Exploratory Data Analysis

Mohammad Irfan Uddin

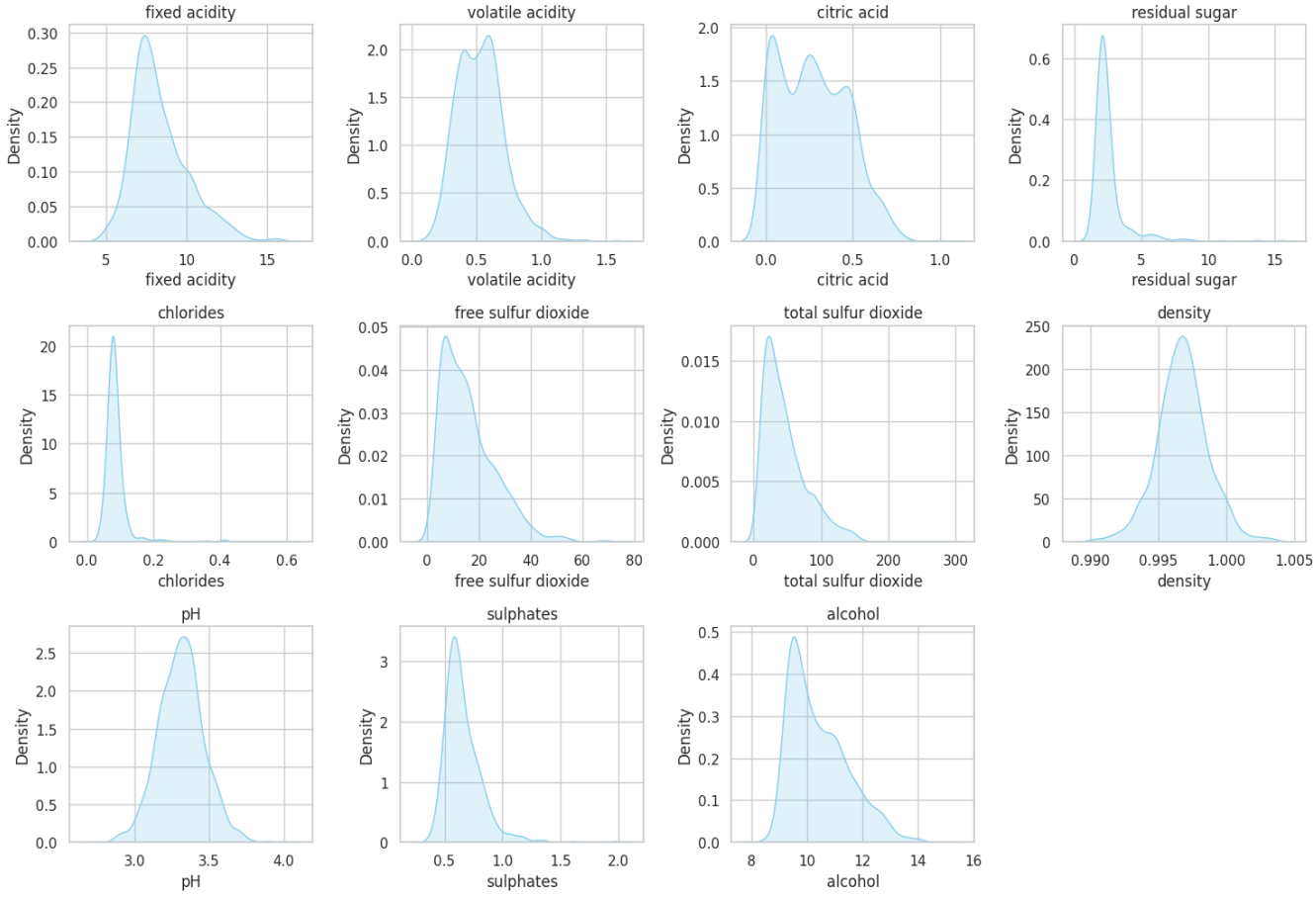
Introduction:

In this exercise, we aim to apply machine learning techniques to a dataset extracted from a wine quality database. The primary objective is to identify potential issues or obstacles that may hinder the successful application of machine learning models. To address these challenges, we plan to conduct an in-depth analysis of the raw data, including an examination of feature distributions, identification of missing values, and the implementation of preprocessing methods to enhance the dataset's suitability for machine learning.

Results of the Analysis:

The dataset under consideration, the Wine Quality Dataset, is composed of 11 features and contains 1599 samples. The target variable is wine quality, which is a categorical attribute. One notable aspect of this dataset is the absence of missing values, simplifying the preprocessing phase. However, to check our code is properly working, I have removed some data from the dataset and our model detected these as missing values.

To gain insights into the distribution of feature values, Kernel Density Estimation (KDE) plots were employed. These plots showcase the probability density function of each feature, providing a visual representation of their distributions. The figures below illustrate the KDE plots for each feature before preprocessing.



Normalization is a crucial preprocessing step, ensuring that all features contribute equally to the machine learning model by bringing them to a common scale. In our exercise, we utilized the Min-Max Scaler from the scikit-learn library to normalize the feature values.