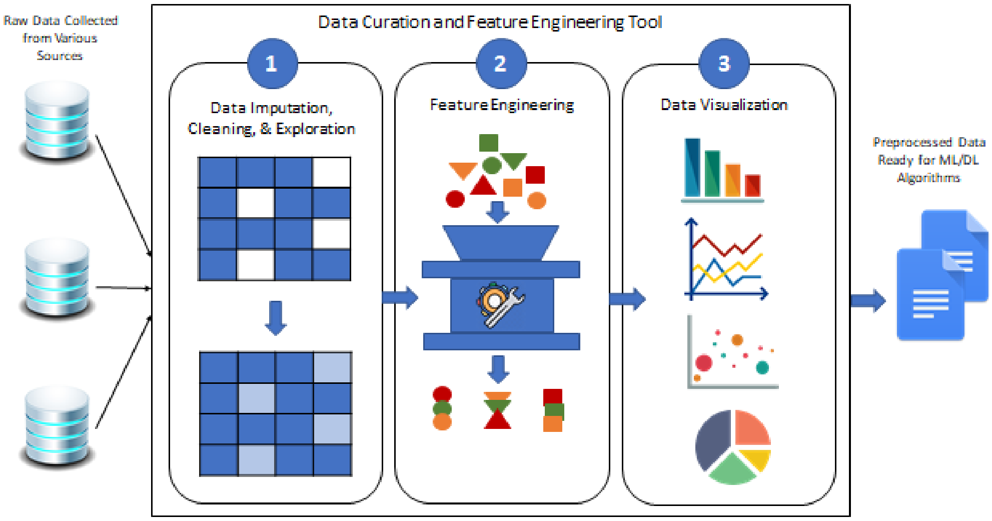

Raw Data Collected from Various Sources
Data Curation and Feature Engineering Tool
1
Data Imputation, Cleaning, & Exploration
2
Feature Engineering
3
Data Visualization
Preprocessed Data Ready for ML/DL Algorithms

DCFE Web Application User Guide

For Vadu HDSS Dataset

DST Sponsored Project-ICPS Division

Department of Computer Applications
NIT-Trichy
Prepared (Rev. 21/01/22)

Table of Contents

Introduction

In the process of data analysis, it is observed that 60% of the time is spent in cleaning and preprocessing the data. To overcome this drawback and to reduce the time spent in preprocessing we have developed a tool named Data Curation and Feature Engineering Tool (DCFE Tool). Epidemiologists, Researchers or any kind of Data analysts can use this web based application tool to upload, explore and analyze the epidemic/pandemic or health care datasets. The tool provides the provision for the user to upload datasets which are in .csv or .xlsx file format. The complete detail of the uploaded data like the size, No. of rows and columns, Percentage of missing value, Numerical and Categorical and date-time attributes can be viewed using ‘Dataset Overview’. The ‘data exploration’ projects the statistical values such as mean, median, skewness, kurtosis, SD and No. of missing values of each attribute in the dataset.

Data imputation and Feature engineering are the two primary modules of this tool. The missing value present in the dataset is handled using the data imputation module where the missing values can be completely dropped or it can be filled using any suitable method like ‘Attribute wise drop NaN’, ‘Attribute wise Fill NaN’, Machine learning model based imputation like ‘KNN’ and ‘Iterative imputation’. In Feature Engineering module, the user can use two ways of transformation. Both numerical and categorical data are transformed into new feature which are ready for any kind of algorithmic computation. The different encoding methods used here for transformation of both numerical and categorical features are Binning, Normalization, Log transform, Label, Ordinal, Binary, One-hot , Count frequency encoding.

The user can generate new features using Date-time features. If user wants to delete an unwanted feature, ‘Remove unwanted feature’ function can be used. If a user wants to roll back to the previous action performed ‘Undo’ function can be used. The user can download his/her processed data using the function ‘download processed data’ or the user can also download the original uploaded dataset. For repeating the process with another dataset then the user can click ‘Home’ function to return to the front page of the tool.

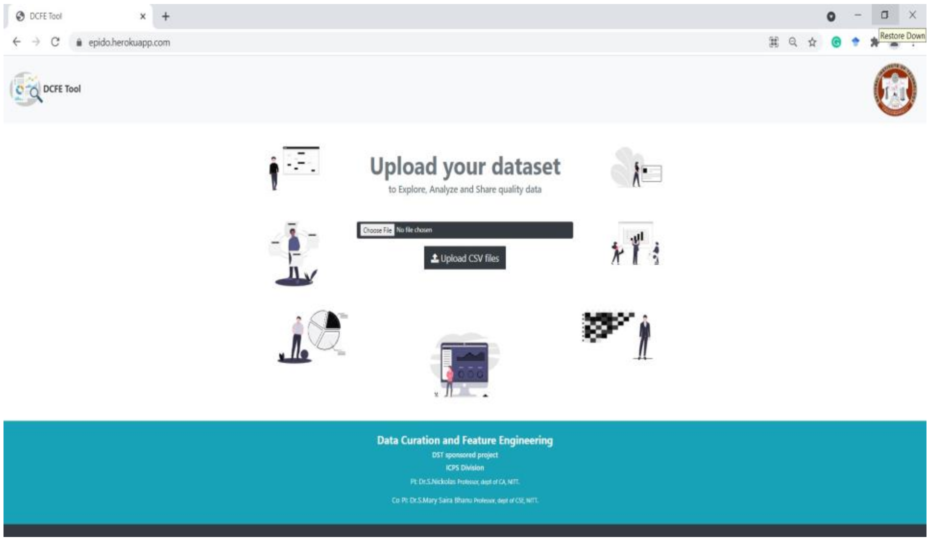
To visualize the uploaded data, the user can use the ‘Visualization’. The function will give provision to use different charts such as bar, scatter, pie charts, which will take the features as inputs to display the counts of each category.

Further, the user can visualize the percentage of categorical and numerical features in the uploaded data using a pie chart. The user can also visualize the skewness, kurtosis and No. of missing value of each attribute in the dataset.

The correlation between the features in the dataset is shown using Heatmap. Finally, the outlier in each attribute of the data can be visualized using the box plot.

This document guides a user through the necessary steps to preprocess and download the curated data which can be used for any further computation with any Machine learning algorithm.

Front page of the project



Upload your dataset

to Explore, Analyze and Share quality data

Data Curation and Feature Engineering

DST sponsored project

ICPS Division

PI: Dr.S.Nickolas Professor, dept of CA, NITT.

Co PI: Dr.S.Mary Saira Bhanu Professor, dept of CSE, NITT.

DCFE Web Application Tool Link

Users can directly use our Web Application DCFE Tool by clicking on the following link: https://epido.herokuapp.com/

Upload your dataset

The front page of our tool has a provision to upload datasets. The dataset must be in either .csv or .xlsx file format.

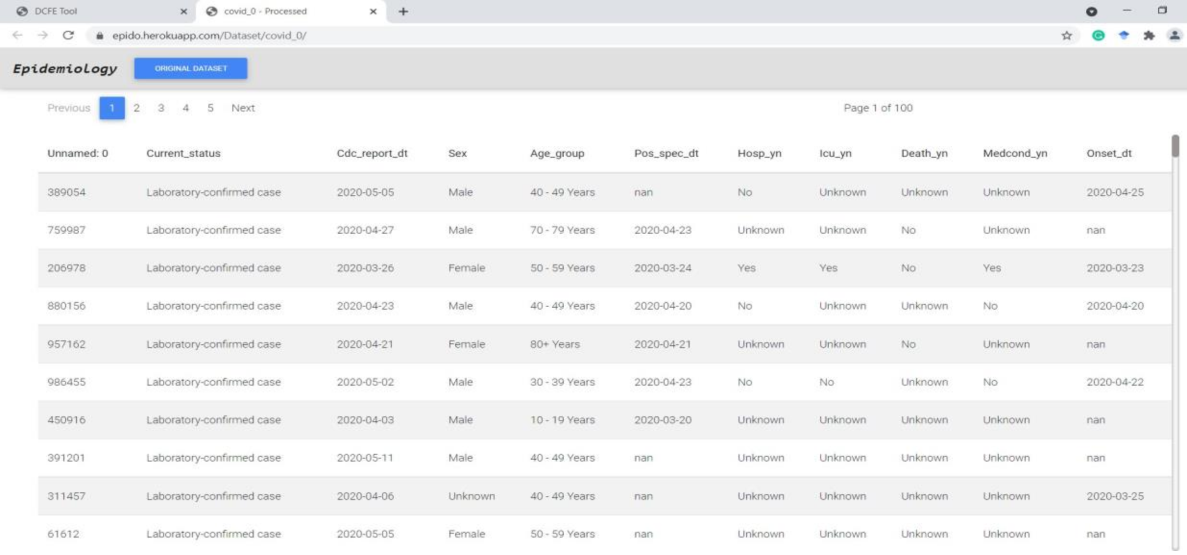
Choose File

If the user clicks 'Choose File' button, it will redirect the user to the file location to choose appropriate dataset file.

Upload CSV files

Users can click this button to upload the dataset.

Dataset view



Unnamed: 0	Current_status	Cdc_report_dt	Sex	Age_group	Pos_spec_dt	Hosp_yn	Icu_yn	Death_yn	Medcond_yn	Onset_dt
389054	Laboratory-confirmed case	2020-05-05	Male	40 - 49 Years	nan	No	Unknown	Unknown	Unknown	2020-04-25
759987	Laboratory-confirmed case	2020-04-27	Male	70 - 79 Years	2020-04-23	Unknown	Unknown	No	Unknown	nan
206978	Laboratory-confirmed case	2020-03-26	Female	50 - 59 Years	2020-03-24	Yes	Yes	No	Yes	2020-03-23
880156	Laboratory-confirmed case	2020-04-23	Male	40 - 49 Years	2020-04-20	No	Unknown	Unknown	No	2020-04-20
957162	Laboratory-confirmed case	2020-04-21	Female	80+ Years	2020-04-21	Unknown	Unknown	No	Unknown	nan
986455	Laboratory-confirmed case	2020-05-02	Male	30 - 39 Years	2020-04-23	No	No	Unknown	No	2020-04-22
450916	Laboratory-confirmed case	2020-04-03	Male	10 - 19 Years	2020-03-20	Unknown	Unknown	Unknown	Unknown	nan
391201	Laboratory-confirmed case	2020-05-11	Male	40 - 49 Years	nan	Unknown	Unknown	Unknown	Unknown	nan
311457	Laboratory-confirmed case	2020-04-06	Unknown	40 - 49 Years	nan	Unknown	Unknown	Unknown	Unknown	2020-03-25
61612	Laboratory-confirmed case	2020-05-05	Female	50 - 59 Years	nan	Unknown	Unknown	Unknown	Unknown	nan

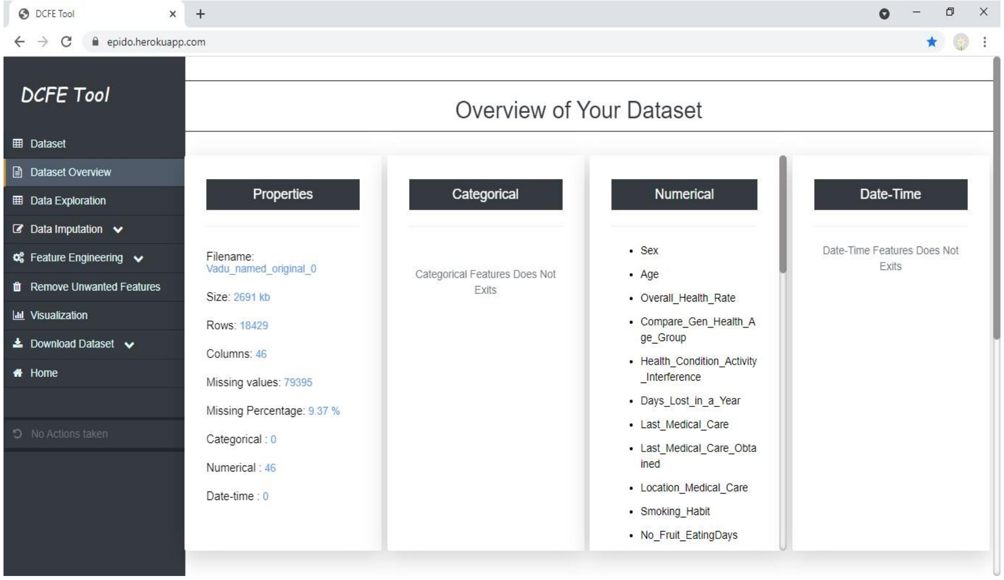
Dataset

User can view the uploaded dataset by selecting the menu **‘Dataset’**, where he/she can view the complete dataset page wise.

Original Dataset

If the user wants to compare the processed file with the original file, then the button **’ORIGINAL DATASET’** can be clicked to download the original file.

Dataset Overview



Overview of Your Dataset

Properties

Filename: Vadu_named_original_0

Size: 2691 kb

Rows: 18429

Columns: 46

Missing values: 79395

Missing Percentage: 9.37 %

Categorical : 0

Numerical : 46

Date-time : 0

Categorical

Categorical Features Does Not Exits

Numerical

- Sex
- Age
- Overall_Health_Rate
- Compare_Gen_Health_Age_Group
- Health_Condition_Activity_Interference
- Days_Lost_in_a_Year
- Last_Medical_Care
- Last_Medical_Care_Obtained
- Location_Medical_Care
- Smoking_Habit
- No_Fruit_EatingDays

Date-Time

Date-Time Features Does Not Exits

Properties

If the user wants to know about the details of the uploaded .csv or .xlxs file, the menu **'Properties'** will provide all the necessary information like Filename, Size, Rows, Columns, Missing values, categorical, Numerical and Date-time features.

Categorical

If the user wants to view how many categorical data is there in a dataset, the menu **'Categorical'** will expose those details.

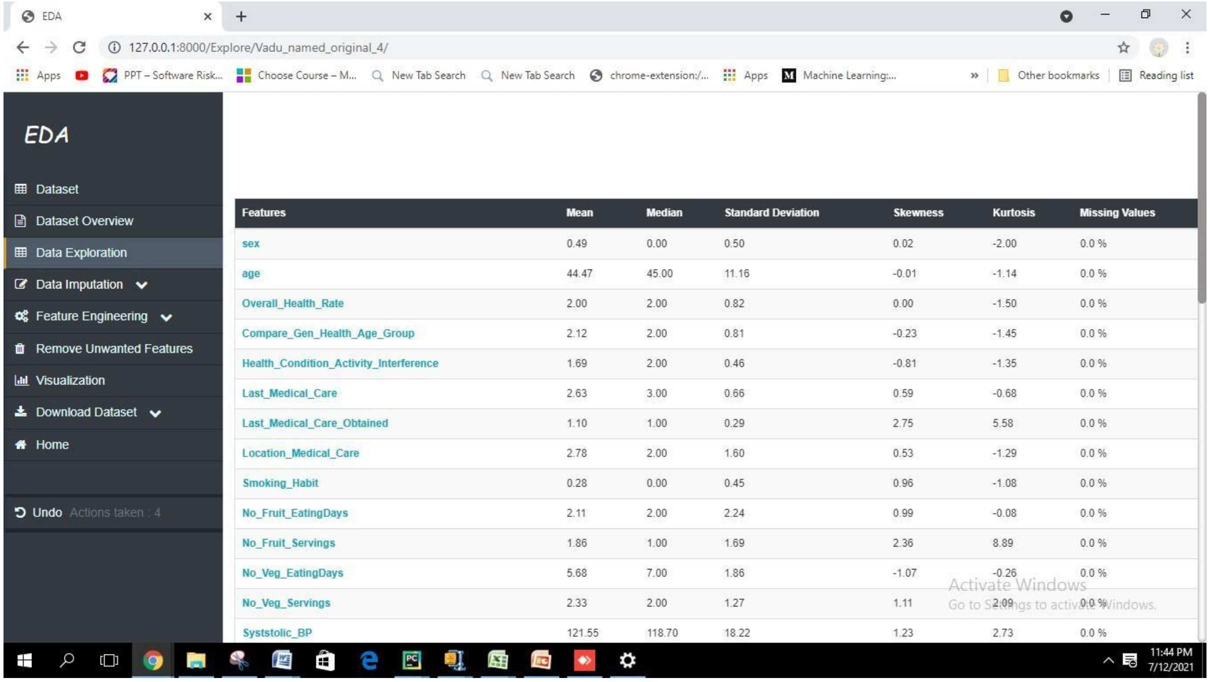
Numerical

If the user wants to see how many numerical data present in a dataset, the menu **'Numerical'** will expose those details.

Date-time

Once the file is uploaded the user is provided with the provision **('Date-time')** to see whether the dataset contains Date-time features or not.

Dataset Exploration



Features	Mean	Median	Standard Deviation	Skewness	Kurtosis	Missing Values
sex	0.49	0.00	0.50	0.02	-2.00	0.0 %
age	44.47	45.00	11.16	-0.01	-1.14	0.0 %
Overall_Health_Rate	2.00	2.00	0.82	0.00	-1.50	0.0 %
Compare_Gen_Health_Age_Group	2.12	2.00	0.81	-0.23	-1.45	0.0 %
Health_Condition_Activity_Interference	1.69	2.00	0.46	-0.81	-1.35	0.0 %
Last_Medical_Care	2.63	3.00	0.66	0.59	-0.68	0.0 %
Last_Medical_Care_Obtained	1.10	1.00	0.29	2.75	5.58	0.0 %
Location_Medical_Care	2.78	2.00	1.60	0.53	-1.29	0.0 %
Smoking_Habit	0.28	0.00	0.45	0.96	-1.08	0.0 %
No_Fruit_EatingDays	2.11	2.00	2.24	0.99	-0.08	0.0 %
No_Fruit_Servings	1.86	1.00	1.69	2.36	8.89	0.0 %
No_Veg_EatingDays	5.68	7.00	1.86	-1.07	-0.26	0.0 %
No_Veg_Servings	2.33	2.00	1.27	1.11	2.09	0.0 %
Syststolic_BP	121.55	118.70	18.22	1.23	2.73	0.0 %

Mean

If a user wants to know the mean/average values of each category, then the menu 'Mean' will generate the values for each feature.

Median

If a user wants to know the median value of each feature, then the menu 'Median' will showcase the values.

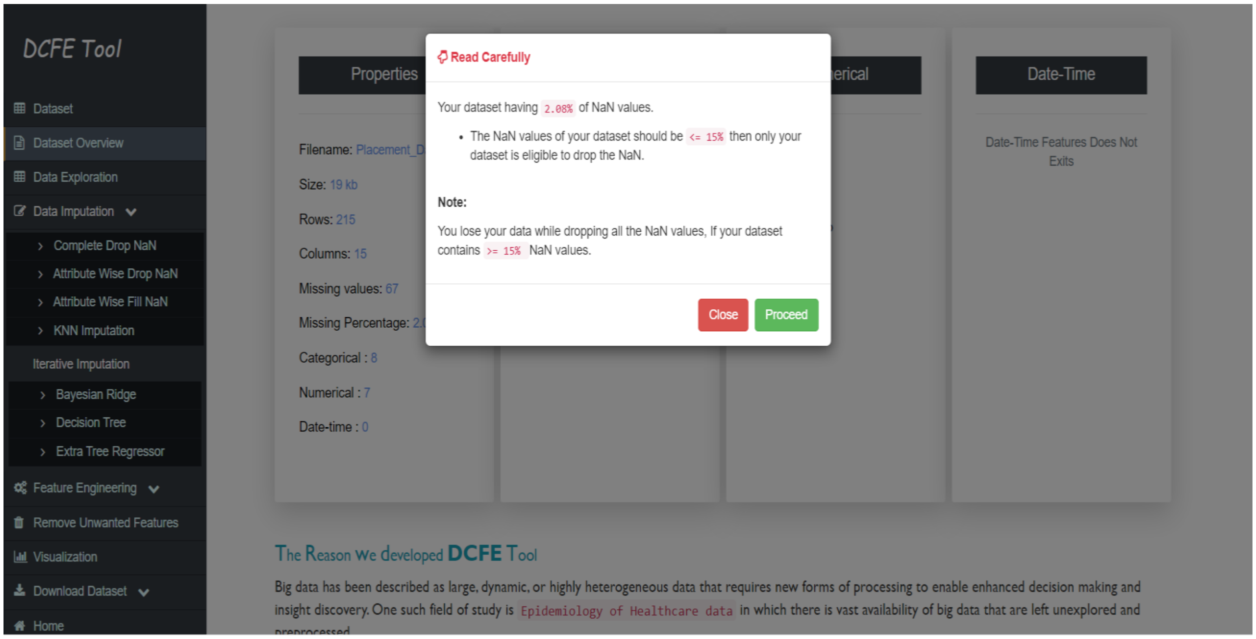
Standard deviation, Skewness, kurtosis

If a user wants to know the median values of each feature, then the menu '**Standard deviation, Skewness, kurtosis** will showcase the values.

Missing values percentage

The user will come to know about the missing percentage for each attribute.

Data Imputation-Complete Drop NaN



Filename: Placement_D

Size: 19 kb

Rows: 215

Columns: 15

Missing values: 67

Missing Percentage: 2.

Categorical : 8

Numerical : 7

Date-time : 0

Date-Time Features Does Not Exits

Read Carefully

Your dataset having 2.08% of NaN values.

- The NaN values of your dataset should be <= 15% then only your dataset is eligible to drop the NaN.

Note:

You lose your data while dropping all the NaN values, If your dataset contains >= 15% NaN values.

The Reason We developed **DCFE** Tool

Big data has been described as large, dynamic, or highly heterogeneous data that requires new forms of processing to enable enhanced decision making and insight discovery. One such field of study is `Epidemiology of Healthcare data` in which there is vast availability of big data that are left unexplored and preprocessed

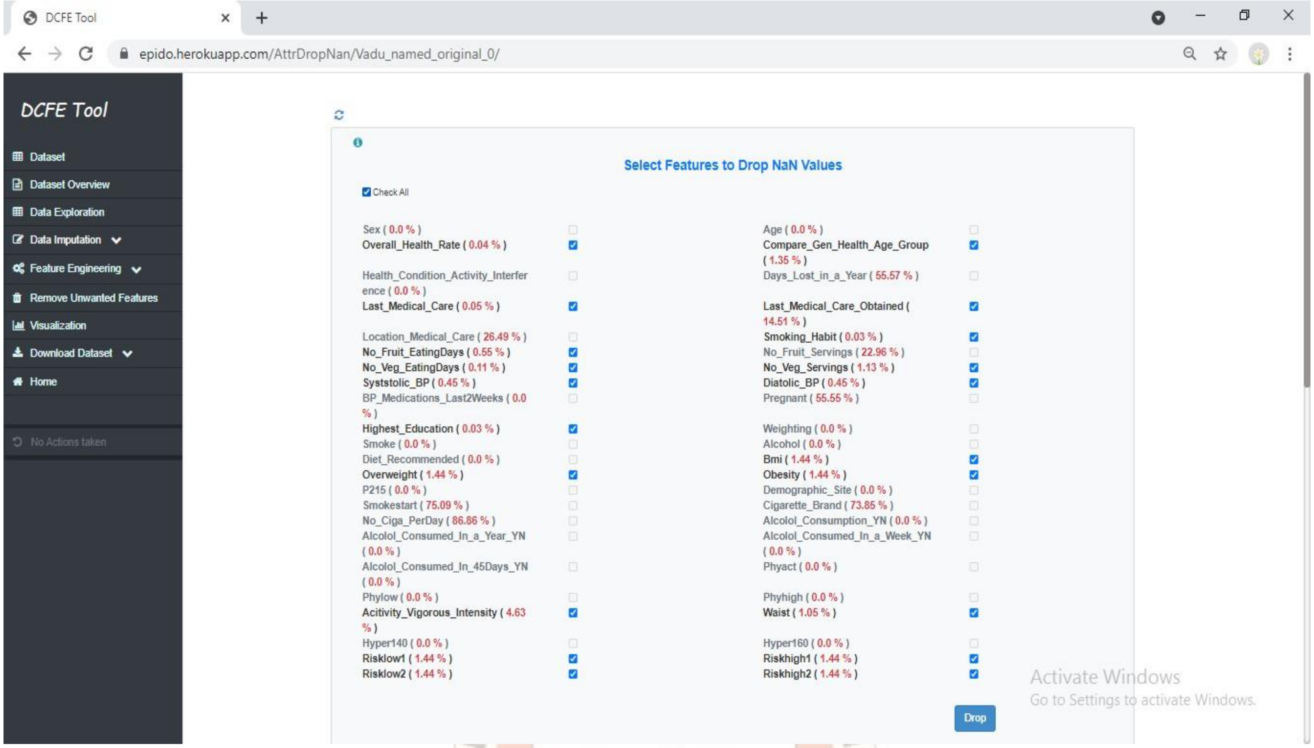
Proceed

When the users click the link **'Complete Drop NaN'** it will open a Dialog Box showing the rules for dropping the NaN values. The users can click the button **'Proceed'** after reading the rules for dropping the NaN values.

Close

The users can click **'close'** button after finish reading the mentioned rules.

Data Imputation- Attribute Wise Drop NaN



Select Features to Drop NaN Values

☑ Check All

Sex (0.0 %)	□	Age (0.0 %)	□
Overall_Health_Rate (0.04 %)	☑	Compare_Gen_Health_Age_Group (1.35 %)	☑
Health_Condition_Activity_Interference (0.0 %)	□	Days_Lost_in_a_Year (55.57 %)	□
Last_Medical_Care (0.05 %)	☑	Last_Medical_Care_Obtained (14.51 %)	☑
Location_Medical_Care (26.49 %)	□	Smoking_Habit (0.03 %)	☑
No_Fruit_EatingDays (0.55 %)	☑	No_Fruit_Servings (22.96 %)	□
No_Veg_EatingDays (0.11 %)	☑	No_Veg_Servings (1.13 %)	☑
Syststolic_BP (0.45 %)	☑	Diatolic_BP (0.45 %)	☑
BP_Medications_Last2Weeks (0.0 %)	□	Pregnant (55.55 %)	□
Highest_Education (0.03 %)	☑	Weighting (0.0 %)	□
Smoke (0.0 %)	□	Alcohol (0.0 %)	□
Diet_Recommended (0.0 %)	□	Bmi (1.44 %)	☑
Overweight (1.44 %)	☑	Obesity (1.44 %)	☑
P215 (0.0 %)	□	Demographic_Site (0.0 %)	□
Smokestart (75.09 %)	□	Cigarette_Brand (73.85 %)	□
No_Ciga_PerDay (86.86 %)	□	Alcolol_Consumption_YN (0.0 %)	□
Alcolol_Consumed_In_a_Year_YN (0.0 %)	□	Alcolol_Consumed_In_a_Week_YN (0.0 %)	□
Alcolol_Consumed_In_45Days_YN (0.0 %)	□	Phyact (0.0 %)	□
Phylow (0.0 %)	□	Phyhigh (0.0 %)	□
Acitivity_Vigorous_Intensity (4.63 %)	☑	Waist (1.05 %)	☑
Hyper140 (0.0 %)	□	Hyper160 (0.0 %)	□
Risklow1 (1.44 %)	☑	Riskhigh1 (1.44 %)	☑
Risklow2 (1.44 %)	☑	Riskhigh2 (1.44 %)	☑

Drop

Select Features to Drop NaN values

If the user wants to drop a feature which has NaN values, then this menu should be clicked by the user. The check boxes are there for each feature. To drop the missing values in specific attributes/features when there are less than 15% of missing values in it. The features with less than 15% missing values are shown in bold embossed i.e enabled to check the check box.

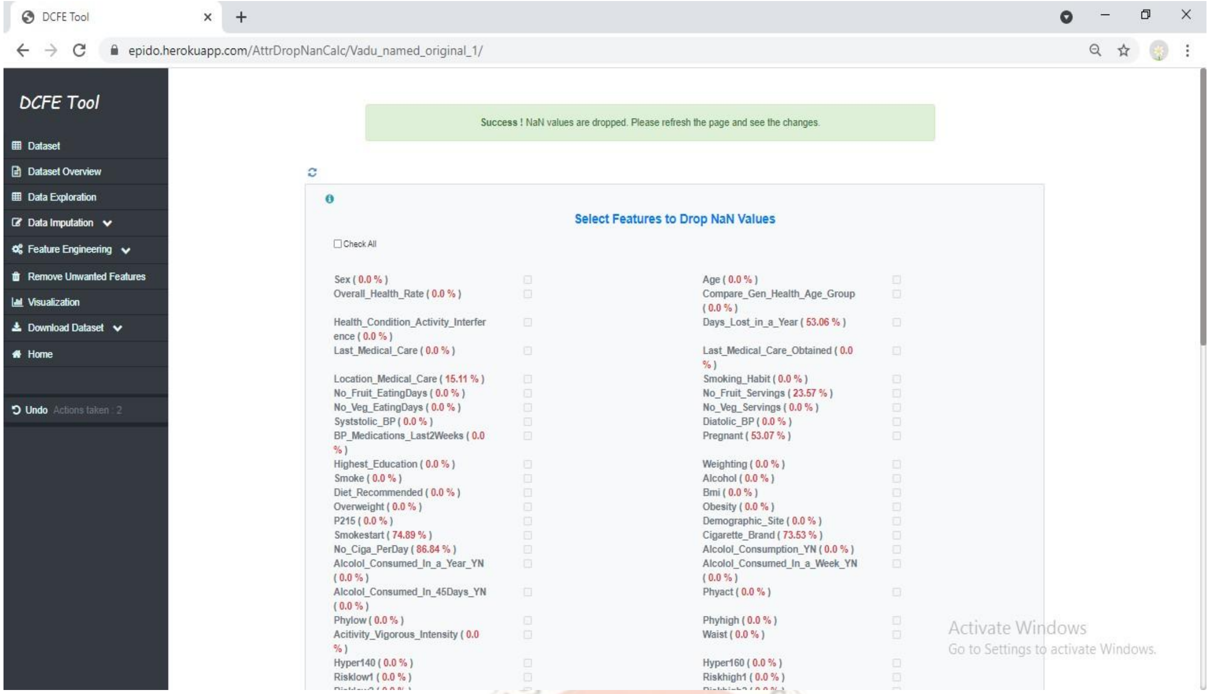
Check all

If a user wants to perform drop nan values on all attributes, then 'check all' option can be used.

Drop

If user wants to do drop the feature after selecting the feature, the user has to click the 'Drop' button. Then a message will be displayed that the selected feature is deleted. (**Success! NaN values are dropped. Please refresh the page and see the changes**)

Data Imputation- Attribute Wise Drop NaN



Success ! NaN values are dropped. Please refresh the page and see the changes.

Select Features to Drop NaN Values

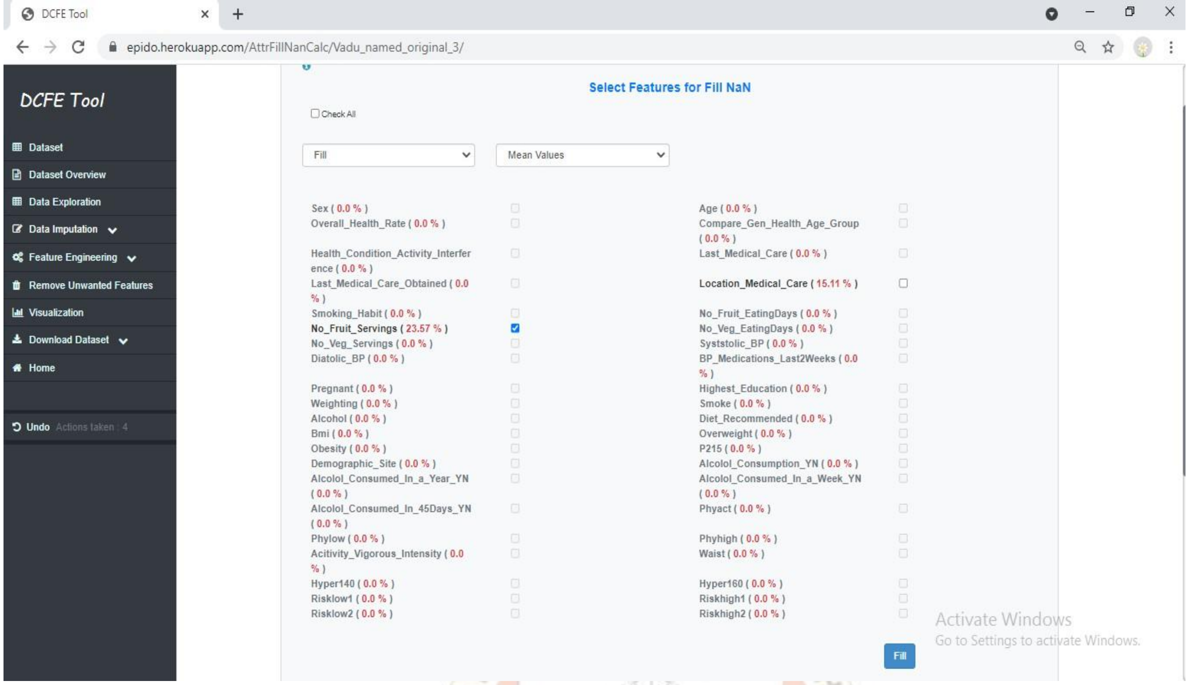
☐ Check All

Sex (0.0 %)	☐	Age (0.0 %)	☐
Overall_Health_Rate (0.0 %)	☐	Compare_Gen_Health_Age_Group (0.0 %)	☐
Health_Condition_Activity_Interference (0.0 %)	☐	Days_Lost_in_a_Year (53.06 %)	☐
Last_Medical_Care (0.0 %)	☐	Last_Medical_Care_Obtained (0.0 %)	☐
Location_Medical_Care (15.11 %)	☐	Smoking_Habit (0.0 %)	☐
No_Fruit_EatingDays (0.0 %)	☐	No_Fruit_Servings (23.57 %)	☐
No_Veg_EatingDays (0.0 %)	☐	No_Veg_Servings (0.0 %)	☐
Syststolic_BP (0.0 %)	☐	Diatolic_BP (0.0 %)	☐
BP_Medications_Last2Weeks (0.0 %)	☐	Pregnant (53.07 %)	☐
Highest_Education (0.0 %)	☐	Weighting (0.0 %)	☐
Smoke (0.0 %)	☐	Alcohol (0.0 %)	☐
Diet_Recommended (0.0 %)	☐	Bmi (0.0 %)	☐
Overweight (0.0 %)	☐	Obesity (0.0 %)	☐
P215 (0.0 %)	☐	Demographic_Site (0.0 %)	☐
Smokestart (74.89 %)	☐	Cigarette_Brand (73.53 %)	☐
No_Ciga_PerDay (86.84 %)	☐	Alcolol_Consumption_YN (0.0 %)	☐
Alcolol_Consumed_In_a_Year_YN (0.0 %)	☐	Alcolol_Consumed_In_a_Week_YN (0.0 %)	☐
Alcolol_Consumed_In_45Days_YN (0.0 %)	☐	Phyact (0.0 %)	☐
Phylow (0.0 %)	☐	Phyhigh (0.0 %)	☐
Acitivity_Vigorous_Intensity (0.0 %)	☐	Waist (0.0 %)	☐
Hyper140 (0.0 %)	☐	Hyper160 (0.0 %)	☐
Risklow1 (0.0 %)	☐	Riskhigh1 (0.0 %)	☐

Success! NaN values are dropped. Please refresh the page and see the changes

The user can view the update after processing of ‘Attribute wise drop NaN’ module. The missing values in those attributes are deleted using ‘Attribute wise drop’ module. After dropping the missing values there are 0% of missing values in these attributes.

Data Imputation- Attribute wise Fill NaN



Select Features for Fill NaN

☐ Check All

Fill ⌄ Mean Values ⌄

Sex (0.0 %)	☐	Age (0.0 %)	☐
Overall_Health_Rate (0.0 %)	☐	Compare_Gen_Health_Age_Group (0.0 %)	☐
Health_Condition_Activity_Interference (0.0 %)	☐	Last_Medical_Care (0.0 %)	☐
Last_Medical_Care_Obtained (0.0 %)	☐	Location_Medical_Care (15.11 %)	☐
Smoking_Habit (0.0 %)	☐	No_Fruit_EatingDays (0.0 %)	☐
No_Fruit_Servings (23.57 %)	☑	No_Veg_EatingDays (0.0 %)	☐
No_Veg_Servings (0.0 %)	☐	Syststolic_BP (0.0 %)	☐
Diatolic_BP (0.0 %)	☐	BP_Medications_Last2Weeks (0.0 %)	☐
Pregnant (0.0 %)	☐	Highest_Education (0.0 %)	☐
Weighting (0.0 %)	☐	Smoke (0.0 %)	☐
Alcohol (0.0 %)	☐	Diet_Recommended (0.0 %)	☐
Bmi (0.0 %)	☐	Overweight (0.0 %)	☐
Obesity (0.0 %)	☐	P215 (0.0 %)	☐
Demographic_Site (0.0 %)	☐	Alcolol_Consumption_YN (0.0 %)	☐
Alcolol_Consumed_In_a_Year_YN (0.0 %)	☐	Alcolol_Consumed_In_a_Week_YN (0.0 %)	☐
Alcolol_Consumed_In_45Days_YN (0.0 %)	☐	Phyact (0.0 %)	☐
Phylow (0.0 %)	☐	Phyhigh (0.0 %)	☐
Acitivity_Vigorous_Intensity (0.0 %)	☐	Waist (0.0 %)	☐
Hyper140 (0.0 %)	☐	Hyper160 (0.0 %)	☐
Risklow1 (0.0 %)	☐	Riskhigh1 (0.0 %)	☐
Risklow2 (0.0 %)	☐	Riskhigh2 (0.0 %)	☐

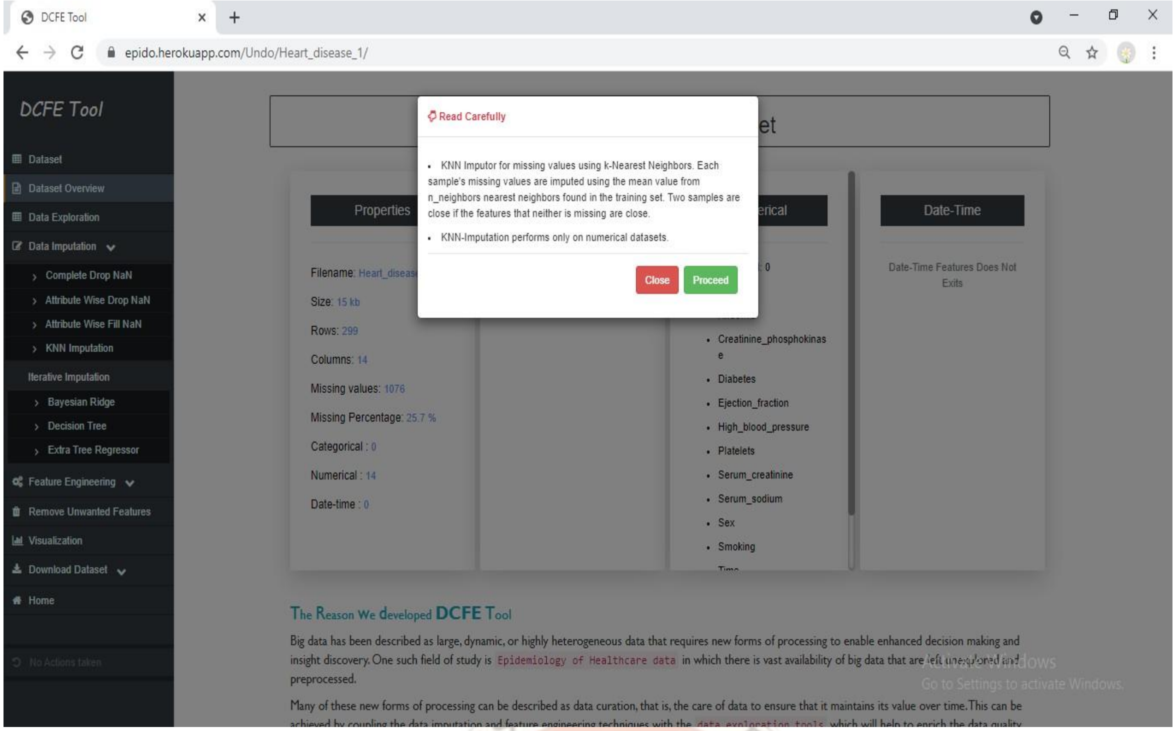
Fill

If the user wants to fill NaN, then they can click the value which is shown in the drop down list as ‘Fill’. Another option shown in the drop down list is ‘Replace’. So either Fill or Replace value can be chosen by the user to perform ‘filling missing values’ operation.

Forward Fill, Backward Fill, Mean and Mode

The user is provided with four different methods (Forward Fill, Backward Fill, Mean and Mode) to fill NaN. The options are provided through drop down list.

Data Imputation- KNN Imputation



Read Carefully

- KNN Imputor for missing values using k-Nearest Neighbors. Each sample's missing values are imputed using the mean value from n_neighbors nearest neighbors found in the training set. Two samples are close if the features that neither is missing are close.
- KNN-Imputation performs only on numerical datasets.

Filename: Heart_diseas

Size: 15 kb

Rows: 299

Columns: 14

Missing values: 1076

Missing Percentage: 25.7 %

Categorical : 0

Numerical : 14

Date-time : 0

- Creatinine_phosphokinase
- Diabetes
- Ejection_fraction
- High_blood_pressure
- Platelets
- Serum_creatinine
- Serum_sodium
- Sex
- Smoking

Date-Time Features Does Not Exits

The Reason We developed DCFE Tool

Big data has been described as large, dynamic, or highly heterogeneous data that requires new forms of processing to enable enhanced decision making and insight discovery. One such field of study is `Epidemiology of Healthcare data` in which there is vast availability of big data that are left unexplored and preprocessed.

Many of these new forms of processing can be described as data curation, that is, the care of data to ensure that it maintains its value over time. This can be achieved by coupling the data imputation and feature engineering techniques with the `data exploration tools` which will help to enrich the data quality

Activate Windows
Go to Settings to activate Windows.

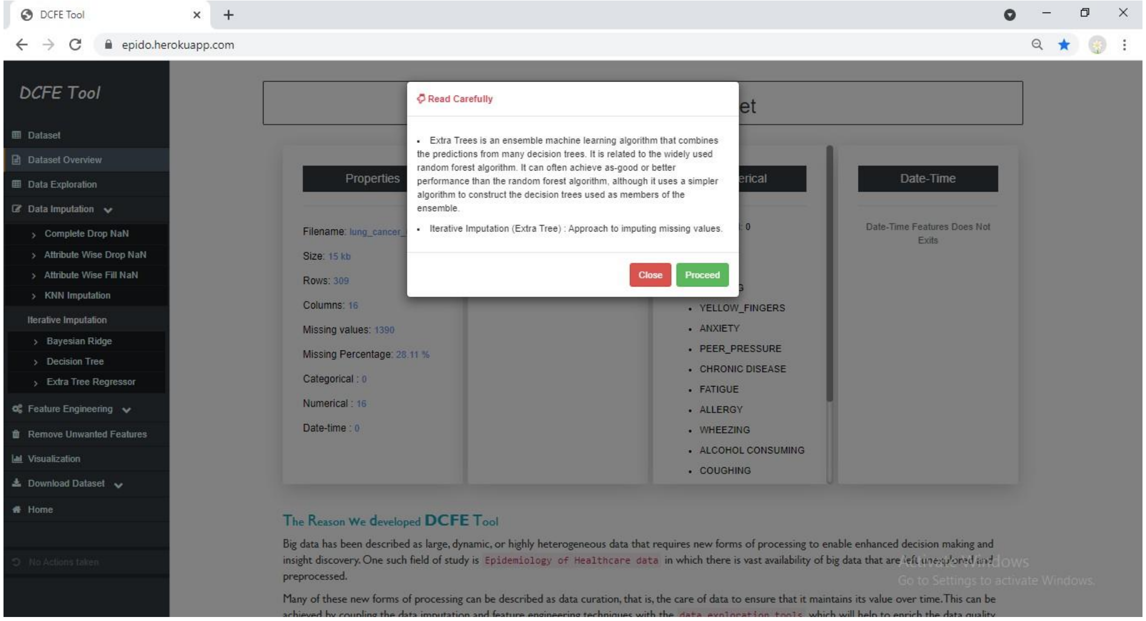
Close

'Close' button is used to close the dialog box after the rules are read by any user.

Proceed

If the user wants to fill missing values using KNN imputation, then click 'Proceed' button. The missing values are imputed using KNN imputation where the missing values are filled using average value of the 'n' nearest neighbor values.

Iterative Imputation- Extra Tree Regressor Imputation



Read Carefully

- Extra Trees is an ensemble machine learning algorithm that combines the predictions from many decision trees. It is related to the widely used random forest algorithm. It can often achieve as-good or better performance than the random forest algorithm, although it uses a simpler algorithm to construct the decision trees used as members of the ensemble.
- Iterative Imputation (Extra Tree) : Approach to imputing missing values.

Filename: lung_cancer_
Size: 15 kb
Rows: 309
Columns: 16
Missing values: 1390
Missing Percentage: 28.11 %
Categorical : 0
Numerical : 16
Date-time : 0

- YELLOW_FINGERS
- ANXIETY
- PEER_PRESSURE
- CHRONIC DISEASE
- FATIGUE
- ALLERGY
- WHEEZING
- ALCOHOL CONSUMING
- COUGHING

Date-Time Features Does Not Exits

The Reason We developed DCFE Tool

Big data has been described as large, dynamic, or highly heterogeneous data that requires new forms of processing to enable enhanced decision making and insight discovery. One such field of study is `Epidemiology of Healthcare data` in which there is vast availability of big data that are left unexplored and preprocessed.

Many of these new forms of processing can be described as data curation, that is, the care of data to ensure that it maintains its value over time. This can be achieved by coupling the data imputation and feature engineering techniques with the `data exploration tools` which will help to enrich the data quality

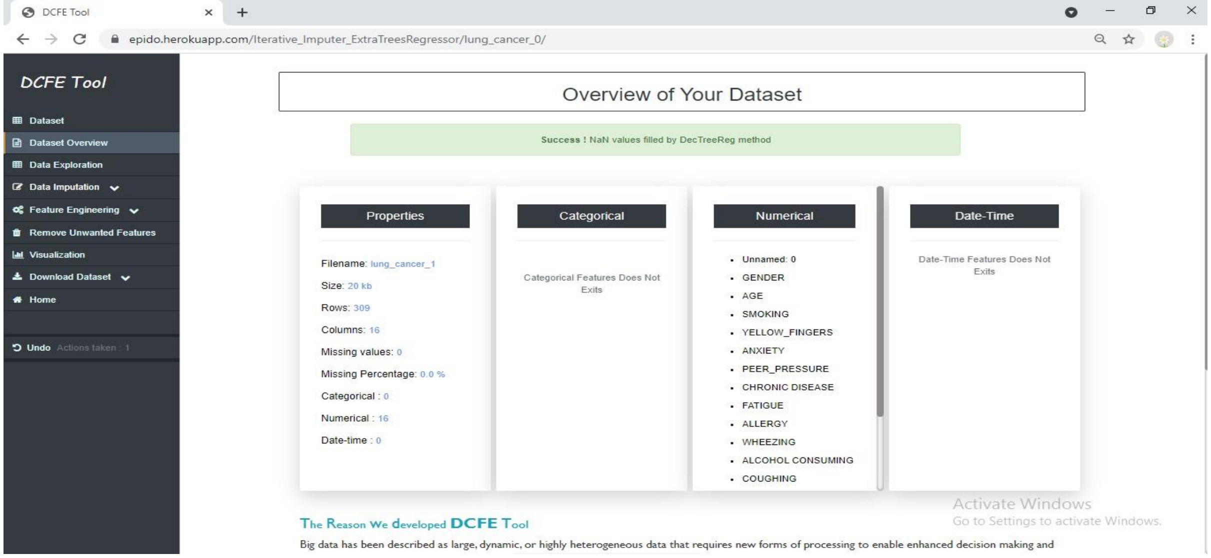
Close

'Close' button is used to close the dialog box after the rules are read by any user.

Proceed

If the user wants to fill missing values using Extra Tree Regressor imputation, then click 'Proceed' button. The missing values are imputed using Extra Tree Regressor.

Iterative Imputation- Decision Tree Regressor Imputation



Overview of Your Dataset

Success ! NaN values filled by DecTreeReg method

Properties

Filename: lung_cancer_1

Size: 20 kb

Rows: 309

Columns: 16

Missing values: 0

Missing Percentage: 0.0 %

Categorical : 0

Numerical : 16

Date-time : 0

Categorical

Categorical Features Does Not Exits

Numerical

- Unnamed: 0
- GENDER
- AGE
- SMOKING
- YELLOW_FINGERS
- ANXIETY
- PEER_PRESSURE
- CHRONIC DISEASE
- FATIGUE
- ALLERGY
- WHEEZING
- ALCOHOL CONSUMING
- COUGHING

Date-Time

Date-Time Features Does Not Exits

The Reason We developed **DCFE** Tool

Big data has been described as large, dynamic, or highly heterogeneous data that requires new forms of processing to enable enhanced decision making and

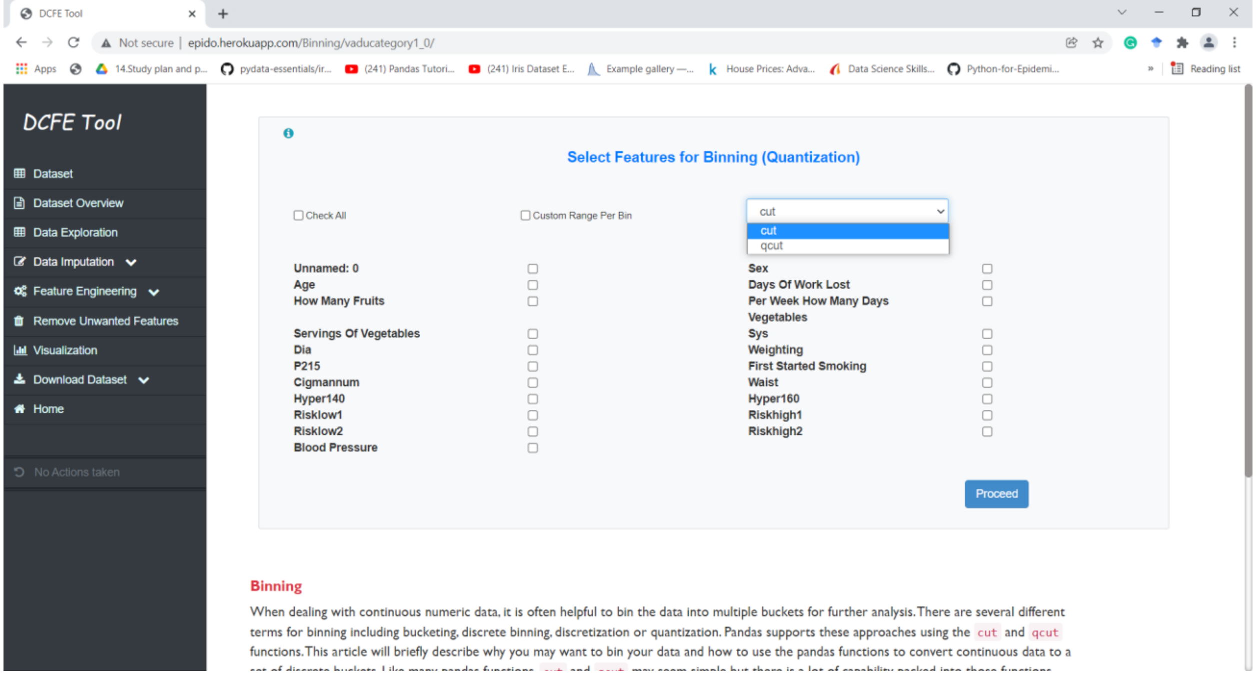
Close

‘Close’ button is used to close the dialog box after the rules are read by any user.

Proceed

If the user wants to fill missing values using Decision Tree Regressor imputation, then click ‘Proceed’ button. The missing values are imputed using Decision Tree Regressor imputation method. The changes can be viewed in the Dataset Overview in which Missing percentage will be shown as ‘Zero’

Feature Engineering- Binning



Select Features for Binning (Quantization)

□ Check All　□ Custom Range Per Bin　cut (cut, qcut)

Unnamed: 0	□	Sex	□
Age	□	Days Of Work Lost	□
How Many Fruits	□	Per Week How Many Days Vegetables	□
Servings Of Vegetables	□	Sys	□
Dia	□	Weighting	□
P215	□	First Started Smoking	□
Cigmannum	□	Waist	□
Hyper140	□	Hyper160	□
Risklow1	□	Riskhigh1	□
Risklow2	□	Riskhigh2	□
Blood Pressure	□		

Binning

When dealing with continuous numeric data, it is often helpful to bin the data into multiple buckets for further analysis. There are several different terms for binning including bucketing, discrete binning, discretization or quantization. Pandas supports these approaches using the `cut` and `qcut` functions. This article will briefly describe why you may want to bin your data and how to use the pandas functions to convert continuous data to a set of discrete buckets. Like many pandas functions, `cut` and `qcut` may seem simple but there is a lot of capability packed into those functions.

Check All

The user can chose any attribute for doing Binning. Binning can be done in numerical data alone. Individual check option is also there for any user.

Custom range Per Bin

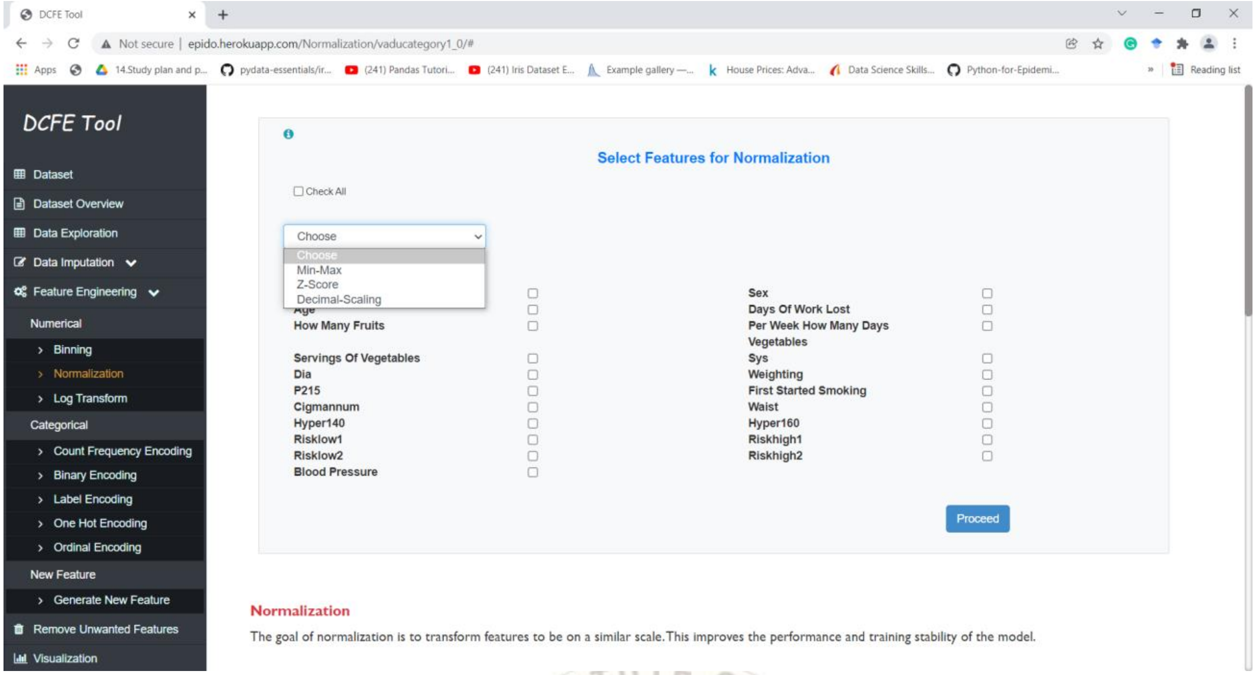
The user can chose/customize the range of each bin for cut (Equal sized Discretization) method. The maximum and minimum values can be known through visualization of Box plot for every attribute.

Next method shown in the drop down list is qcut (Quantile based discretization) method. The user can also chose qcut method for Binning process.

Proceed

If user wants to start any method for processing, then click ‘Proceed’ button for Binning any continuous values. After clicking ‘proceed’ button, it will redirect the user to Dataset view menu to see the changes in the dataset.

Feature Engineering- Normalization



Select Features for Normalization

□ Check All

Choose
- Choose
- Min-Max
- Z-Score
- Decimal-Scaling

	□	Sex	□
Age	□	Days Of Work Lost	□
How Many Fruits	□	Per Week How Many Days Vegetables	□
Servings Of Vegetables	□	Sys	□
Dia	□	Weighting	□
P215	□	First Started Smoking	□
Cigmannum	□	Waist	□
Hyper140	□	Hyper160	□
Risklow1	□	Riskhigh1	□
Risklow2	□	Riskhigh2	□
Blood Pressure	□		

Proceed

Normalization

The goal of normalization is to transform features to be on a similar scale.This improves the performance and training stability of the model.

Check All

The user can chose any attribute for doing Normalization. Normalization can be done in numerical data alone. Individual check option is also there for any user.

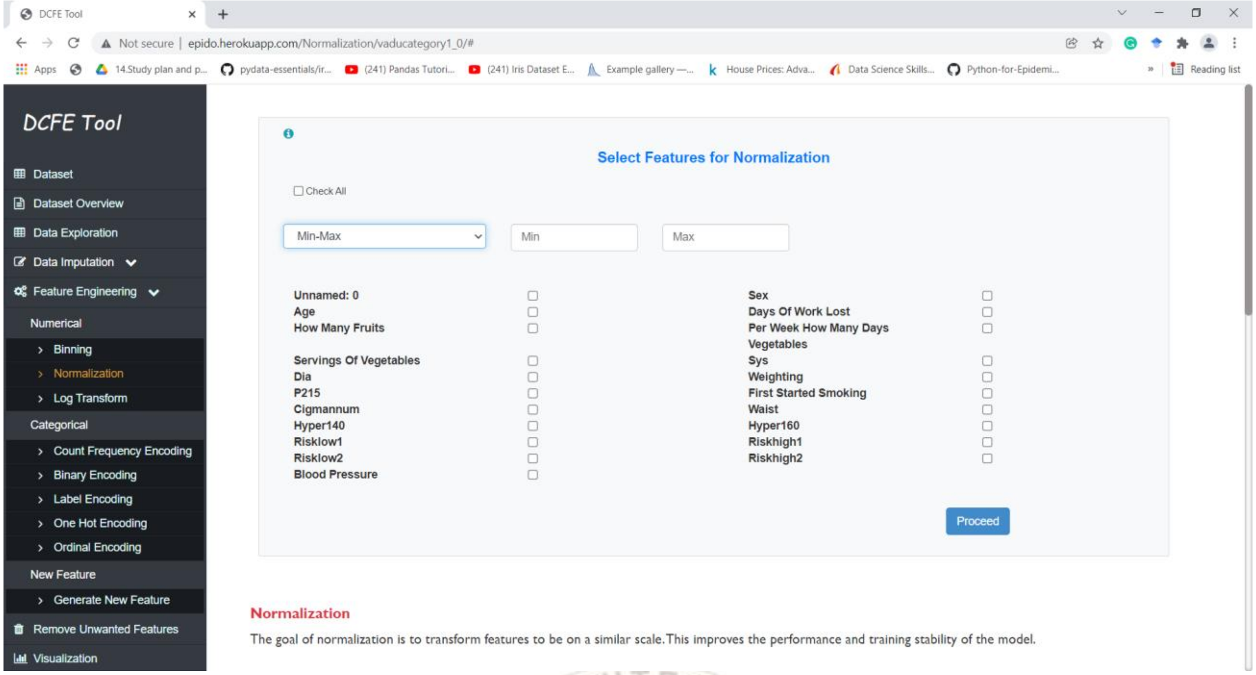
Drop down list

The drop down list has three options for normalization of any numerical data. The user can click any method to perform normalization depending upon their uploaded data.

Proceed

The user will click ‘proceed’ button for doing normalization. The updated values can be viewable from the Dataset menu option. Every change can also be reflected in visualization part.

Feature Engineering- Min – Max normalization



Select Features for Normalization

□ Check All

Min-Max | Min | Max

Unnamed: 0 □
Age □
How Many Fruits □
Servings Of Vegetables □
Dia □
P215 □
Cigmannum □
Hyper140 □
Risklow1 □
Risklow2 □
Blood Pressure □

Sex □
Days Of Work Lost □
Per Week How Many Days Vegetables □
Sys □
Weighting □
First Started Smoking □
Waist □
Hyper160 □
Riskhigh1 □
Riskhigh2 □

Proceed

Normalization

The goal of normalization is to transform features to be on a similar scale.This improves the performance and training stability of the model.

Check All

The user can chose any attribute for doing Min-Max Normalization. Normalization can be done in numerical data alone. Individual check option is also there for any user.

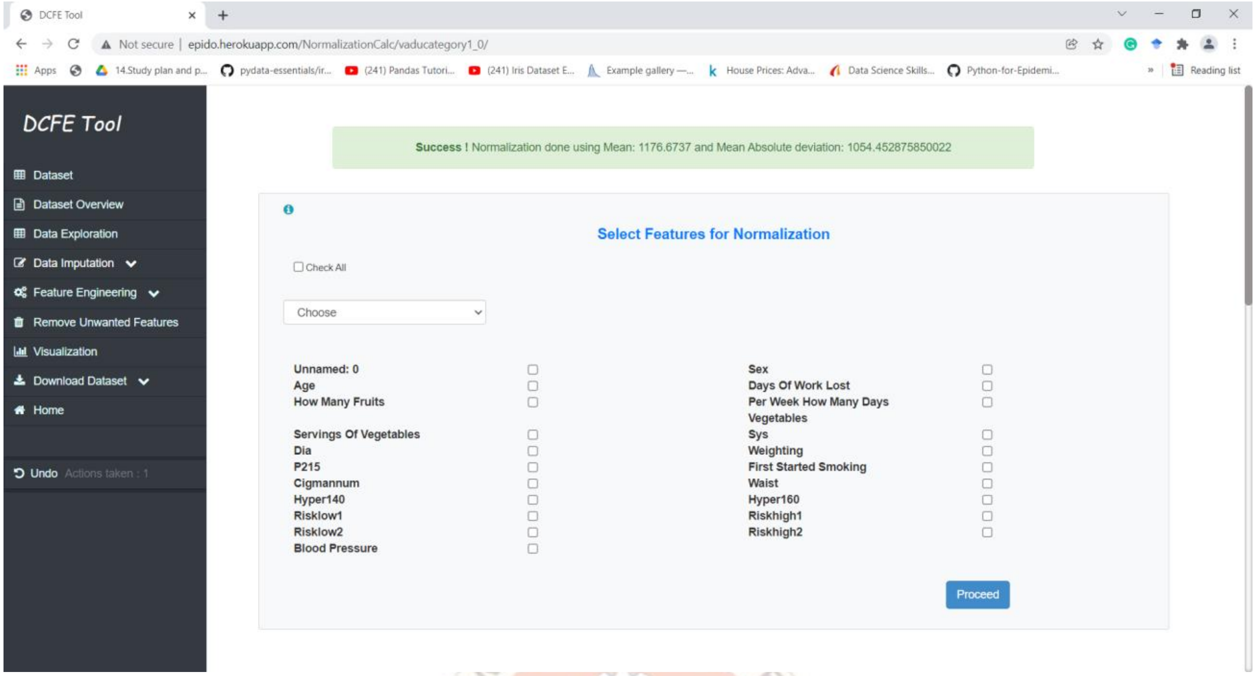
Drop down list

The Min-Max option can be clicked by any user.

Max and Min value

The user can identify the minimum and maximum value using Box –Plot which is shown in visualization part.

Feature Engineering- Z Score normalization



Success ! Normalization done using Mean: 1176.6737 and Mean Absolute deviation: 1054.452875850022

Select Features for Normalization

□ Check All

Choose

Unnamed: 0	□	Sex	□
Age	□	Days Of Work Lost	□
How Many Fruits	□	Per Week How Many Days Vegetables	□
Servings Of Vegetables	□	Sys	□
Dia	□	Weighting	□
P215	□	First Started Smoking	□
Cigmannum	□	Waist	□
Hyper140	□	Hyper160	□
Risklow1	□	Riskhigh1	□
Risklow2	□	Riskhigh2	□
Blood Pressure	□		

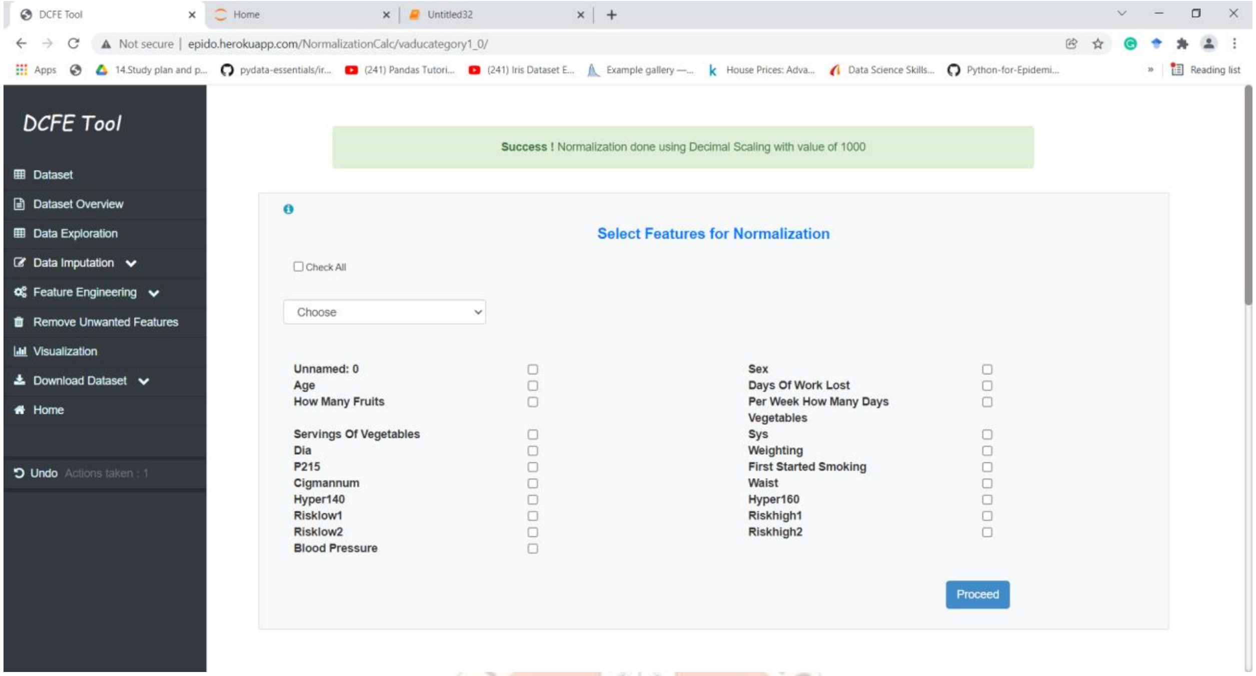
Check All

The user can choose any attribute for doing Z-score Normalization. Z-Score normalization can be done in numerical data alone. Individual check option is also there for any user.

Proceed

The ‘Proceed’ button can provide the user to perform Z-score normalization on the checked attribute. The mean and mean absolute deviation is automatically assigned for chosen attribute.

Feature Engineering-Decimal Scaling



Success ! Normalization done using Decimal Scaling with value of 1000

Select Features for Normalization

□ Check All

Unnamed: 0	□	Sex	□
Age	□	Days Of Work Lost	□
How Many Fruits	□	Per Week How Many Days Vegetables	□
Servings Of Vegetables	□	Sys	□
Dia	□	Weighting	□
P215	□	First Started Smoking	□
Cigmannum	□	Waist	□
Hyper140	□	Hyper160	□
Risklow1	□	Riskhigh1	□
Risklow2	□	Riskhigh2	□
Blood Pressure	□		

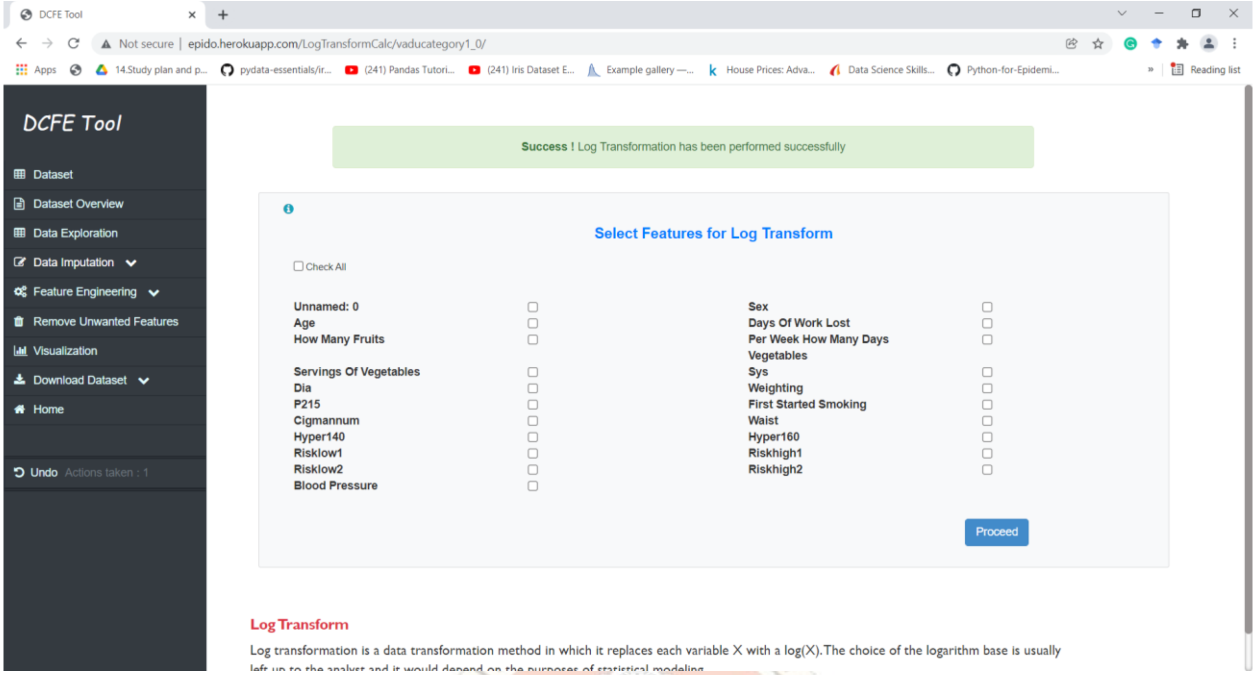
Check All

The user can chose any attribute for doing Decimal Scaling Normalization. Individual check option is also there for any user.

Proceed

The ‘Proceed’ button can provide the user to perform Decimal Scaling normalization on the checked attribute. The value for Decimal Scaling is automatically generated.

Feature Engineering-Log Transform



Success ! Log Transformation has been performed successfully

Select Features for Log Transform

□ Check All

Unnamed: 0 □
Age □
How Many Fruits □
Servings Of Vegetables □
Dia □
P215 □
Cigmannum □
Hyper140 □
Risklow1 □
Risklow2 □
Blood Pressure □

Sex □
Days Of Work Lost □
Per Week How Many Days Vegetables □
Sys □
Weighting □
First Started Smoking □
Waist □
Hyper160 □
Riskhigh1 □
Riskhigh2 □

Proceed

Log Transform

Log transformation is a data transformation method in which it replaces each variable X with a log(X).The choice of the logarithm base is usually left up to the analyst and it would depend on the purposes of statistical modeling

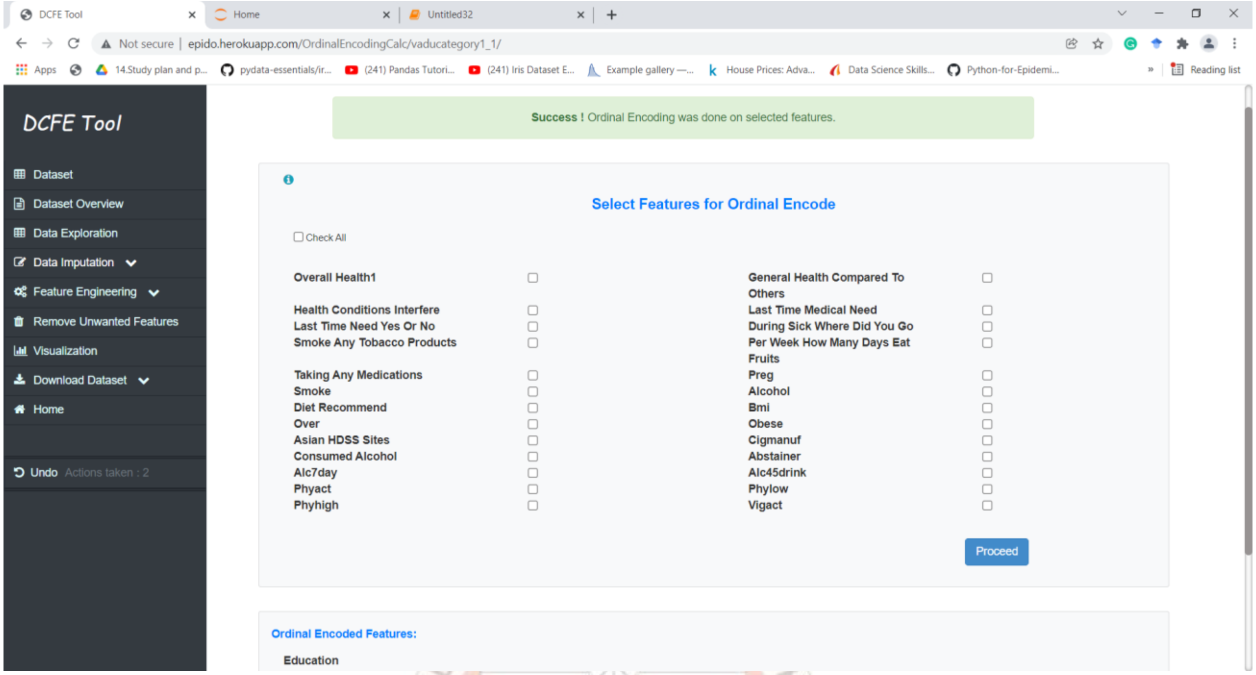
Check All

The user can select any attribute for doing Log Transform. Individual check option is also there for any user.

Proceed

The 'Proceed' button can provide the user to perform Log Transform on the selected attribute. The value is automatically generated which replaces each variable X with a log(X).

Feature Engineering-Ordinal Encoding



Success ! Ordinal Encoding was done on selected features.

Select Features for Ordinal Encode

□ Check All

Overall Health1 □
Health Conditions Interfere □
Last Time Need Yes Or No □
Smoke Any Tobacco Products □
Taking Any Medications □
Smoke □
Diet Recommend □
Over □
Asian HDSS Sites □
Consumed Alcohol □
Alc7day □
Phyact □
Phyhigh □

General Health Compared To Others □
Last Time Medical Need □
During Sick Where Did You Go □
Per Week How Many Days Eat Fruits □
Preg □
Alcohol □
Bmi □
Obese □
Cigmanuf □
Abstainer □
Alc45drink □
Phylow □
Vigact □

Ordinal Encoded Features:

Education

Check All

The user can choose any attribute for performing Ordinal Encoding. Individual check option is also there for any user. The categorical features alone displayed in the display list of features.

Proceed

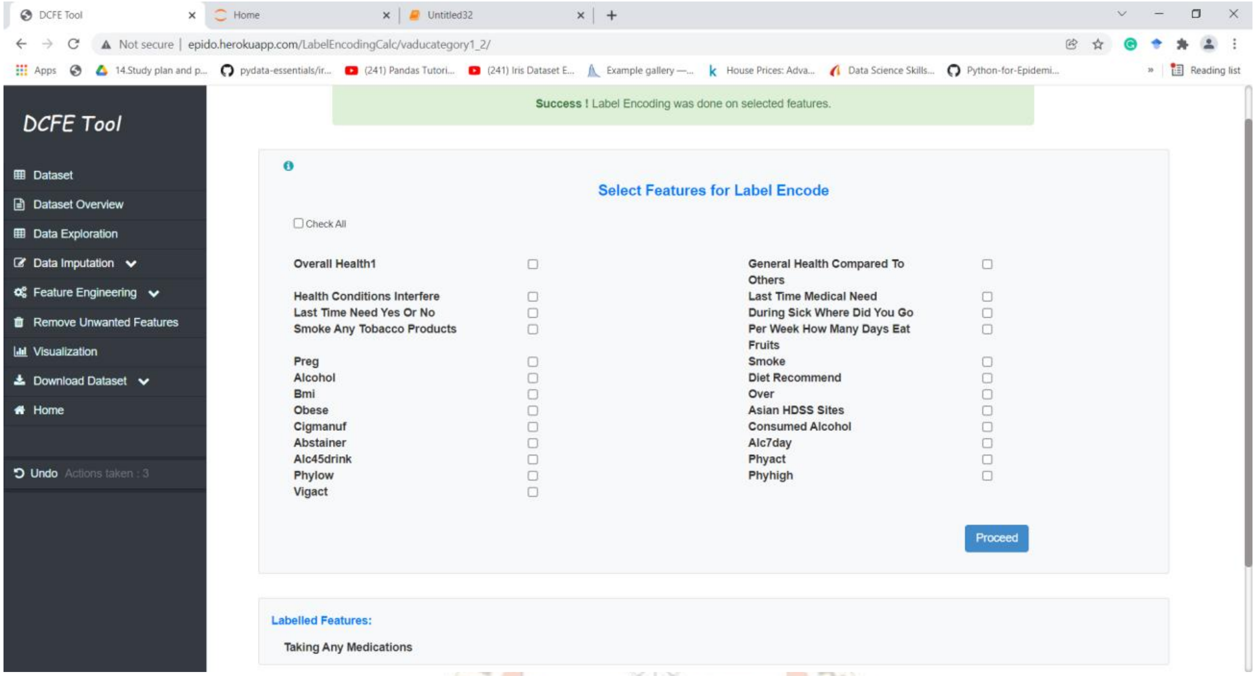
The **‘Proceed’** button can provide the user to perform ordinal encoding on the checked attribute.

Message: Success! Ordinal Encoding was done on selected features. The user can see the mentioned message which is displayed after finishing the proceed button. The user can select the ‘Dataset’ menu option to know the changes made in the dataset.

Ordinal Encoded Features

The user can view the newly generated feature with its new attribute name.

Feature Engineering- Label Encoding



Success ! Label Encoding was done on selected features.

Select Features for Label Encode

□ Check All

Overall Health1 □
Health Conditions Interfere □
Last Time Need Yes Or No □
Smoke Any Tobacco Products □
Preg □
Alcohol □
Bmi □
Obese □
Cigmanuf □
Abstainer □
Alc45drink □
Phylow □
Vigact □

General Health Compared To Others □
Last Time Medical Need □
During Sick Where Did You Go □
Per Week How Many Days Eat Fruits □
Smoke □
Diet Recommend □
Over □
Asian HDSS Sites □
Consumed Alcohol □
Alc7day □
Phyact □
Phyhigh □

Labelled Features:

Taking Any Medications

Check All

The user can choose any attribute for performing Label Encoding. Individual check option is also there for any user. The categorical features alone displayed in the display list of features.

Proceed

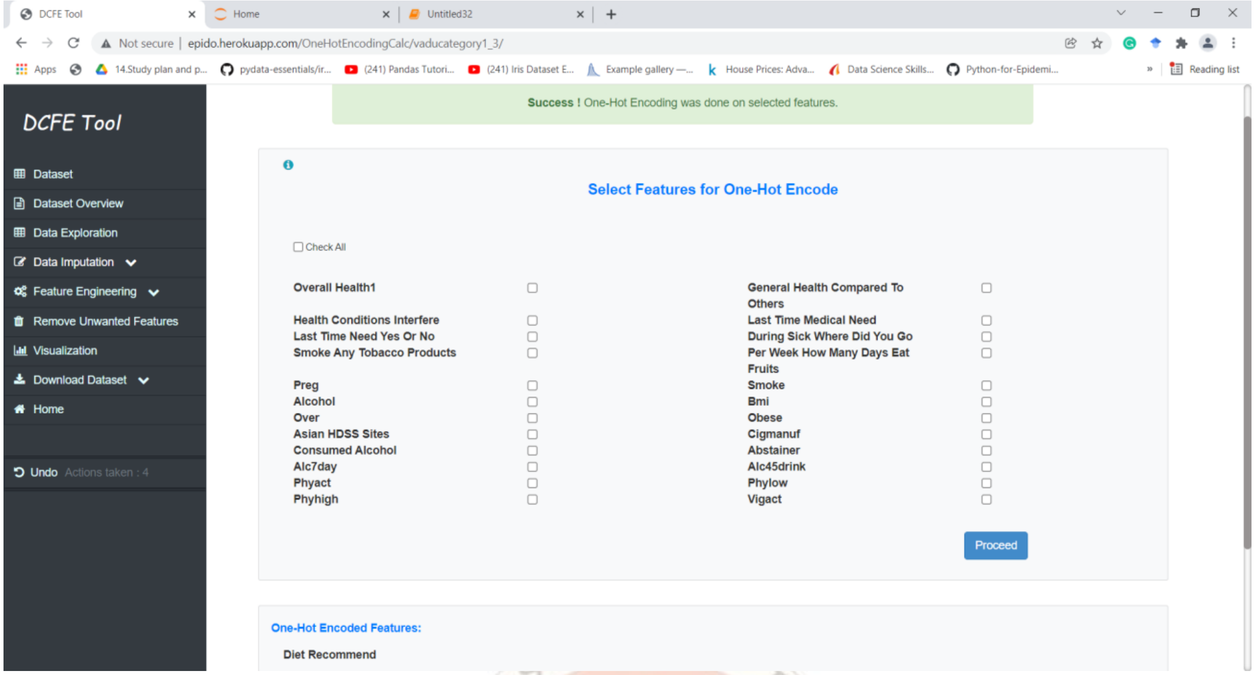
The **‘Proceed’** button can provide the user to perform Label encoding on the checked attribute.

Message: Success! Label Encoding was done on selected features. The user can see the mentioned message which is displayed after finishing the proceed button. The user can select the ‘Dataset’ menu option to know the changes made in the dataset.

Labelled Features

The user can view the newly generated feature with its new attribute name.

Feature Engineering-One-Hot Encoding



Success ! One-Hot Encoding was done on selected features.

Select Features for One-Hot Encode

□ Check All

Overall Health1	□	General Health Compared To Others	□
Health Conditions Interfere	□	Last Time Medical Need	□
Last Time Need Yes Or No	□	During Sick Where Did You Go	□
Smoke Any Tobacco Products	□	Per Week How Many Days Eat Fruits	□
Preg	□	Smoke	□
Alcohol	□	Bmi	□
Over	□	Obese	□
Asian HDSS Sites	□	Cigmanuf	□
Consumed Alcohol	□	Abstainer	□
Alc7day	□	Alc45drink	□
Phyact	□	Phylow	□
Phyhigh	□	Vigact	□

One-Hot Encoded Features:

Diet Recommend

Check All

The user can choose any attribute for performing One-Hot Encoding. Individual check option is also there for any user. The categorical features alone displayed in the display list of features.

Proceed

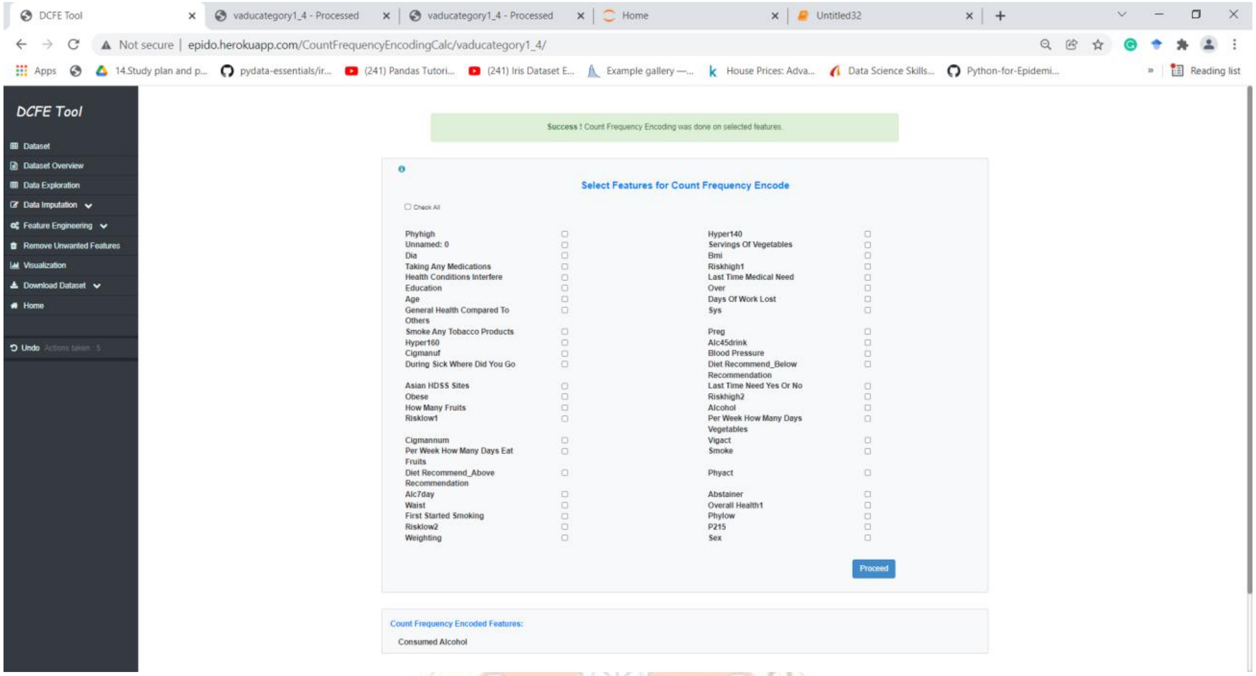
The **‘Proceed’** button can provide the user to perform One-Hot encoding on the checked attribute.

Message: Success! One-Hot Encoding was done on selected features. The user can see the mentioned message which is displayed after finishing the proceed button. The user can select the ‘Dataset’ menu option to know the changes made in the dataset.

One-Hot Encoded Features

The user can view the newly generated feature with its new attribute name.

Feature Engineering- Count Frequency



Success ! Count Frequency Encoding was done on selected features.

Select Features for Count Frequency Encode

□ Check All

Phyhigh	□	Hyper140	□
Unnamed: 0	□	Servings Of Vegetables	□
Dia	□	Bmi	□
Taking Any Medications	□	Riskhigh1	□
Health Conditions Interfere	□	Last Time Medical Need	□
Education	□	Over	□
Age	□	Days Of Work Lost	□
General Health Compared To Others	□	Sys	□
Smoke Any Tobacco Products	□	Preg	□
Hyper160	□	Alc45drink	□
Cigmanuf	□	Blood Pressure	□
During Sick Where Did You Go	□	Diet Recommend_Below Recommendation	□
Asian HDSS Sites	□	Last Time Need Yes Or No	□
Obese	□	Riskhigh2	□
How Many Fruits	□	Alcohol	□
Risklow1	□	Per Week How Many Days Vegetables	□
Cigmannum	□	Vigact	□
Per Week How Many Days Eat Fruits	□	Smoke	□
Diet Recommend_Above Recommendation	□	Phyact	□
Alc7day	□	Abstainer	□
Waist	□	Overall Health1	□
First Started Smoking	□	Phylow	□
Risklow2	□	P215	□
Weighting	□	Sex	□

Proceed

Count Frequency Encoded Features:

Consumed Alcohol

Check All

The user can choose any attribute for performing Count Frequency Encoding. Individual check option is also there for any user. The categorical features alone displayed in the display list of features.

Proceed

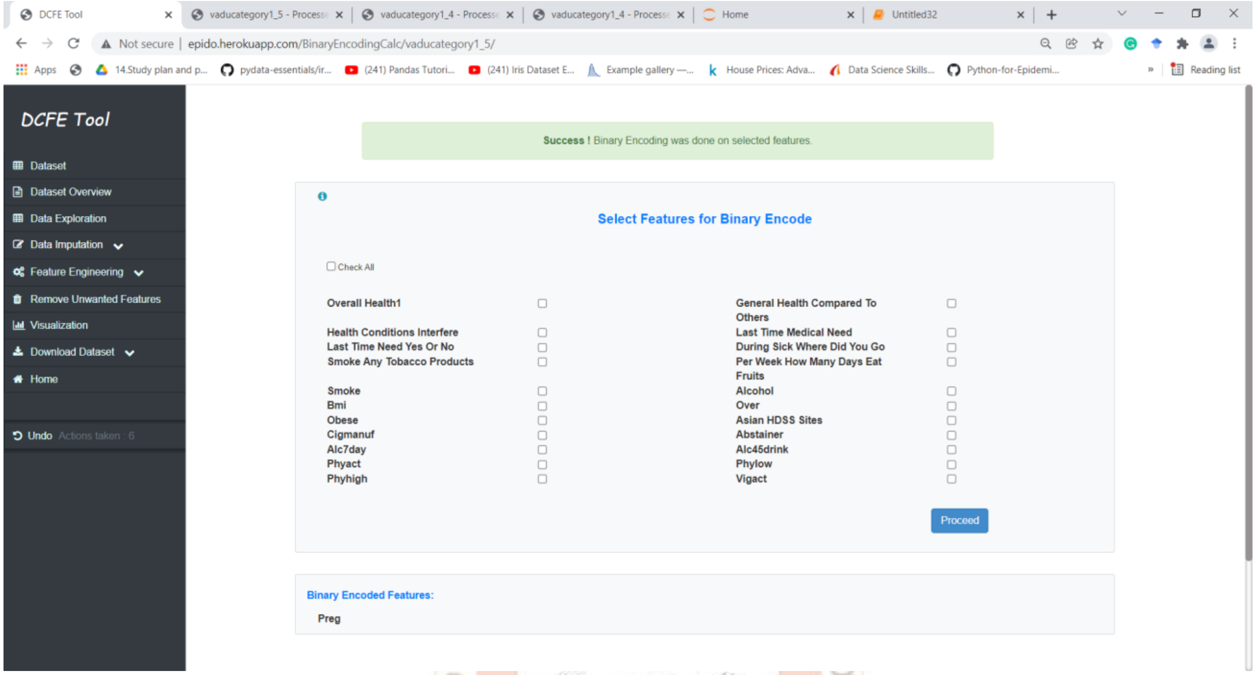
The **'Proceed'** button can provide the user to perform count frequency encoding on the checked attribute.

Message: Success! Count frequency Encoding was done on selected features. The user can see the mentioned message which is displayed after finishing the proceed button. The user can select the 'Dataset' menu option to know the changes made in the dataset.

Count-frequency Encoded Features

The user can view the newly generated feature with its new attribute name.

Feature Engineering-Binary Encoding



Success ! Binary Encoding was done on selected features.

Select Features for Binary Encode

□ Check All

Overall Health1	□	General Health Compared To Others	□
Health Conditions Interfere	□	Last Time Medical Need	□
Last Time Need Yes Or No	□	During Sick Where Did You Go	□
Smoke Any Tobacco Products	□	Per Week How Many Days Eat Fruits	□
Smoke	□	Alcohol	□
Bmi	□	Over	□
Obese	□	Asian HDSS Sites	□
Cigmanuf	□	Abstainer	□
Alc7day	□	Alc45drink	□
Phyact	□	Phylow	□
Phyhigh	□	Vigact	□

Proceed

Binary Encoded Features:

Preg

Check All

The user can choose any attribute for performing Binary Encoding. Individual check option is also there for any user. The categorical features alone displayed in the display list of features.

Proceed

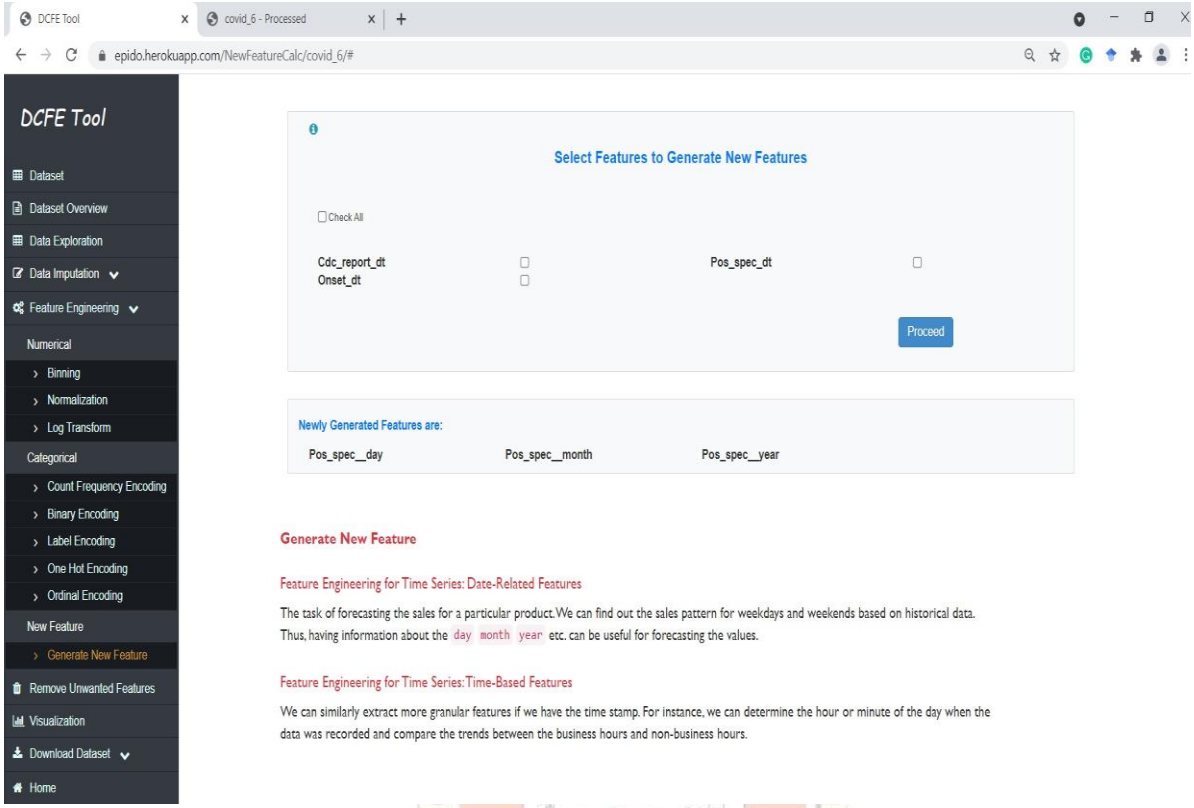
The **'Proceed'** button can provide the user to perform binary encoding on the checked attribute.

Message: Success! Binary Encoding was done on selected features. The user can see the mentioned message which is displayed after finishing the proceed button. The user can select the 'Dataset' menu option to know the changes made in the dataset.

Binary Encoded Features

The user can view the newly generated feature with its new attribute name.

Feature Engineering-New feature- Generate new feature



Select Features to Generate New Features

□ Check All

Cdc_report_dt	□	Pos_spec_dt	□
Onset_dt	□		

Proceed

Newly Generated Features are:

Pos_spec__day Pos_spec__month Pos_spec__year

Generate New Feature

Feature Engineering for Time Series: Date-Related Features

The task of forecasting the sales for a particular product. We can find out the sales pattern for weekdays and weekends based on historical data. Thus, having information about the `day` `month` `year` etc. can be useful for forecasting the values.

Feature Engineering for Time Series: Time-Based Features

We can similarly extract more granular features if we have the time stamp. For instance, we can determine the hour or minute of the day when the data was recorded and compare the trends between the business hours and non-business hours.

Check All

The user can choose any attribute to perform new feature generation only on Date-Time features. Individual check option is also there for any user. The Date-Time features alone displayed in the display list of features.

Proceed

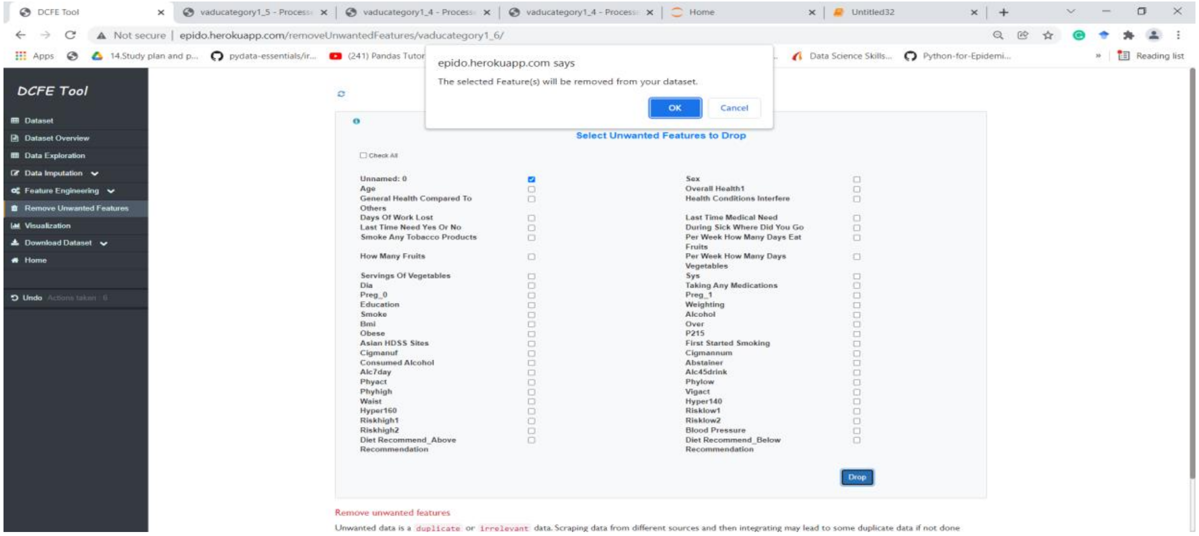
The **'Proceed'** button can provide the user to perform new feature generation on the checked attribute.

Message: Success! New features are generated on selected features. The user can see the mentioned message which is displayed after finishing the proceed button. The user can select the 'Dataset' menu option to know the changes made in the dataset.

Newly generated Features are:

The user can view the newly generated feature with its new attribute name.

Remove unwanted feature



epido.herokuapp.com/removeUnwantedFeatures/vaducategory1_6/

epido.herokuapp.com says

The selected Feature(s) will be removed from your dataset.

OK Cancel

Select Unwanted Features to Drop

□ Check All

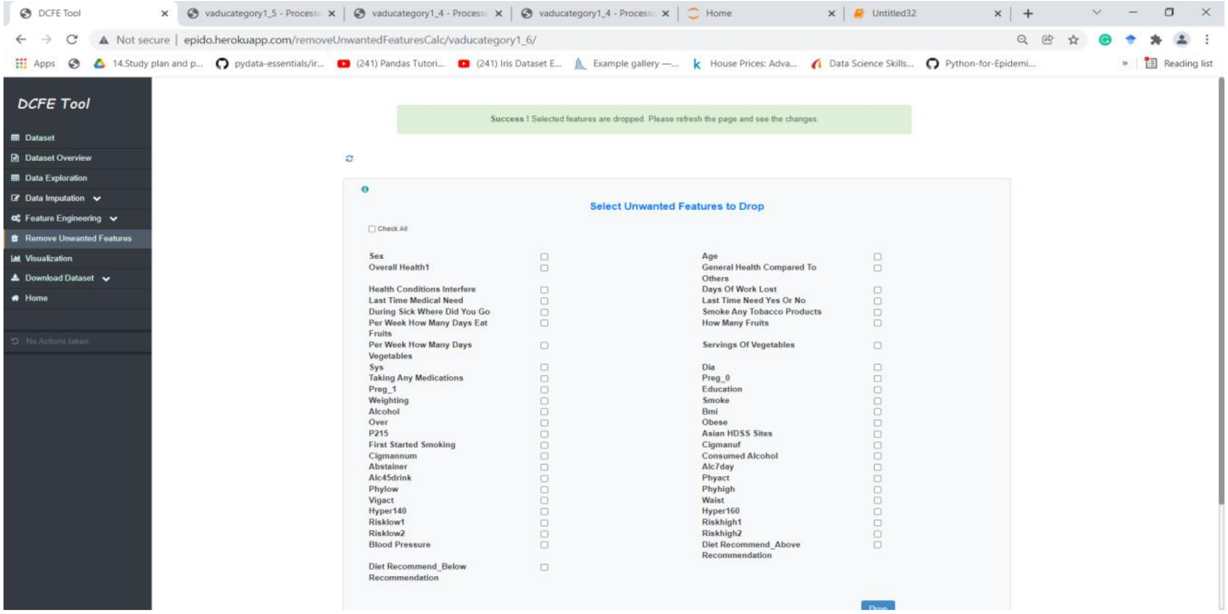
Unnamed: 0 ☑
Age □
General Health Compared To Others □
Days Of Work Lost □
Last Time Need Yes Or No □
Smoke Any Tobacco Products □
How Many Fruits □
Servings Of Vegetables □
Dia □
Preg_0 □
Education □
Smoke □
Bmi □
Obese □
Asian HDSS Sites □
Cigmanuf □
Consumed Alcohol □
Alc7day □
Phyact □
Phyhigh □
Waist □
Hyper160 □
Riskhigh1 □
Riskhigh2 □
Diet Recommend_Above Recommendation □

Sex □
Overall Health1 □
Health Conditions Interfere □
Last Time Medical Need □
During Sick Where Did You Go □
Per Week How Many Days Eat Fruits □
Per Week How Many Days Vegetables □
Sys □
Taking Any Medications □
Preg_1 □
Weighting □
Alcohol □
Over □
P215 □
First Started Smoking □
Cigmannum □
Abstainer □
Alc45drink □
Phylow □
Vigact □
Hyper140 □
Risklow1 □
Risklow2 □
Blood Pressure □
Diet Recommend_Below Recommendation □

Drop

Remove unwanted features

Unwanted data is a `duplicate` or `irrelevant` data. Scraping data from different sources and then integrating may lead to some duplicate data if not done



epido.herokuapp.com/removeUnwantedFeaturesCalc/vaducategory1_6/

Success ! Selected features are dropped. Please refresh the page and see the changes

Select Unwanted Features to Drop

□ Check All

Sex □
Overall Health1 □
Health Conditions Interfere □
Last Time Medical Need □
During Sick Where Did You Go □
Per Week How Many Days Eat Fruits □
Per Week How Many Days Vegetables □
Sys □
Taking Any Medications □
Preg_1 □
Weighting □
Alcohol □
Over □
P215 □
First Started Smoking □
Cigmannum □
Abstainer □
Alc45drink □
Phylow □
Vigact □
Hyper140 □
Risklow1 □
Risklow2 □
Blood Pressure □
Diet Recommend_Below Recommendation □

Age □
General Health Compared To Others □
Days Of Work Lost □
Last Time Need Yes Or No □
Smoke Any Tobacco Products □
How Many Fruits □
Servings Of Vegetables □
Dia □
Preg_0 □
Education □
Smoke □
Bmi □
Obese □
Asian HDSS Sites □
Cigmanuf □
Consumed Alcohol □
Alc7day □
Phyact □
Phyhigh □
Waist □
Hyper160 □
Riskhigh1 □
Riskhigh2 □
Diet Recommend_Above Recommendation □

Drop

Check All

The user can choose any attribute to perform ‘remove unwanted feature’ among all features. Individual check option is also there for any user. All features are displayed in the display list of features

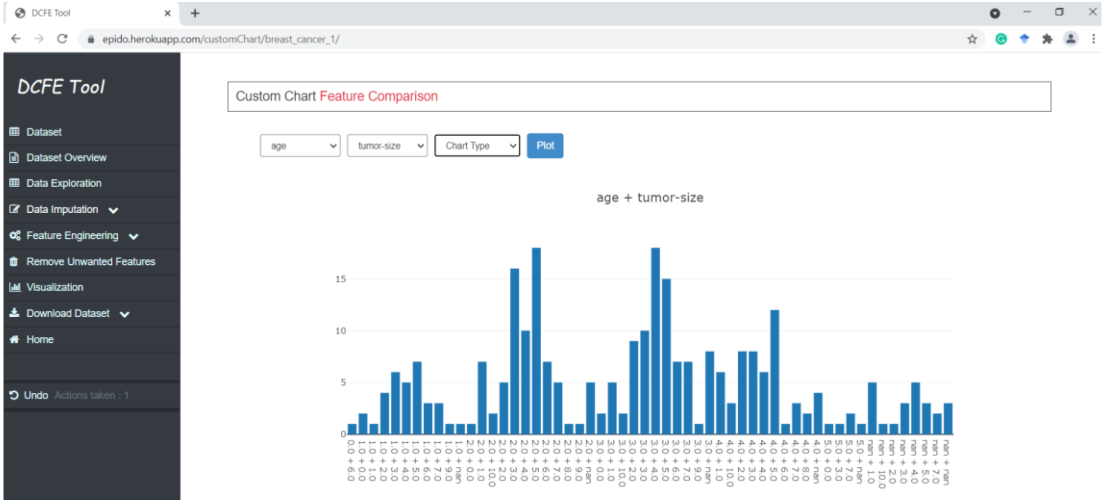
Drop

‘Drop’ button is provided to the user, to proceed with the operation of removing unwanted features.

Message: Success! Selected features are dropped. Please refresh the page and see the changes.

This message is shown to the user to know the status of the function.

Visualization-Feature Comparison



Drop-Down list Feature Comparison

The drop-down list is provided for the user to choose X axis and Y axis features.

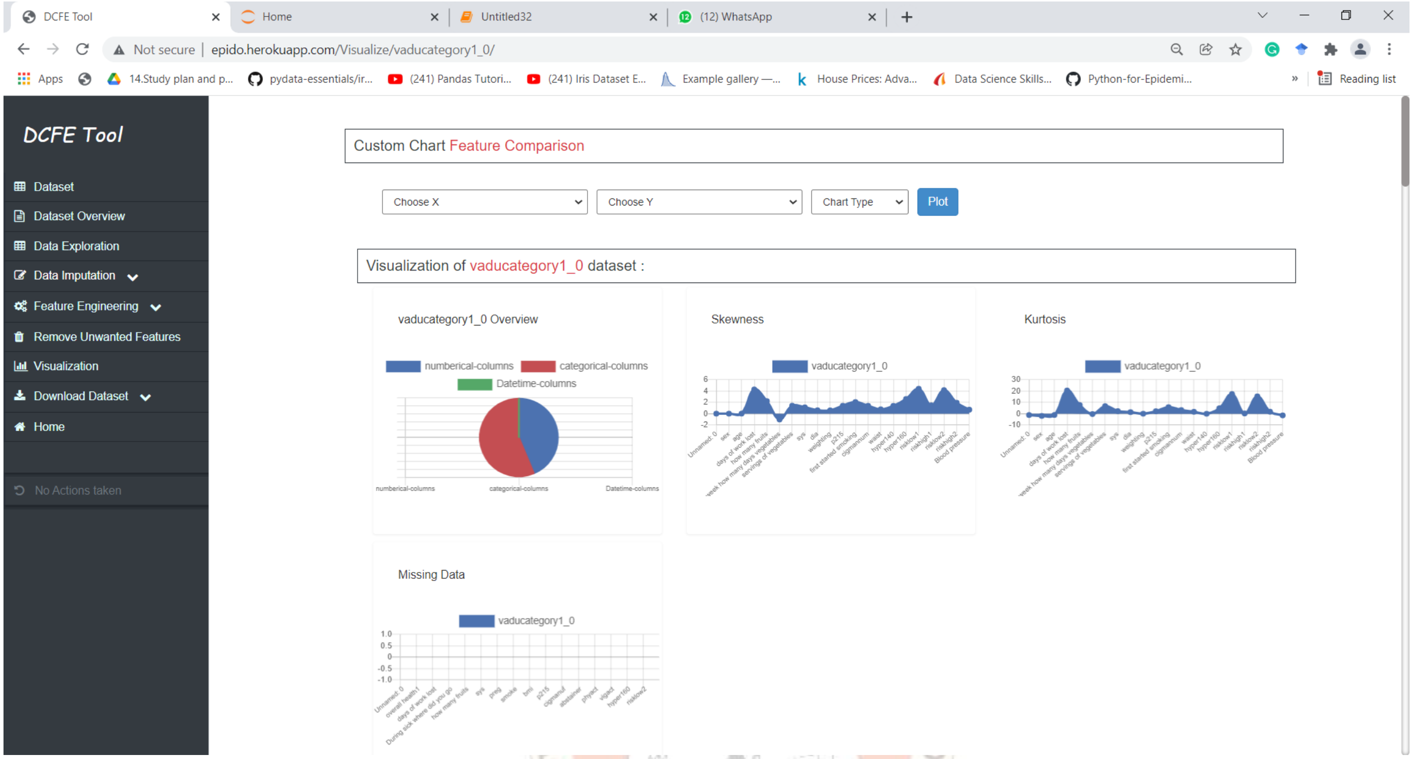
Chart Type

The chart type drop down list for the user is provided for selecting the various charts to plot such as Pie chart, Bar chart, Scatter chart and line chart.

Plot

‘**Plot**’ Button is assigned for triggering the function to perform with the selected chart type.

Visualization- Data Exploration



Dataset overview

If the user wants to visualize the full dataset i.e how many numerical and categorical features present in the dataset,that can be viewed here as scatter plot.

Skewness

The users are provided with the provision to view the skewness of each variable in the dataset.

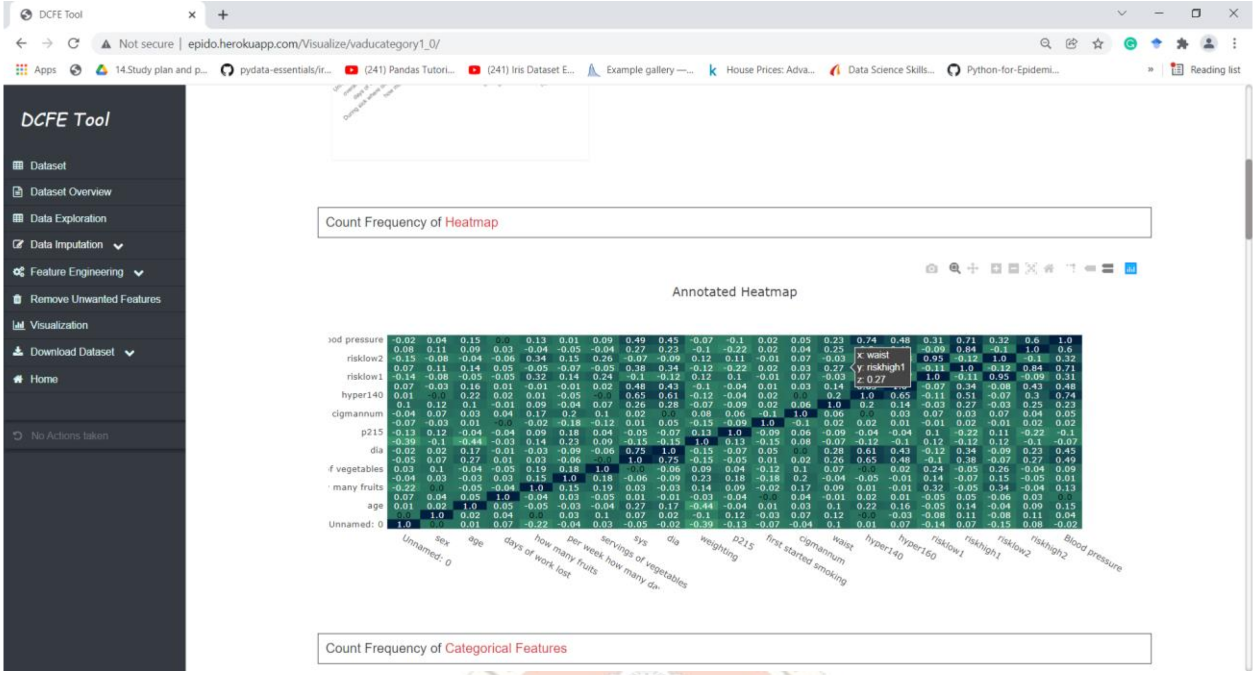
Kurtosis

The users are provided with the provision to view the kurtosis of each variable in the dataset.

Missing Data

The users can view how many values are missed in each feature.

Visualization-Annotated Heat Map

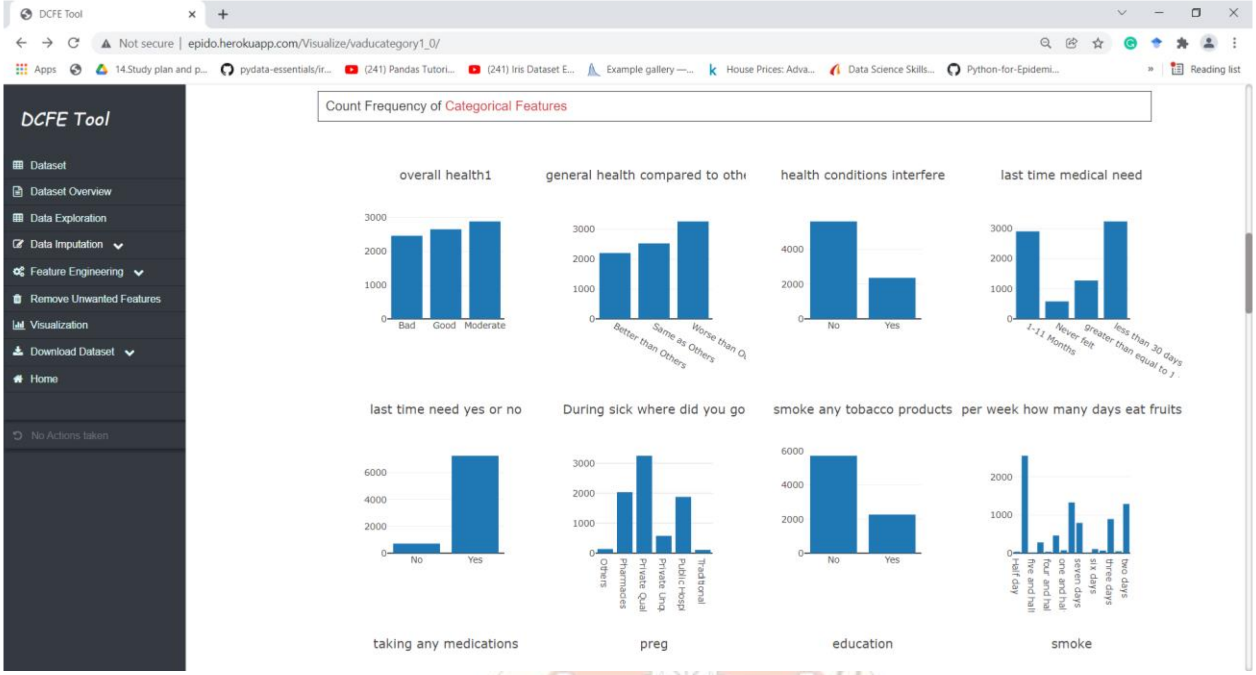


If the user wants to know the correlation between every attribute, then the above heatmap is used to know the correlation between attributes.

Hovering of the cursor will display the exact percentage values.

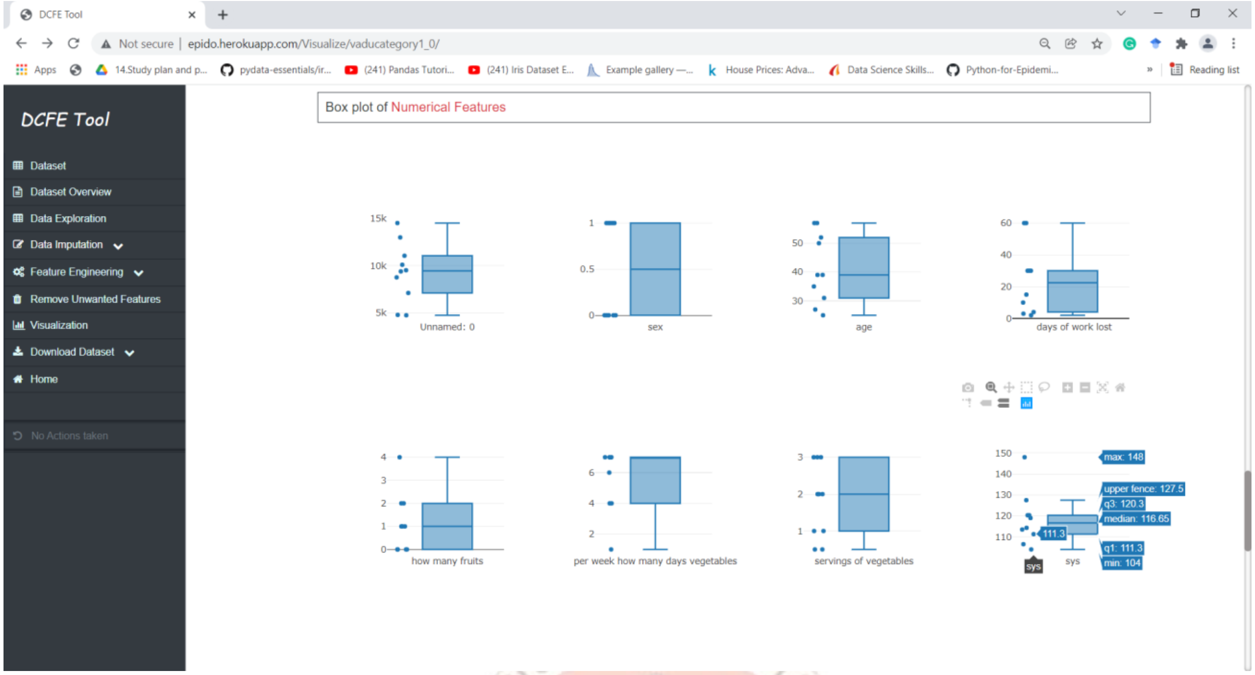
The user can view X axis and Y axis selected features and Z axis will show the correlated value.

Visualization- Count Frequency of categorical features



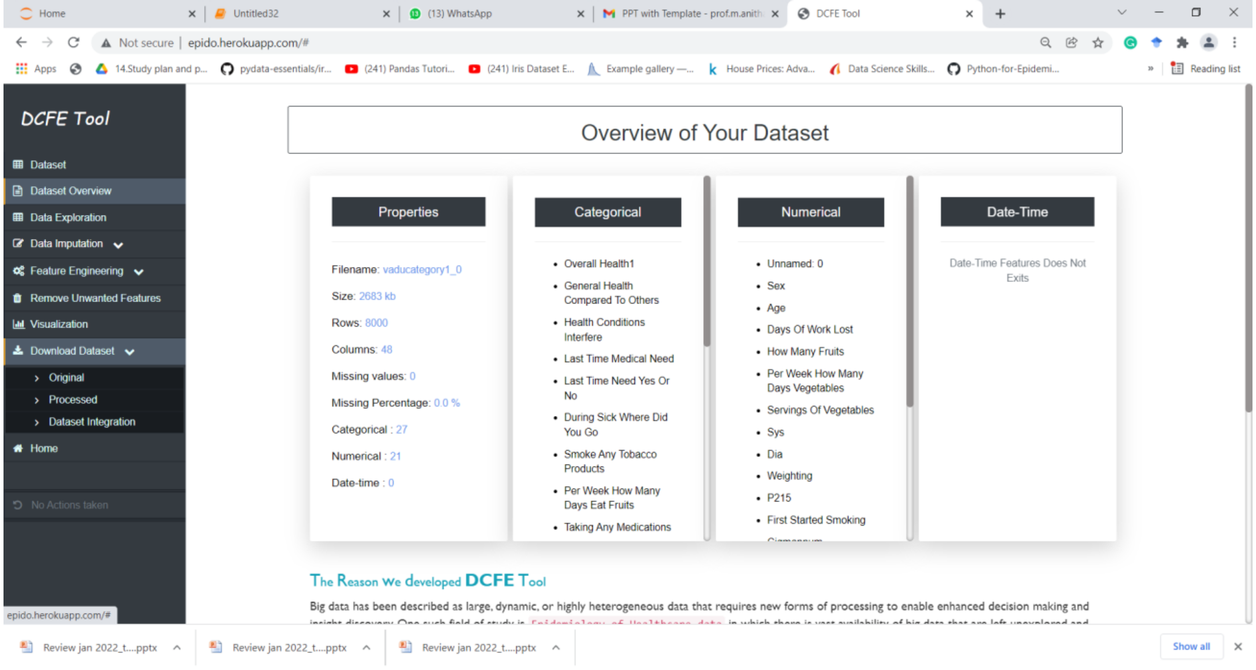
The users can view the counts of each category of every attribute using Bar chart.

Visualization-Box-Plot



The user can view the Box-Plot for each attribute. The minimum and maximum values of each attribute can be shown clearly. The left and right skewed values are also viewable to the user.

Download Dataset



Overview of Your Dataset

Properties

Filename: vaducategory1_0

Size: 2683 kb

Rows: 8000

Columns: 48

Missing values: 0

Missing Percentage: 0.0 %

Categorical : 27

Numerical : 21

Date-time : 0

Categorical

- Overall Health1
- General Health Compared To Others
- Health Conditions Interfere
- Last Time Medical Need
- Last Time Need Yes Or No
- During Sick Where Did You Go
- Smoke Any Tobacco Products
- Per Week How Many Days Eat Fruits
- Taking Any Medications

Numerical

- Unnamed: 0
- Sex
- Age
- Days Of Work Lost
- How Many Fruits
- Per Week How Many Days Vegetables
- Servings Of Vegetables
- Sys
- Dia
- Weighting
- P215
- First Started Smoking

Date-Time

Date-Time Features Does Not Exits

The Reason We developed DCFE Tool

Big data has been described as large, dynamic, or highly heterogeneous data that requires new forms of processing to enable enhanced decision making and

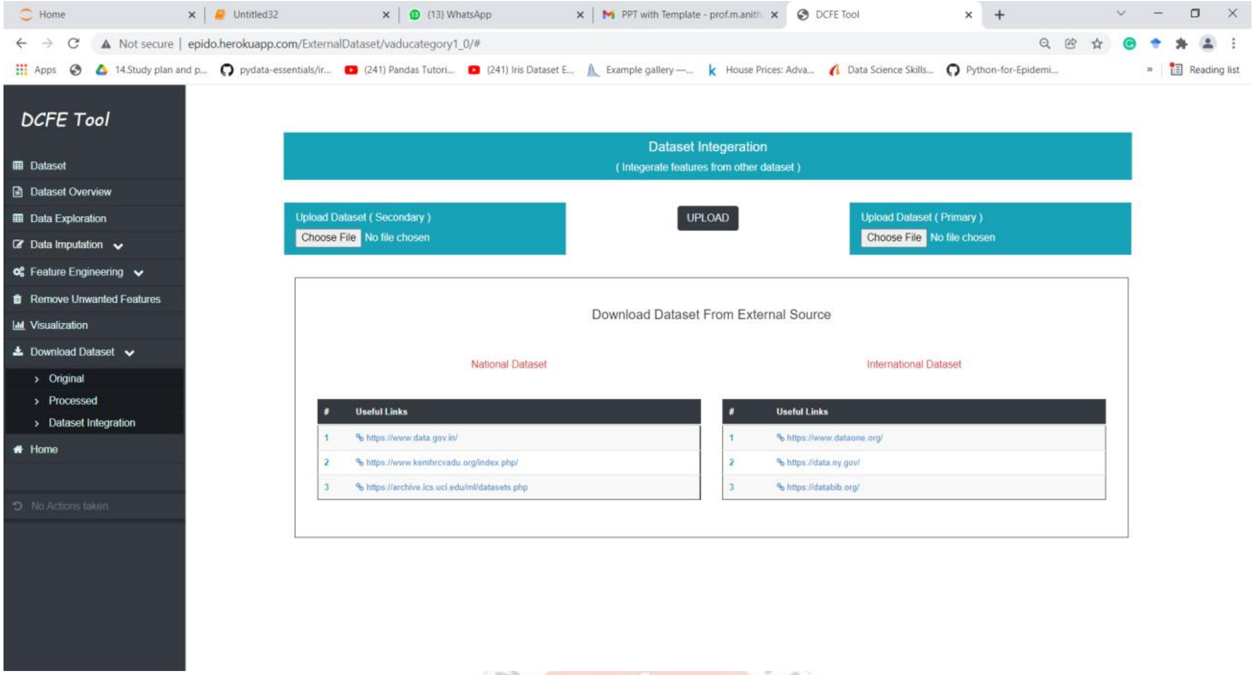
Original Dataset

If the user wants to download the original dataset for any comparison of values with the processed file, then they are provided with Download option of original dataset. The Downloaded dataset can be stored in the local file location by the user.

Processed Dataset

If the user has finished every functions mentioned in the Tool and if it is in curated form, then it is ready to download.
The user is allowed to click the processed link and the file is downloaded and stored in the local file location itself for further usage.

Download Dataset-Dataset Integration



Download Dataset From External Source

National Dataset

#	Useful Links
1	https://www.data.gov.in/
2	https://www.kemhrcvadu.org/index.php/
3	https://archive.ics.uci.edu/ml/datasets.php

International Dataset

#	Useful Links
1	https://www.dataone.org/
2	https://data.ny.gov/
3	https://databib.org/

Choose file

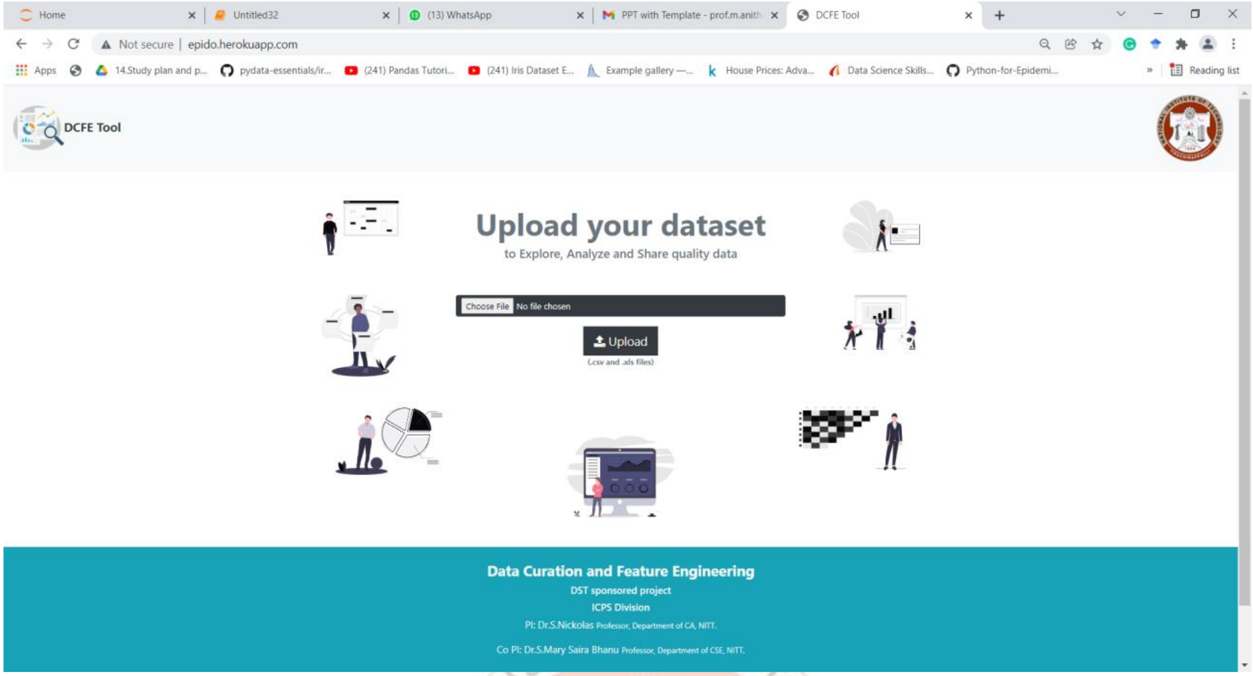
The user can select .csv files for uploading as both secondary and primary. The file must have common attribute between them in-order to do integration of data.

Download Dataset from External Source

The dataset can be downloaded from external source and a provision is given in the Download Dataset menu itself for the user.

The user is provided with the provision of downloading National and International datasets very easily through Data collection interface

Home



Upload your dataset

to Explore, Analyze and Share quality data

Data Curation and Feature Engineering

DST sponsored project

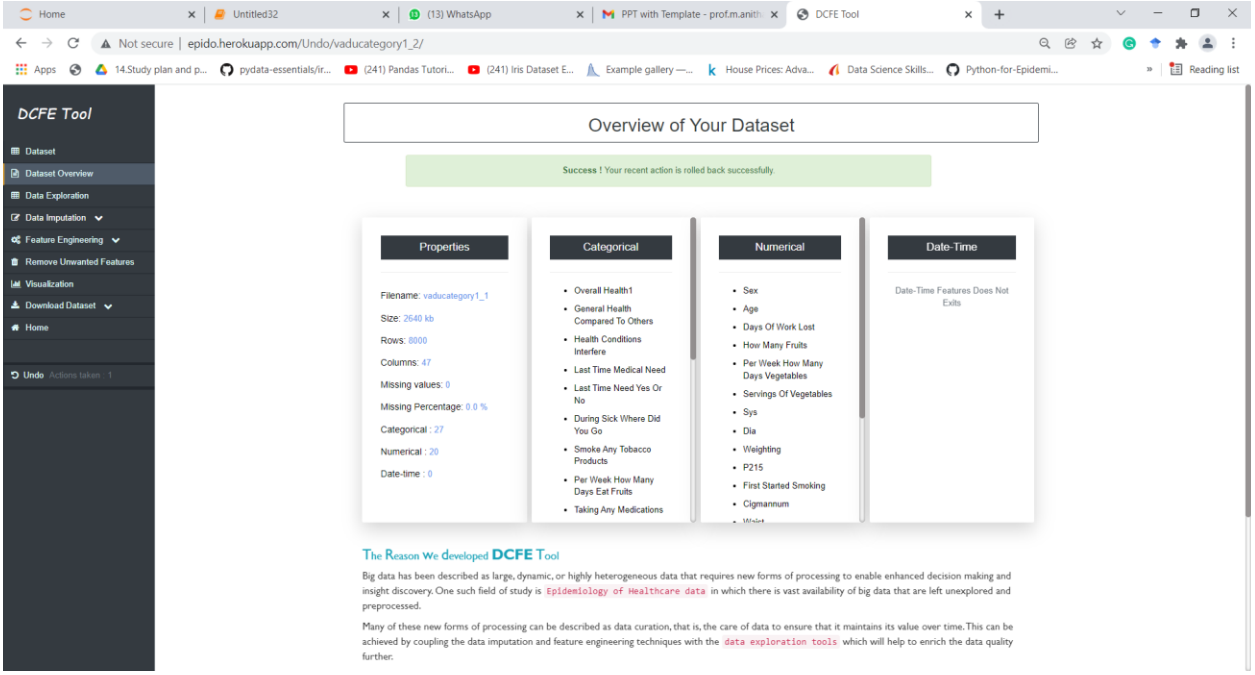
ICPS Division

PI: Dr.S.Nickolas Professor, Department of CA, NITT.

Co PI: Dr.S.Mary Saira Bhanu Professor, Department of CSE, NITT.

Home Link can be used by any user in-order to return to Home Page of the project. If the user is willing to move in case of uploading some other dataset, then this function link is useful for the user.

Undo



Overview of Your Dataset

Success ! Your recent action is rolled back successfully.

Properties

Filename: vaducategory1_1

Size: 2640 kb

Rows: 8000

Columns: 47

Missing values: 0

Missing Percentage: 0.0 %

Categorical : 27

Numerical : 20

Date-time : 0

Categorical

- Overall Health1
- General Health Compared To Others
- Health Conditions Interfere
- Last Time Medical Need
- Last Time Need Yes Or No
- During Sick Where Did You Go
- Smoke Any Tobacco Products
- Per Week How Many Days Eat Fruits
- Taking Any Medications

Numerical

- Sex
- Age
- Days Of Work Lost
- How Many Fruits
- Per Week How Many Days Vegetables
- Servings Of Vegetables
- Sys
- Dia
- Weighting
- P215
- First Started Smoking
- Cigmannum
- Waist

Date-Time

Date-Time Features Does Not Exits

The Reason We developed **DCFE** Tool

Big data has been described as large, dynamic, or highly heterogeneous data that requires new forms of processing to enable enhanced decision making and insight discovery. One such field of study is `Epidemiology of Healthcare data` in which there is vast availability of big data that are left unexplored and preprocessed.

Many of these new forms of processing can be described as data curation, that is, the care of data to ensure that it maintains its value over time. This can be achieved by coupling the data imputation and feature engineering techniques with the `data exploration tools` which will help to enrich the data quality further.

If the user has done different operations, but one point of time they may think to delete an action which may finished already. At that time, the user is provided with an option ‘Undo’ link. If the link is clicked one time the recent action will be deleted. At a time one action can be removed and a counter is also displayed along with the link.