zeroc<>de

learning

Learning Data Analytics Made Easy

USER GUIDE

k-NN ANALYSIS

TABLE OF CONTENTS

INDEX

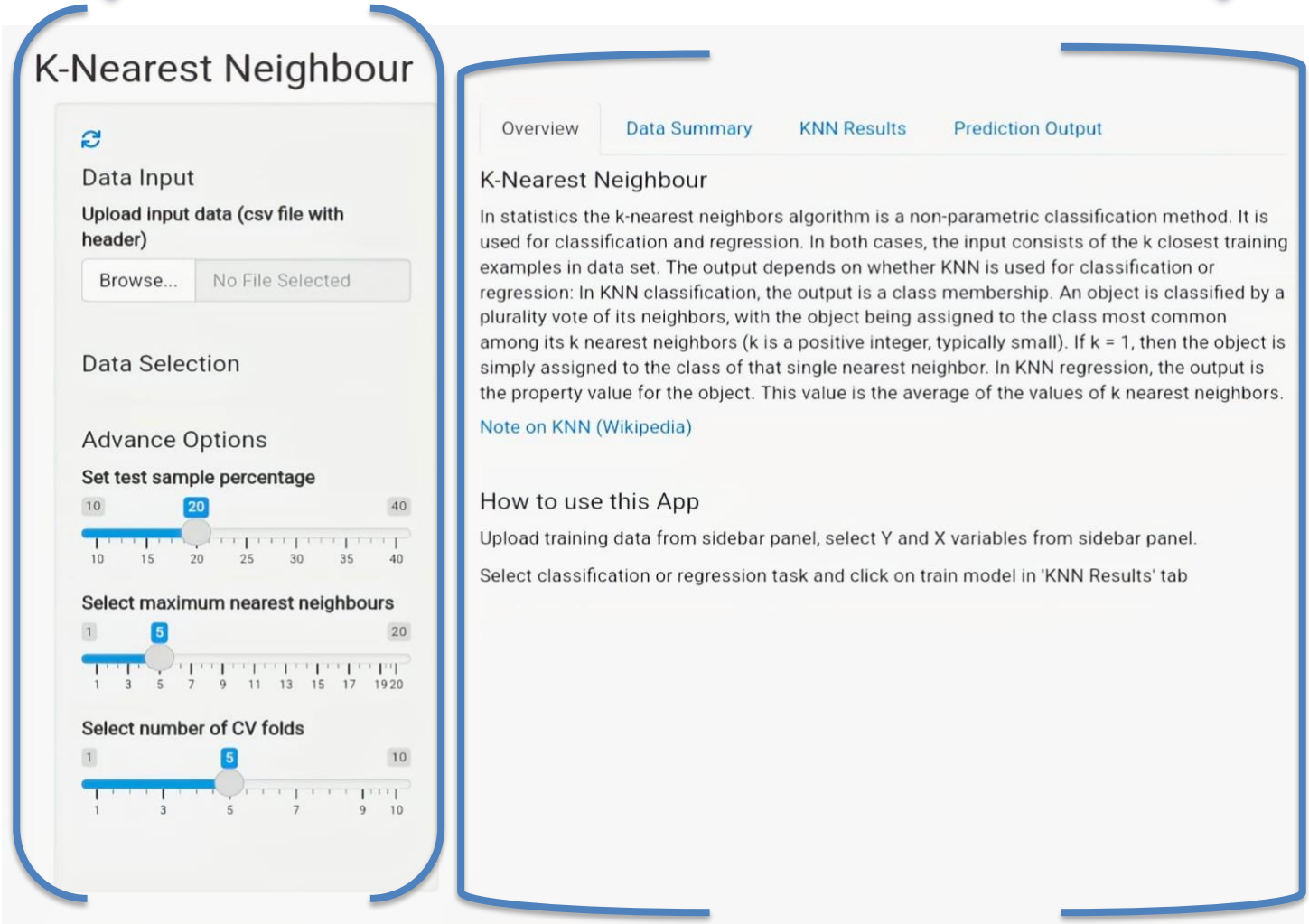
k-NEAREST NEIGHBOUR ANALYSIS

01

KNN, The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. Therefore, you can use the KNN algorithm for applications that require high accuracy but that do not require a human-readable model. The quality of the predictions depends on the distance measure.

LEFT PANEL (INPUT AREA)

OPERATIONAL ANALYSIS TAB (MAIN PANEL)



LEFT PANEL (INP)

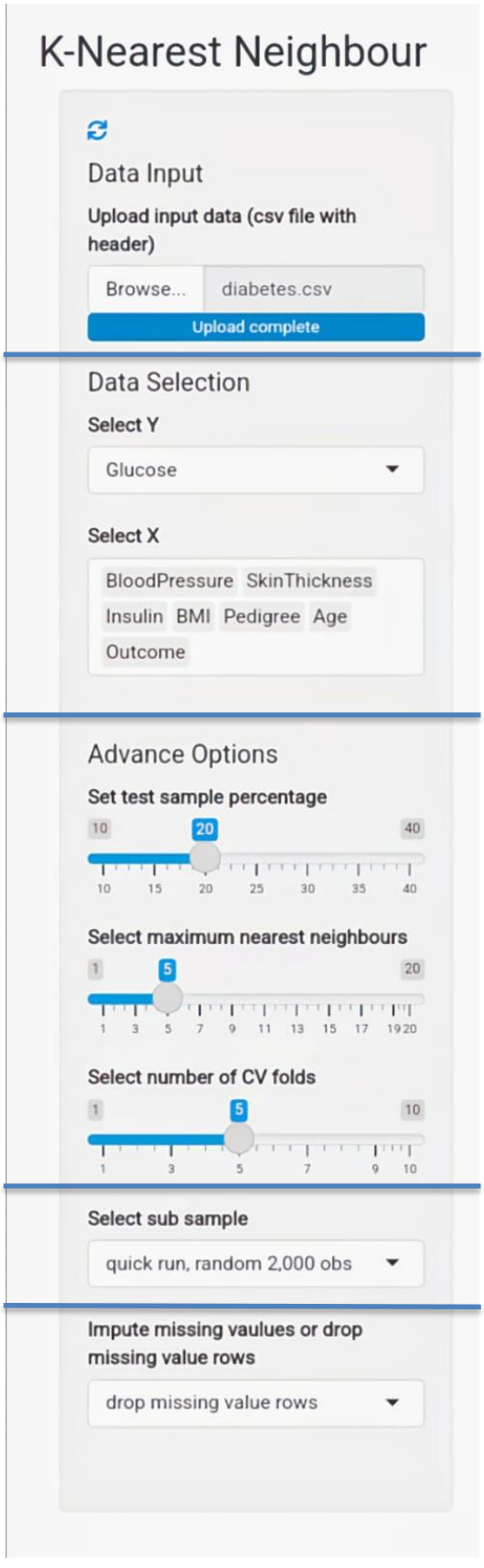
Upload your dataset here

Select your favorable variables required to base the analysis

Apply any changes under advanced options according to the needs.

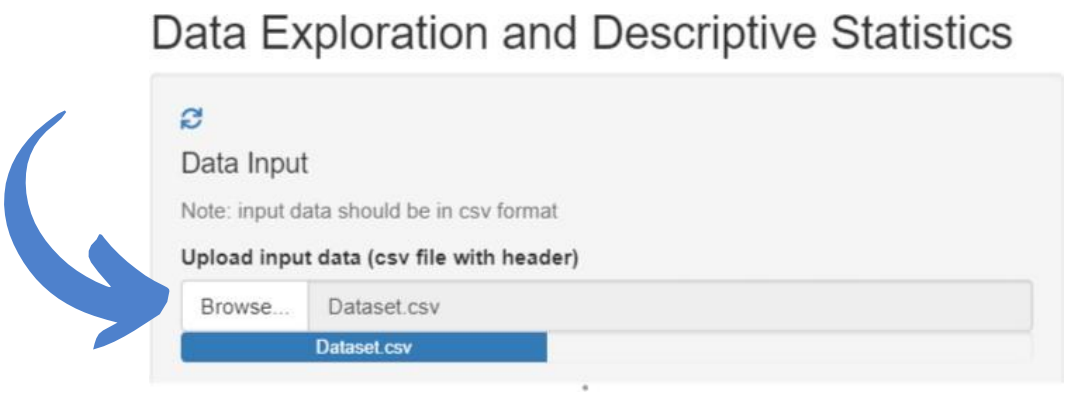
Select the subsamples or the whole data for testing.

Deal with missing values either drop or immute it.



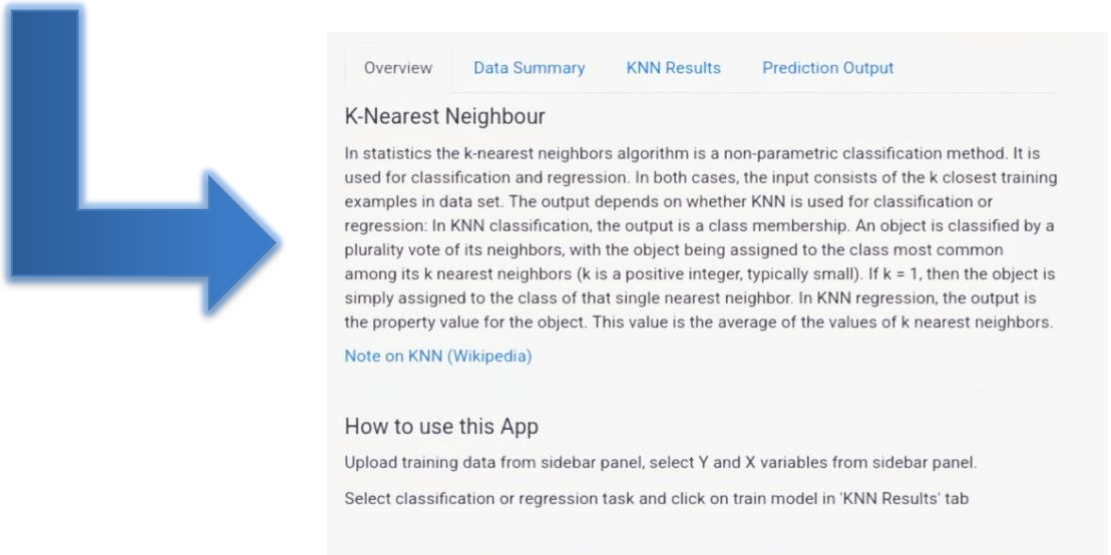
DATA INPUT (UPLOADING DATASET)

- Click on browse
- Select the datafile that is in the form of csv format.(Ex program.csv)
- Browse the file and select the data to train your model for prediction.
- Top rows of the dataset should be of ‘variable names’.



OVERVIEW TAB

This tab provides you with relevant study resources, tutorials, sample datasets and a short overview to start with, which helps you understand and comprehend your data correctly. This tab also provides you the basic idea about k-NN, gives sample data and provides the description about k-NN analysis.

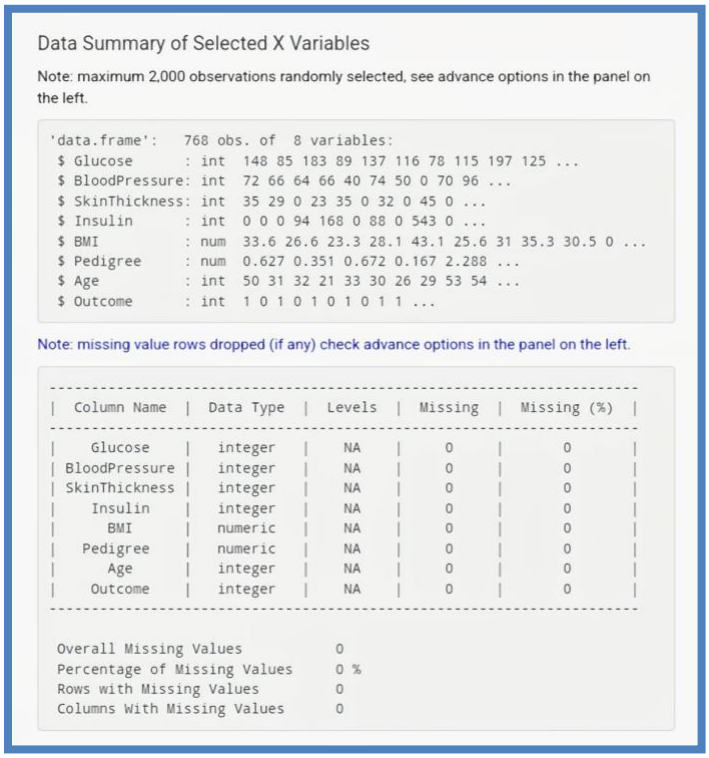


DATA SUMMARY TAB

It is very important to understand our data completely to infer meaningful insights and to get an overview of all the data points as a whole, but it is quite impossible to analyze thousand data points manually.

The **'Data Summary'** option enables you to get a comprehensive evaluation through statistical measures that help us form the basis of our analysis.

It will display all the 'descriptive analytics' measures including data variables, Data frame , overall analysis for all the data variables present in the dataset. we can review the uploaded data and the contents of it, A brief summary of the data can be seen it includes range of data values,missing and null values etc. It also segregates dataset variables into respective data types, such as integer, whole numbers, character etc.



Data Summary of Selected X Variables

Note: maximum 2,000 observations randomly selected, see advance options in the panel on the left.

```
'data.frame':   768 obs. of  8 variables:
 $ Glucose      : int  148 85 183 89 137 116 78 115 197 125 ...
 $ BloodPressure: int  72 66 64 66 40 74 50 0 70 96 ...
 $ SkinThickness: int  35 29 0 23 35 0 32 0 45 0 ...
 $ Insulin      : int  0 0 0 94 168 0 88 0 543 0 ...
 $ BMI          : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ Pedigree     : num  0.627 0.351 0.672 0.167 2.288 ...
 $ Age          : int  50 31 32 21 33 30 26 29 53 54 ...
 $ Outcome      : int  1 0 1 0 1 0 1 0 1 1 ...
```

Note: missing value rows dropped (if any) check advance options in the panel on the left.

Column Name	Data Type	Levels	Missing	Missing (%)
Glucose	integer	NA	0	0
BloodPressure	integer	NA	0	0
SkinThickness	integer	NA	0	0
Insulin	integer	NA	0	0
BMI	numeric	NA	0	0
Pedigree	numeric	NA	0	0
Age	integer	NA	0	0
Outcome	integer	NA	0	0

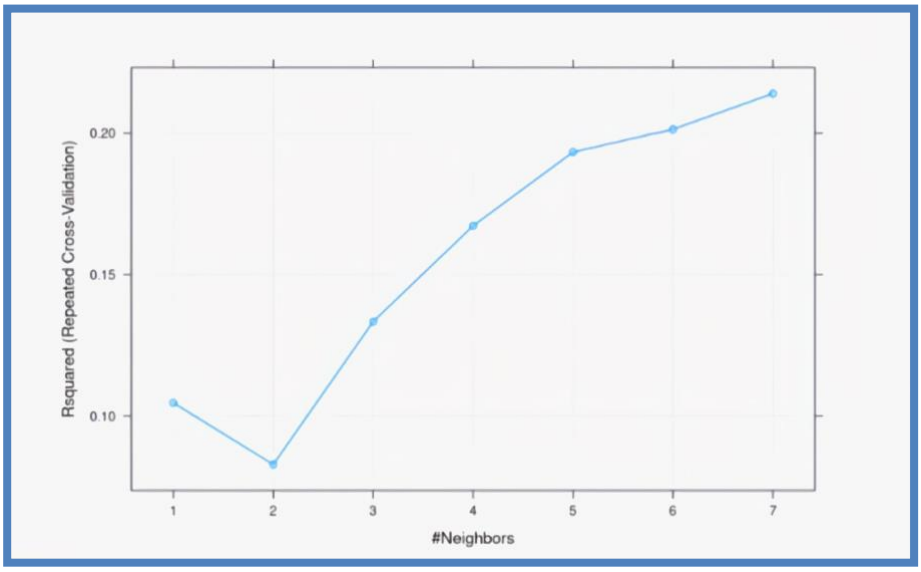
Overall Missing Values	0
Percentage of Missing Values	0 %
Rows with Missing Values	0
Columns With Missing Values	0

This includes the minimum value maximum value ,data frames,variables and related info about the summary after analysis is finished.

Info about missing values

Use the left panel to transform selected variables as per the requirement of analysis ,correspondingly the data summary will also change.

k-NN RESULTS TAB



The k-nearest neighbors (k-NN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows. During training phase, k-NN arranges the data (sort of indexing process) in order to find the closest neighbors efficiently during the inference phase. Otherwise, it would have to compare each new case during inference with the whole dataset making it quite inefficient. Train the model either for regression or classification based on the data and the prediction.

K-NN REGRESSION : k-NN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

K-NN CLASSICATION : In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small).

Use the left panel to modify/deal with the outliers identified here.