Documentation on Drug Prediction using Naïve bayes

**Dataset** : https://www.kaggle.com/jeevanrh/drug200csv/kernels

**Overview :** The dataset contains information about various attributes of a set of drugs. The dataset includes details such as Age, Cholesterol , Sex , BP , Na\_to\_K.

**Problem Statement :** Fit a model to predict the quality of the fruit using Naïve Bayes .

1. **Import Libraries**

*# To deal with the dataset*

import pandas as pd

import numpy as np

# For data visualization

import matplotlib.pyplot as plt

import seaborn as sns

# To fit and predict the quality of fruit

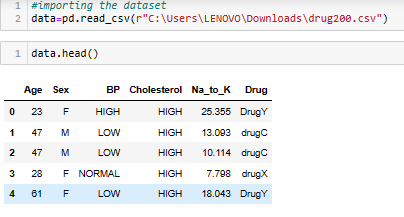
from sklearn.model\_selection import train\_test\_split

from mixed\_naive\_bayes import MixedNB

from sklearn.metrics import accuracy\_score

**Data Collection**

1. Importing the dataset and preview of the data :



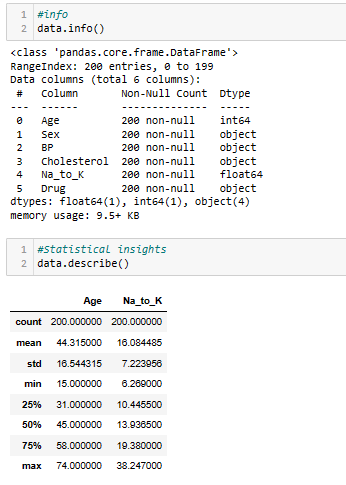
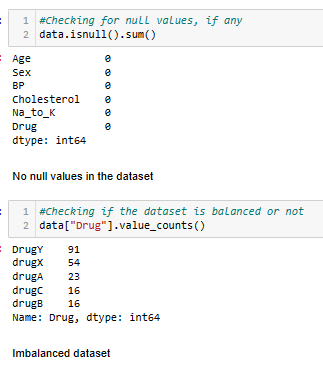
**Data Cleaning / Data manipulation**

1. Data Cleaning :

* Found No null values

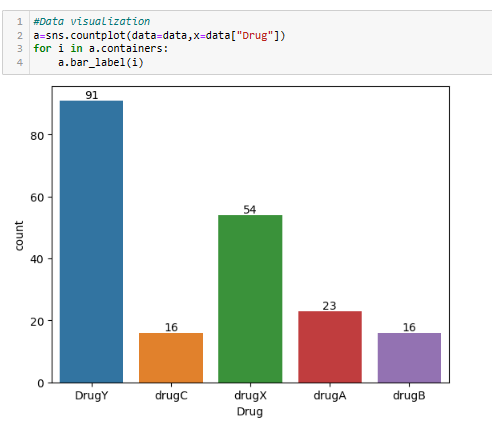
**Exploratory Data Analysis**

1. Statistical sights :

Dataset consists of 2 numerical and 3 categorical columns

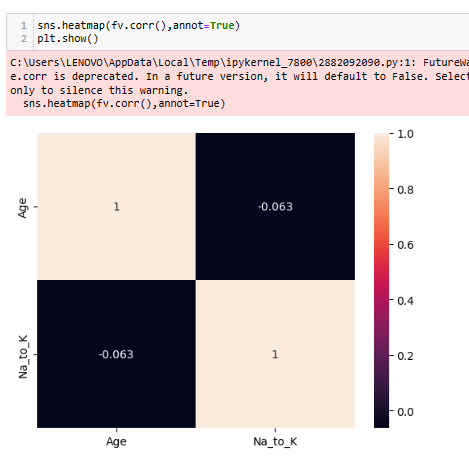
1. Exploratory Data Analysis :



Based on the count plot of class variable, the dataset is imbalanced.

We have 5 different drugs in Drugs column.

Correlation Between feature variable :



There exits not much correlation between feature variables

**Feature Engineering**

**Training the model**

1. Dividing the dataset to train and test :

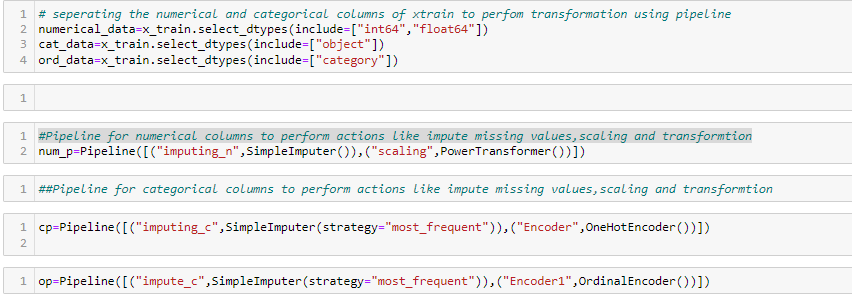


1. Feature engineering :

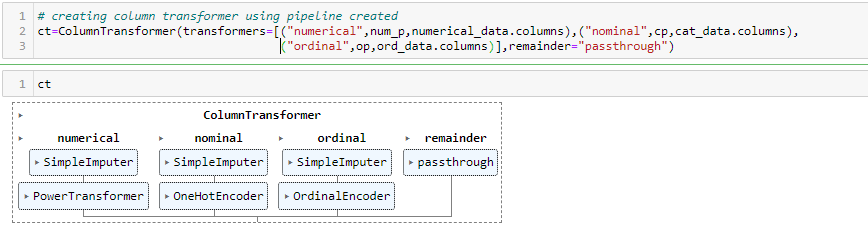
* Extracted required feature variable and class variable from the dataset
* Since the dataset contains categorical and numerical columns, Mixed naïve bayes has been used.

1. Creating Pipelines :

Created different pipelines for different datatypes :

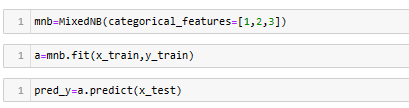


**Column Transformation using pipelines :**

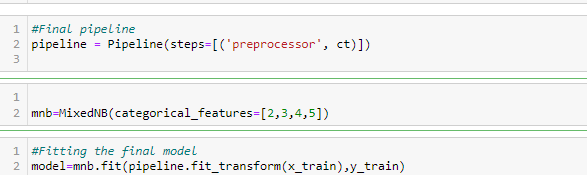
****

**Testing the model**

1. Predicting the Class variable for test data using the model

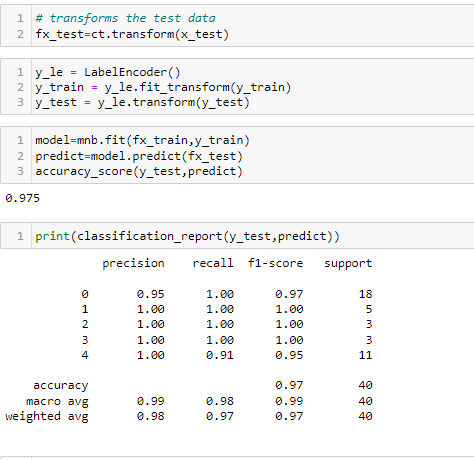


1. Fitting the final model



**Performance of the model**

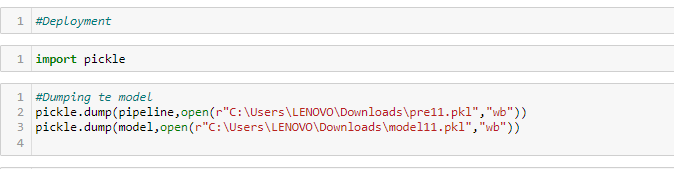
1. Performance of the model :



The accuracy\_score of the model is 97.5% which concludes that the performance is good.

1. **Deployment :**

Dumped the model and loaded in VSCode :



1. Streamlit :

