Neuro Data Engineering
Online EDA Automation

High Level Design Document

Vamsi Alla

23rd October, 2022

INEURON

1. INTRODUCTION

The initial and the most crucial part of any model building is making the data set ready for next step. Howsoever it consumes 70-80% of the time which can be reduced. An initial knowledge of the data is very much required to go for advanced data mining. This work discusses the implementation of unmanned basic exploration of the data in order to reduce the time of the scientist so they can work on other important part of the work. And this work also explains the behind implementation of data mining and can be very much useful for new data analytics.

The solution proposed here is an Automated EDA which can be implemented to perform above mentioned use cases. Initially the Automated EDA will take the dataset from the client , will process the dataset into Dataframe and on this processed DataFrame further operations will be performed. The choice of operations are fairly dependent on the user and the copy of the performed operations will be made available to the client on the provided email.

PROBLEM STATEMENT

Create a web app application to perform Data cleaning, Feature engineering, and EDA. The web application should allow the user to perform various data transformation operations on the dataset with help of prebuild component. Users must be able to drag and drop the existing component at UI to perform any operation.

2. DATASET INFORMATION

Select the File form UI w.r.t file type and perform the following operations .

To create an automated exploration of data at the initial stage will make following processes faster:

To reduce coding so the exploration of the data can be made possible for management

To reduce time involved in basic exploration of data set and data mining

To demonstrate the working of exploring data to the data science enthusiast

To make the code readily available to the data scientists

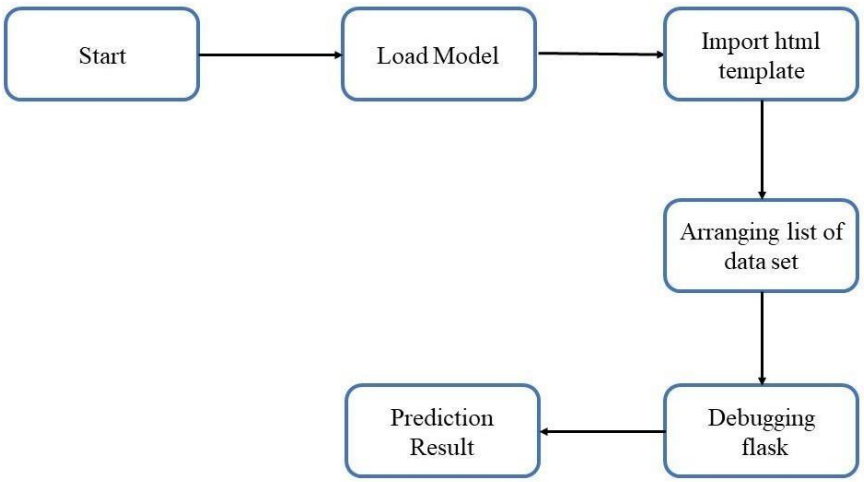
3. TOOLS USED

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn are used to build the whole model.

4. DESIGN DETAILS

4.1. Process flow

4.2. Deployment process



5. CONCLUSION

The project is designed in flask; hence it is accessible to everyone. The above designing process will help automated exploration of data at the initial stage will make following processes faster and reduce time involved in basic exploration of data set and data mining and demonstrate the working of exploring data to the data science enthusiast. The UI is made to be user-friendly so that the user will not need much knowledge of any tools but will just need the information for results.