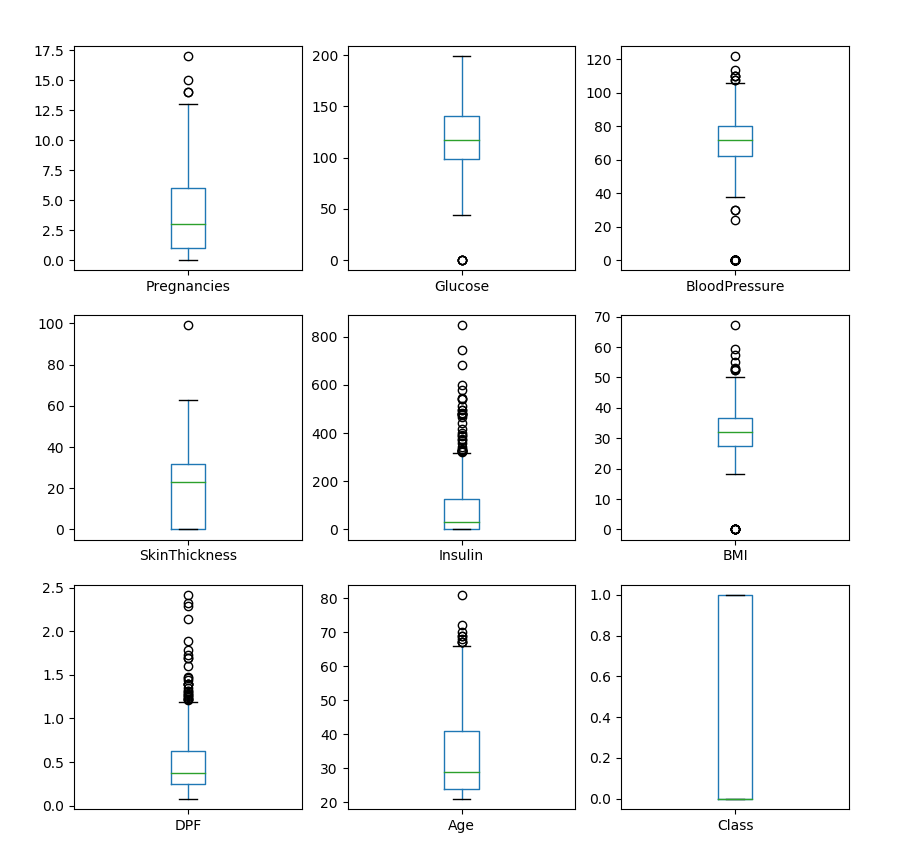
Follow the tutorial videos from week 2 and carry out the following steps

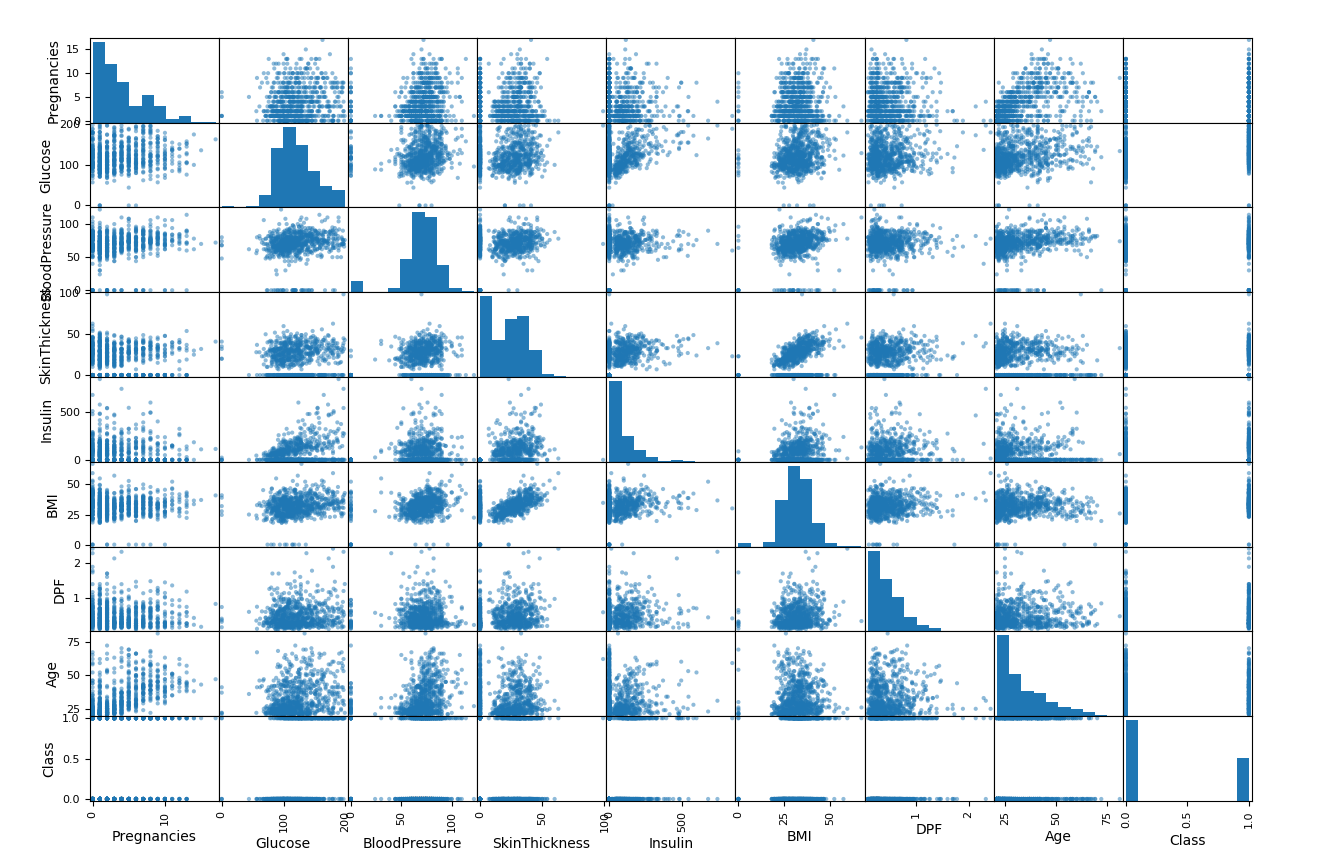
1. Download the code archive and extract the file from the week 3 learning materials. Make sure that you can run the examples code as provided.
2. Write Python code to create summary report of each of the features in the **pima-indians-diabetes.data.csv** file, included in the code archive for this week.

Answer: In my code, we plotted the histograms, density curves, bar plots and scatter matrix. As is shown in the picture.

Histograms

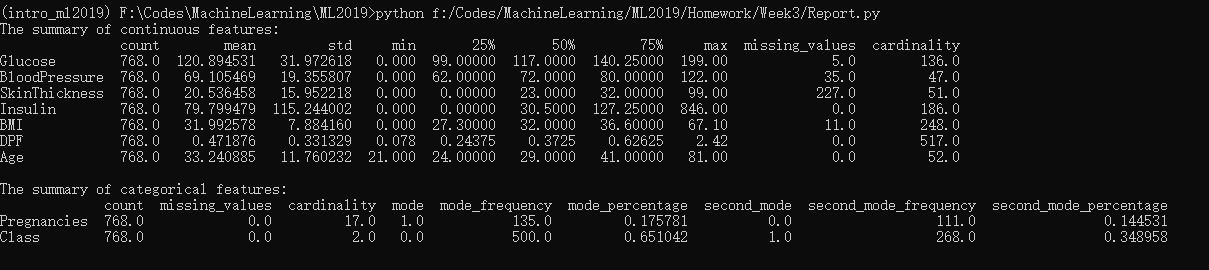
Density curves

Bar Plots



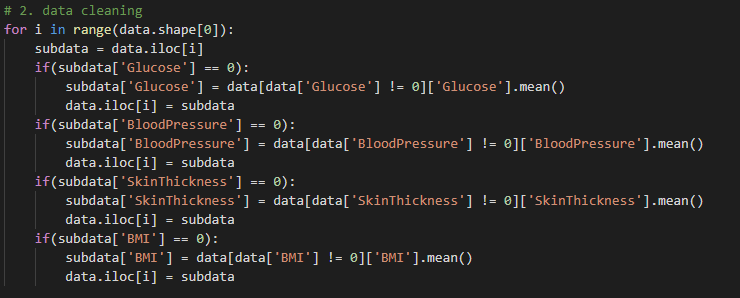
Scatter Matrix

From the density curves shown previously, we can easily see that there are some missing values for ‘Glucose’, ‘BMI’, ‘Skin Thickness’, ‘Blood Pressure’ and ‘Insulin’, because the curve looks like the gaussian distribution but reached the top at the value of 0, which means the missing values are noted as 0. For each of the features, it is needed to get some basic information about them. Two tables about their information are shown for information of categorical features and continuous features.

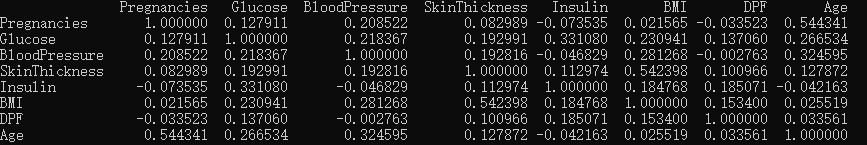


1. Examine any pairwise correlations in the scatter matrix and compute covariance/correlation statistics for any pairs you suspect might be correlated

Answer: First of all data cleaning steps, we need to clean the missing values. My method is to replace missing values with the average number for the same feature with other values. As is shown in the screenshot below.



After cleaning missing values, the second step is to drop correlated features. From the scatter matrix, we can see that the correlation of pregnancies and age, as well as the correlation of skin thickness and BMI is relatively high. So in our code, we printed the correlation matrix the pd.DataFrame.corr() method, as is shown in the picture below.

Correlation Matrix

From the picture we can see that the correlation of age and Pregnancies is 0.54431, and that the correlation of skin thickness and BMI is 0.542393, both of which have great impact on our model. We have to drop several features. Because BMI has better scientific supports, we dropped the skin thickness feature. And age is usually a more common feature in statistics, so we dropped the feature pregnancies. What’s more, for the ‘Insulin’ features, there are too many missing\_values, so we decided to drop the feature ‘Insulin’ in our dataset.

1. Prepare the final dataset for machine learning by removing any redundant columns.

Answer: We have our data cleaned in the third question, and import the decision tree and random forest model we have built last week, and got the accuracy score.

1. Build a classifier for the target variable using the decision tree or random forest model you build in week 2 and train and test this on the cleaned, split dataset.

Answer: We used both the decision tree and random forest classifier, here is the accuracy score.

