COSC 4010 Practical Machine Learning Fall 2023
Exploratory Data Analysis Report

Derek Walton

November 2023

1 Introduction

This exercise covers an important piece of any machine learning process: exploratory data analysis. As often stated in class and by industry speakers, the most challenging part of machine learning is data — i.e., data collection and preprocessing. Exploratory data analysis allows relationships between data that may not be known to be found, identifies missing values, outliers, and other issues that can affect a machine learning pipeline's ability to accurately make predictions. The dataset used for these exercises is the Pima Indian Diabetes Database from the UCI Machine Learning Repository.

2 Data Description

The Pima diabetes dataset comprises tabular data originally sourced from the National Institute of Diabetes and Digestive and Kidney Diseases [UCI Machine Learning, 2016]. It consists of 8 features, all representing continuous values based on specific diabetes diagnostic criteria. The objective is to predict whether a patient has diabetes or not. Of note all individuals in the dataset are female and are over 21 years old, belonging to the Pima Indian heritage.

2.1 Further Details

In Figure 1, a heatmap was generated to check for missing values, with missing values indicated by yellow areas and non-missing values represented by purple. Based on this plot, it's evident that the dataset doesn't contain any missing values.

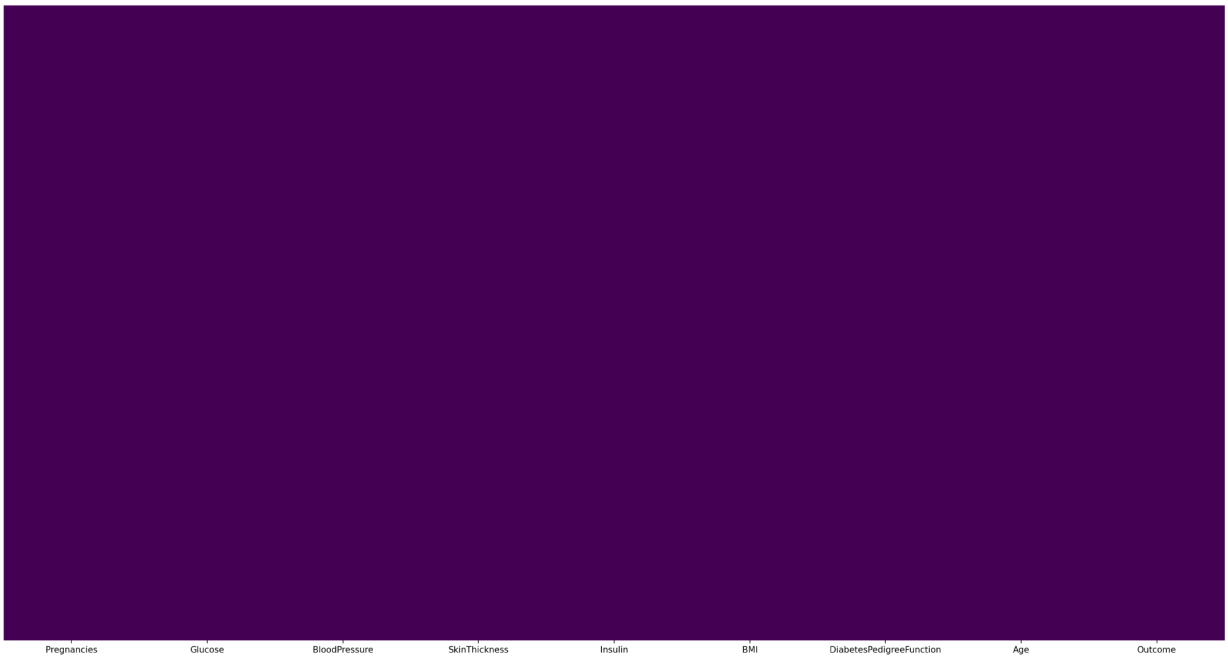

Figure 1

In this dataset, several features, such as insulin (mean ~ 79, max ~ 846), skin thickness (mean ~ 21, max ~ 99), and BMI (mean ~ 32, max ~ 67), may contain outliers. Figure 2, a box plot, illustrates the extent to which these suspected outliers deviate from the mean. The visualization reveals that while the mean for insulin is approximately 10 times lower than the maximum value, a considerable number of observations deviate significantly from the mean. This pattern also holds true for BMI, although its maximum value is not as extreme as that of insulin. Although these values outside the mean impact the dataset, it appears that the mean value for these attributes is not notably elevated when compared to standard diagnostic indicators [Collier, 1989]. Thus, these outliers do not seem to pose a significant issue within the dataset.

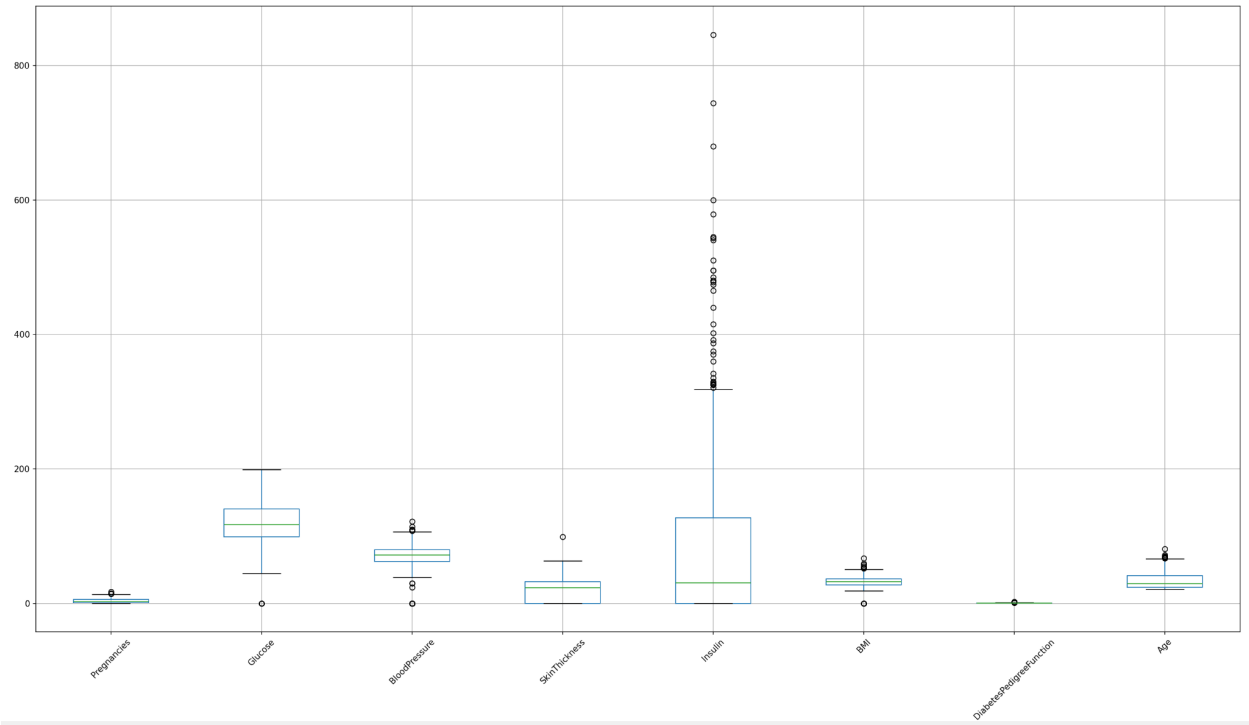

Figure 2

Figure 3 depicts a heatmap that visualizes the data correlation within the dataset. Highly correlated variables possess a linear dependence and may exhibit nearly equal predictive ability for the outcome value of an observation [Vishal, 2018]. Therefore, eliminating one of the correlated features before training can enhance the learning process [Lannge, 2021]. In this heatmap, a value of 1 indicates a high correlation. Upon examination, it appears that none of the features in the dataset exhibit a high correlation. Consequently, all features should be retained for model training.

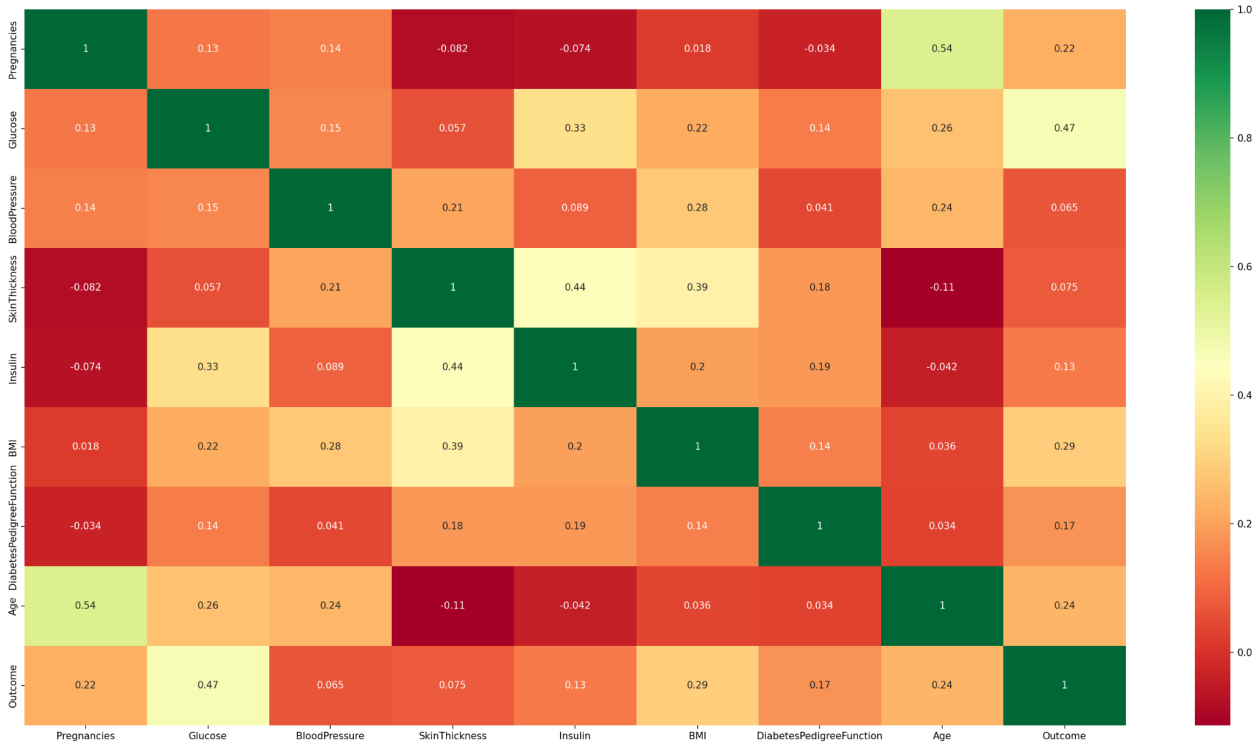

Figure 3

In Figure 4, a set of histograms is employed to assess the variation in feature data distribution compared to the target data. This assessment is crucial as linear models operate under the assumption that the distribution of the independent variable aligns with that of the target variable [Sharma, 2023]. In this scenario, although glucose, blood pressure, and BMI seem to follow a normal distribution, they exhibit a noticeable deviation from the distribution of the target data. This discrepancy should be considered when training a machine learning pipeline, as it might lead to diminished prediction accuracy.

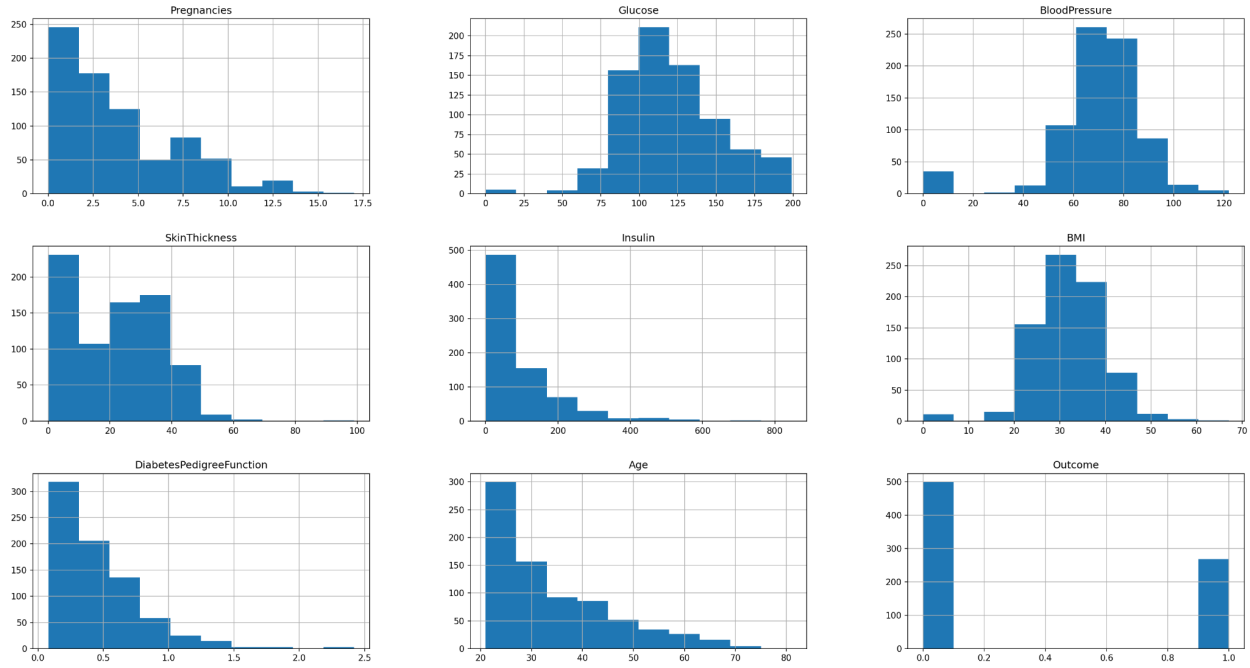

Figure 4

3.1 Experiment Setup

This exercise was conducted using the Python programming language, using the packages pandas, numpy, matplotlib, and seaborn. Selections regarding methods and visualizations aimed at understanding the dataset were guided by an article authored by P. Patil in 2022. In this article, Patil utilizes the white wine quality dataset to showcase common techniques used in the industry for data cleaning, providing explanations of the reasoning behind these practices.

References

UCI Machine Learning. (2016, October 6). Pima Indians Diabetes Database. Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Collier A, Patrick AW, Bell D, Matthews DM, MacIntyre CC, Ewing DJ, Clarke BF. Relationship of skin thickness to duration of diabetes, glycemic control, and diabetic complications in male IDDM patients. Diabetes Care. 1989 May;12(5):309-12. doi: 10.2337/diacare.12.5.309. PMID: 2721339.

Vishal, R., (2018). Feature selection — Correlation and P-value. towardsdatascience.com, September 11. Available: https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf [Accessed: 2021-12-27]

Lannge, E. J. (2021). *Does removal of correlated variables affect the classification accuracy of machine learning algorithms?*. diva-portal. https://www.diva-portal.org/smash/get/diva2:1632660/FULLTEXT01.pdf

Sharma, A. (2023, September 13). *Understanding skewness in data and its impact on Data Analysis (updated 2023)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/07/what-is-skewness-statistics/#:~:text=First%2C%20linear%20models%20work%20on,us%20create%20better%20linear%20models.

Patil, P. (2022, May 30). *What is exploratory data analysis?*. Medium. https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15