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?
2. What is the process of data cleaning and preparation? When and why do we carry out this activity?
3. What does covariance measure and why do we use it? What is an alternative measure and why might we use that instead?
4. What is a scatter plot matrix and what does it tell you about the features in a dataset?
5. What is the problem with including correlated features in a dataset when training a model?
6. What is meant by dimensionality reduction?
7. Why might it be recommended to normalize a continuous feature?
8. What is the purpose of binning? Explain how it works with an example

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.
3. Examine any pairwise correlations in the scatter matrix and compute covariance/correlation statistics for any pairs you suspect might be correlated
4. Prepare the final dataset for machine learning by removing any redundant columns
5. 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.