**Performing EDA and calculating the accuracy of the diabetes dataset**.

Prepared in the partial fulfillment of the Summer Internship Program on Data analysis using python



Under the guidance of

Mrs. K. Meenakshi, APSSDC

Mrs. K. Narmada Mani, APSSDC

Submitted by

Lokeswari Yarlagadda, AP21110011547

SRM University AP

**ACKNOWLEDGEMENT**

I would like to express my heartfelt gratitude to all those who have contributed to the successful completion of my summer internship project at Andhra Pradesh Skill Development Corporation (APSSDC). This opportunity has been an enriching and transformative experience for me, and I am truly thankful for the support, guidance, and encouragement I have received along the way.

First and foremost, I extend my sincere regards Mrs.K.Meenakashi and Mrs. K. Narmada Mani my supervisors and mentors, for providing me with valuable insights, constant guidance, and unwavering support throughout the duration of the internship. Their expertise and encouragement have been instrumental in shaping the direction of this project. I would like to thank the entire team at Andhra Pradesh Skill Development Corporation (APSSDC) for fostering a collaborative and innovative environment. The camaraderie, knowledge sharing, and feedback I received from my colleagues significantly contributed to the development and success of this project.

In conclusion, I am honoured to have been a part of this internship program, and I look forward to leveraging the skills and knowledge gained to contribute positively to future endeavours.

Thank you.

Sincerely,

Lokeswari .Y ,AP21110011547.

lokeswari\_yarlagadda@srmap.edu.

**ABSTRACT**

Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too

high. The objective of the dataset is to diagnostically predict whether or not a patient has

diabetes, based on certain diagnostic measurements included in the dataset. Several

constraints were placed on the selection of these instances from a larger database. In

particular, all patients here are females at least 21 years old of Pima Indian heritage.

The diabetes is one of lethal diseases in the world. It is additional inventor of various

varieties of disorders foe example: coronary failure, blindness, urinary organ diseases In

such case the patient is required to visit a diagnostic center, to get their reports

after consulta.

**TABLE OF CONTENTS**

Sno Content Pg. No

1. Introduction………………………………. 5

2. System Requirements…………………… 6

3. Methodology………………………………. 7-8

4. Implementation & Results…………….9-36

5. Conclusion………………………………… 37

**INTRODUCTION**

Let's start with understanding what exploratory data analysis (EDA) is. It is an approach

to analyzing data sets to summarize their main characteristics, often using statistical

graphics and other data visualization methods. Simply put, it is the process of

investigating data. This blog is a guide to understanding EDA with an example

dataset.

We will explore this dataset and find out factors that contribute the most for diabetes

causation. We will also build Machine Learning Models that can help to predict whether

a person is diabetic or not and try to improve the model by performing Cross Validation

and hyperparameter tuning.

**SYSTEM REQUIREMENTS**

Software:

Excel (for the CSV data)

Jupyter notebook (for analysis and visualizing data)

Hardware:

System: Pentium i3 Processor

Hard Disk: 500 GB

Monitor: 15’’ LED

Input Devices: Keyboard Mouse

 Ram: 2 GB

**METHODOLOGY**

The development of the project Imports and Exports Analysis in India follows a systematic and iterative methodology to ensure the successful implementation of project objectives. The methodology encompassed several phases, including data collection, data preprocessing (including data cleaning) and data visualization.

Data Collection:

Dataset was compiled from public sources including government records and analyzed using Python tools.

Data Preprocessing:

Raw data went through cleaning and processing. In data cleaning we will sort out the null values remove or replace those values accordingly.

Data Visualization:

Here we will represent the data in the form of graphs and charts. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

Variations and changes in import and export values were studied. Relations with foreign exchange were analyzed to understand impact.

**IMPLEMENTATION & RESULTS**

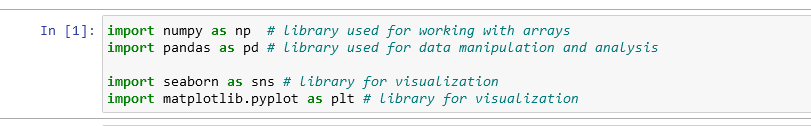
## IMPORTING THE LIBRARIES

import numpy as np # library used for working with arrays

import pandas as pd # library used for data manipulation and analysis

import seaborn as sns # library for visualization

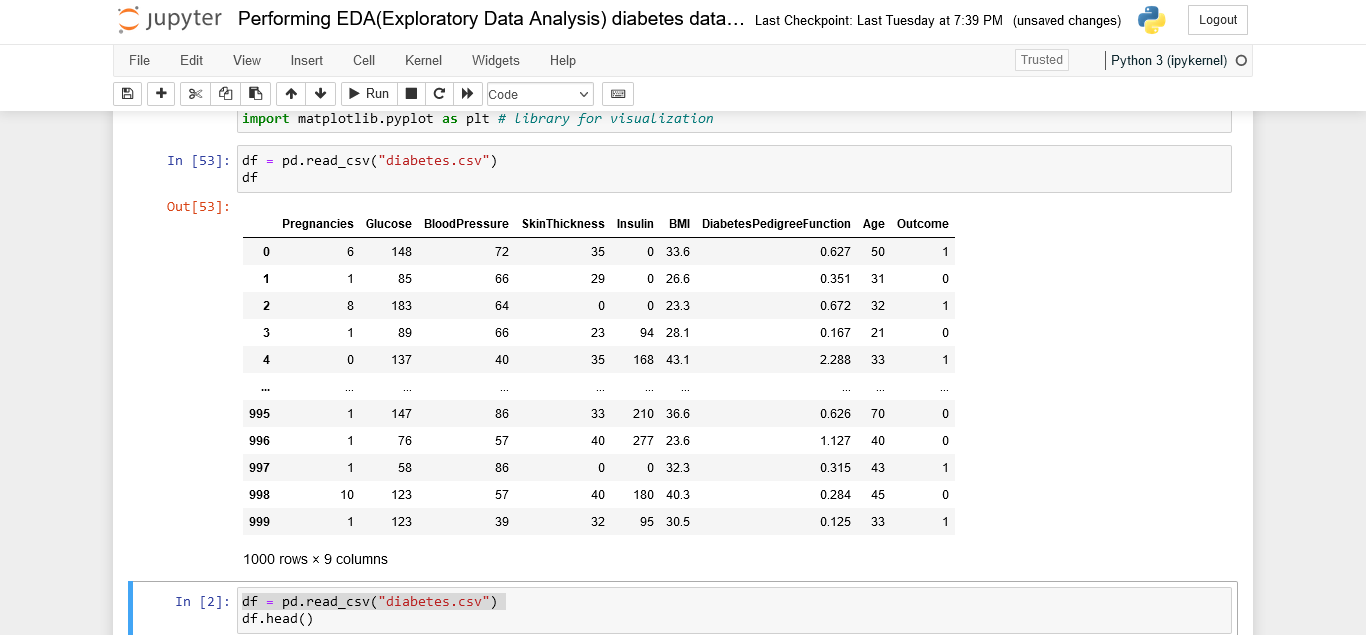
import matplotlib. Pyplot as plt # library for visualization



## READING THE DATASET

df = pd. read\_ csv("diabetes.csv")

READING THE DATA SET



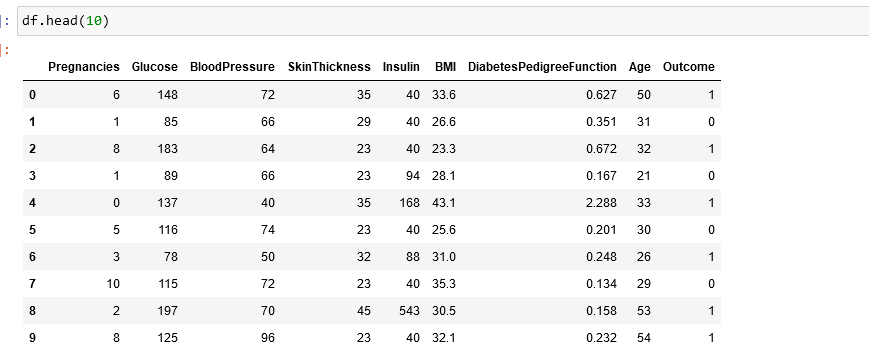
## BASIC FUNCTIONS ON DATA SET

#head () function returns specified no of rows from top

trade= pd. Read .csv ("imports and exports.csv")

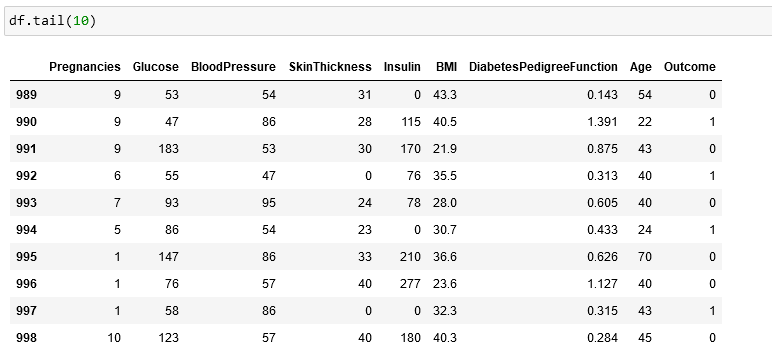
trade. Head (10)

Basic functions on data set



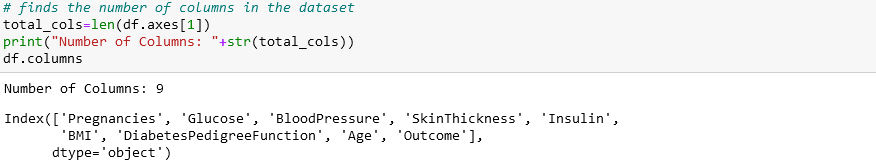
#tail () function returns specified no of rows from bottom

df=pd. read csv ("imports and exports.csv")

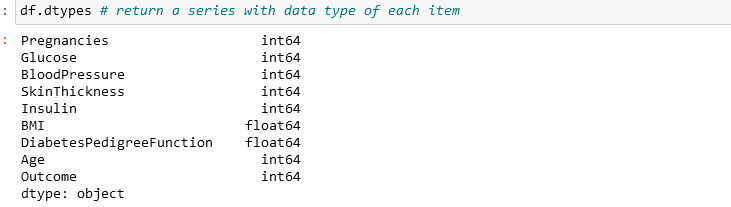


df. shape # to get current shape

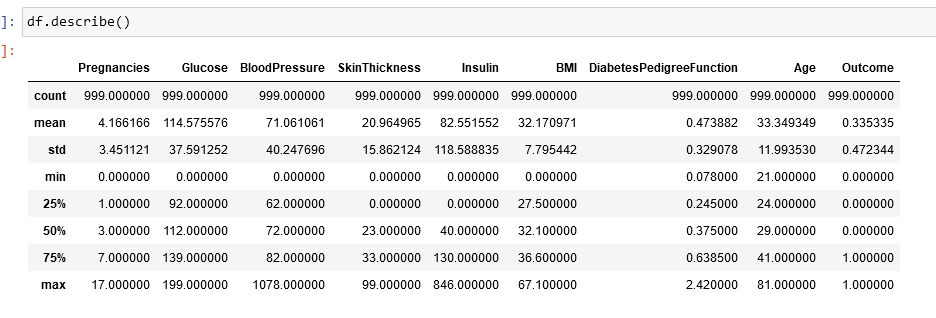
df.columns # to get the columns



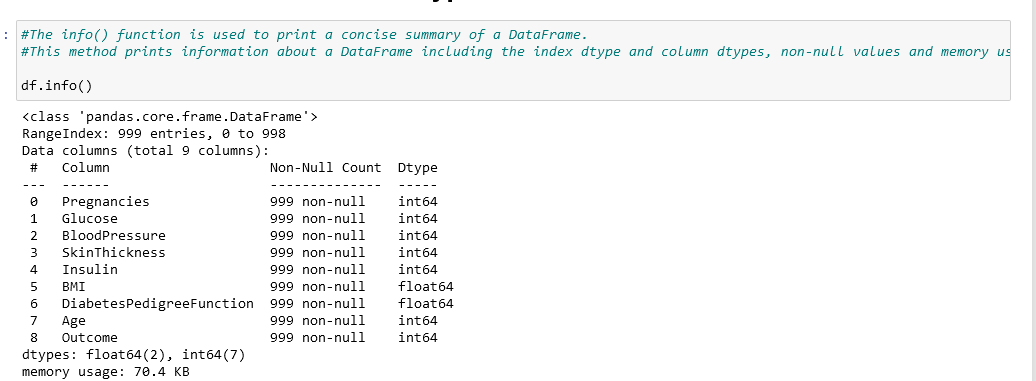
df. dtypes # returns a series with data type of each item



df. describe () # describe functions returns all these values like mean count etc.

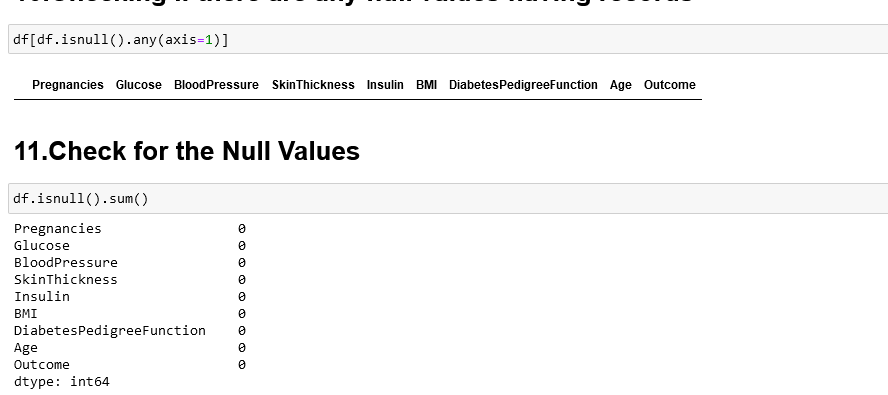


df.info () # data types of all variables in the dataset

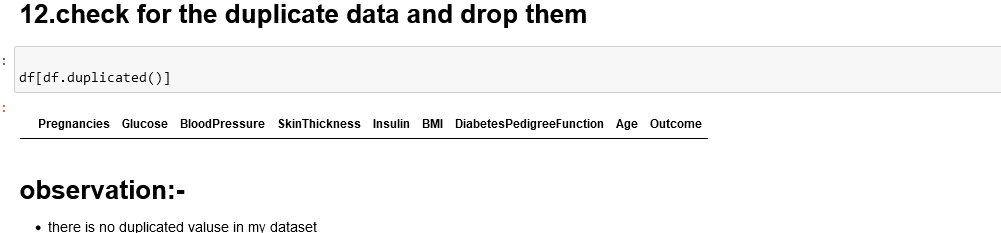


# Checking if there are any null values having records

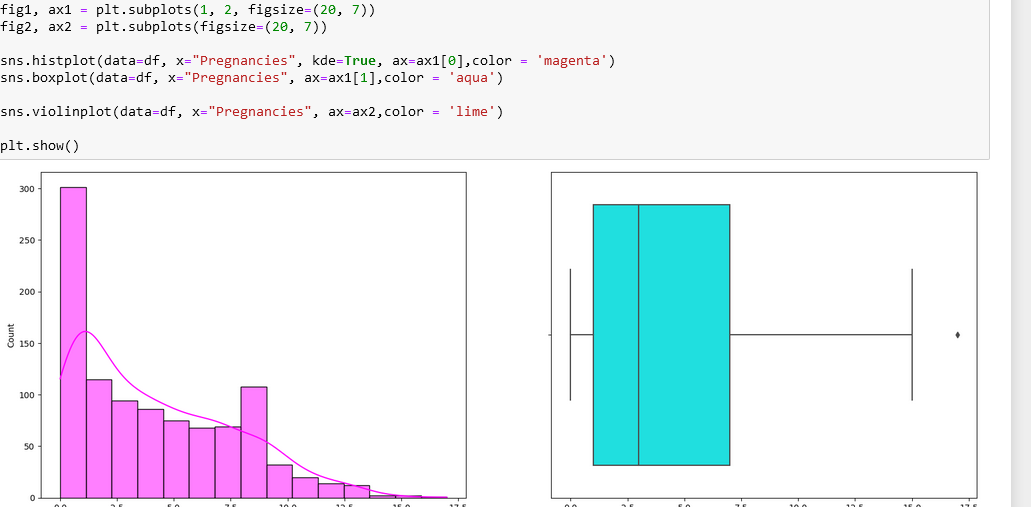
#check for the null values

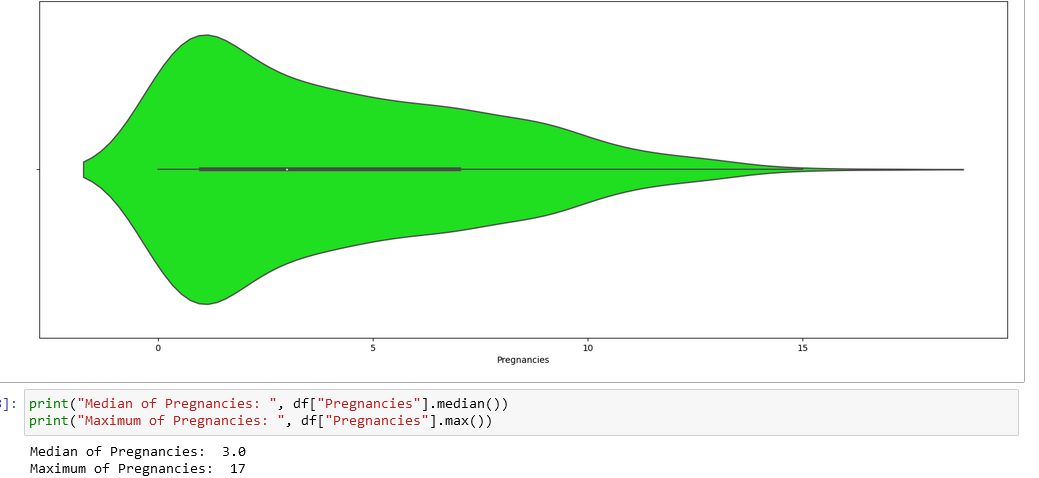


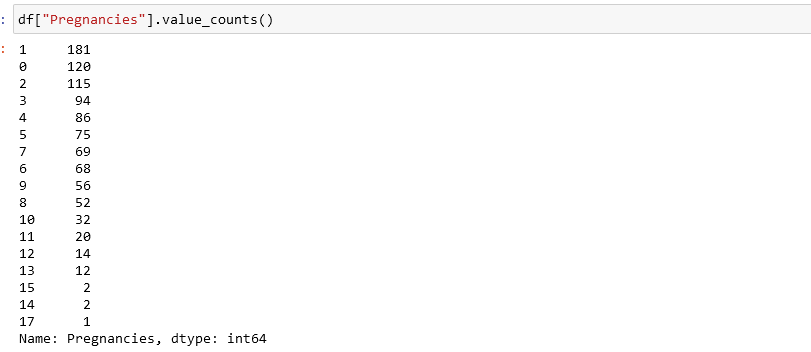
# Check for the duplicates



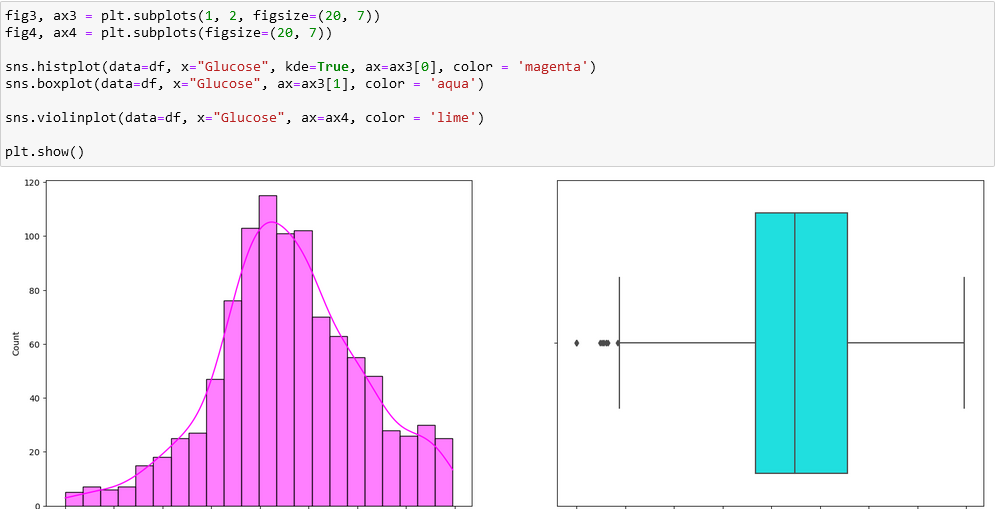
# Analysis of pregnancies

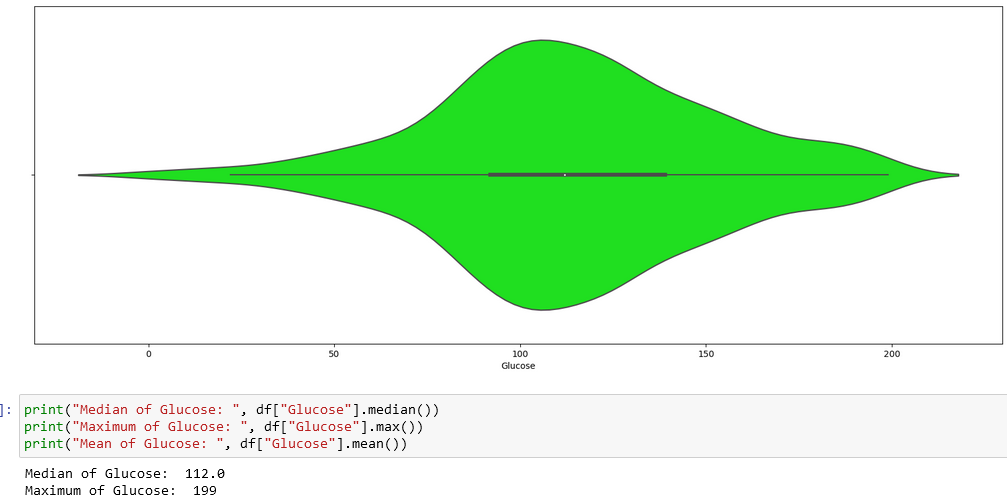






# analysis of glucose





# #Analysis of Blood Pressure

# 

# 

# #Analysis of Insulin

# 

# 

# #Analysis of BMI

# 

# 

# #Analysis of Diabetes Pedigree Function

# 

# #Analysis of Age

# 

# 

# #Exploratory Data Analysis

# 

# 

# #Getting a pairwise distribution between Glucose, Skin thickness and Diabetes pedigree function.

# 

# 

# #Studying the correlation between glucose and insulin using a Scatter Plot

# 

# #Let us explore the possibility of outliers using the Box Plot.

# 

# # Understanding the number of women in different age groups with diabetes.

# 

# #understanding the number of women in different age groups without diabetes

# 

# #And finally let us find and visualize the correlation between all variables

# 

# #Fill missing Values

# 

# 

# # separating the data and labels

# 

# 

# #Analyzing the Data

# #Scatter plot

# 

**Conclusion**

# The glucose levels may not alone help us in diagnosing diabetes. If you continue to do this

# analysis on other features as well, the predictions can be more accurate. Not all plots work

# for all the data, we have to try and test to find which fits the best!

# The highest number of Women without diabetes range between ages 22 to 33. Women

# between the age of 22 to 35 are at the highest risk of diabetes and also the is the highest

# number of those without diabetes.