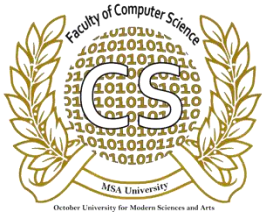


October University for modern Sciences and Arts (MSA)

Graduation Project Documentation

Semester 70 Spring 2020

Faculty of Computer Science

Mengen: A Web-based platform for learning and applying machine learning algorithms

Supervised by: Dr. Islam ElShaarawy

Student Name: Ahmed Ali Abdelkhalek Sayed

Student ID: 160647

Graduation Project Fall 2020

Project Title

Mengen: Web-based platform for learning and applying machine learning algorithms

Prepared by

Ahmed Ali Abdelkhalek Sayed

Supervisor

Dr. Islam ElShaarawy

Abstract

The business industry began plain and simple with managerial process is being done through papers. Throughout the years development of computer and internet has boomed and resulted in a huge shift in the world from using papers to computers to gain business advantage, be more efficient and save money. During this process web services emerged offering business and people easy solutions to problems that would cost a lot of money for the business to implement them. In the recent years the amount of data collected has exploded compared to the years before.

Consequently machine learning became much more used due to the availability of data. This has driven many industries to lean to machine learning and gain advantage over competitors using it. Nowadays machine learning is almost used in every new application and several companies are built just to provide Artificial Intelligence services.

Hence the rise on new technology comes new problems, this project currently aims to clarify the problems that were encountered emergence of Machine Learning and how Artificial Intelligence companies didn't provide solid core solutions that are powerful and at the same time simple, sociable and enjoyable. Machine Learning used to be a scary word for non-programmers, however since the emergence of open source libraries machine learning has been accessible to everyone even non-experienced users.

Table of Contents

Chapter 1: Introduction ***1***

1.1 Introduction **2**

1.2 Purpose **2**

1.3 Domain Background **2**

1.4 Problem Statement **7**

1.4.1 Data Visualization 9

1.4.2 Machine Learning 9

1.4.3 Computing power 9

1.4.4 Lack of machine learning references 11

1.4.5 Money wall 11

1.5 Aim **12**

1.6 Approach **12**

Chapter 2: Literature review ***13***

2.1 Introduction **14**

2.2 How computer scientists learn **14**

2.3 Similar System **15**

2.3.1 Matlab 15

2.3.2 GNU Octave 15

2.3.3 Scilab 16

2.3.4 Sage .....

2.3.5 DataPine..... 16

2.4 Mengen.....18

2.4.1 Introduction..... 18

2.4.2 System features.....19

2.5 Intended Audience and Reading suggestions.....25

..... 25

Chapter 3: Project Specification and Planning.....26

3.1 Programs and libraries.....27

Chapter 4:.....30

4.1 System Functional Requirements – FRs.....31

4.1.1 Registration..... 31

4.1.2 Login..... 31

4.1.3 System homepage.....31

4.1.4 Add friends..... 31

4.1.5 Chat..... 31

4.1.6 Full Website chat.....31

Forums..... 31

4.1.7 Upload dataset files.....32

4.1.8 Groupby function presented in a Bar chart..... 32

4.1.9 Linear regression..... 32

4.1.10 Logistic regression.....32

4.1.11 Decision tree classifier..... 33

4.1.12 Support Vector machine..... 33

4.1.13 Add attribute.....

4.1.14 Remove attribute..... 33
4.1.15 Fill null attributes..... 33
4.1.16 Label encoder..... 34
4.1.17 Get dummies..... 34
4.1.18 Change cells value..... 34
4.1.19 Removing rows that contain certain cell value..... 34
4.1.20 Calculating an attribute..... 34
4.1.21 Show table..... 35
4.1.22 The table Sorting values..... 35
4.1.23 Search for certain value..... 35
4.1.24 Describe info..... 35
4.1.25 Describe ordinal..... 35
4.1.26 Data profile..... 36
4.1.27 Code Logs..... 36
4.1.28 Session-based Website..... 36
4.1.29 Sharing Screen and holding teaching sessions..... 36
4.1.30 Advanced text editor..... 36
4.1.31 Python editor..... 37
4.1.32 R editor..... 37
4.1.33 Calendar..... 37
4.1.34 Download Data Set..... 37
4.1.35 Download code logs..... 37
4.1.36 Create and download pandas profile..... 38
4.1.37 Download Machine learning model..... 38
4.1.38 Session and dataset code..... 38
4.1.39 Algorithms and functions code..... 38
4.1.40 Links and tips.....

4.1.41 Customer support..... 39

Chapter 5: External Design..... 40

5.1 System user interface..... 41

Chapter 6: Internal Design..... 58

6.1 Deployment diagram..... 58

6.2 Main Flow Diagram..... 59

6.3 Use case Diagrams..... 61

6.4 Sequence Diagram..... 66

6.5 Database..... 76

6.5.1 ERD Diagram..... 76

Chapter 7: Testing..... 79

7.1 Software testing..... 80

7.1.1 Testing process..... 80

7.1.2 Unit testing..... 80

7.1.3 Integration testing..... 81

Chapter 8: Conclusion..... 99

Chapter 9: Future Work..... 101

9.1 Adding containers..... 102

9.2 Adding more Preprocessing functionalities..... 102

9.3 Adding more machine learning algorithms..... 102

9.4 Implementing a payment plan 103

9.5 Making a model to predict best steps.....103

9.6 Adding unsupervised, NLP algorithms and ability to build neural networks 103

9.7 Using D3.js library..... 104

9.8 Using Qualas.....104

9.9 Implementing a wizard..... 104

9.10 Implementing dataflow/graphical representation..... 105

9.11 Adding persistent.....105

Chapter 10: References..... 105

List of figures

User Interface

Figure 5- 1. Home page..... 41

Figure 5- 2. Login interface.....41

Figure 5- 3.Register interface..... 42

Figure 5- 4.User home page..... 42

Figure 5- 5.User Profile.....43

Figure 5- 6.Another user profile.....43

Figure 5- 7.User search bar..... 43

Figure 5- 8.Chat page.....44

Figure 5- 9.User to user chat box..... 45

Figure 5- 10.Website chat box.....46

Figure 5- 11.Upload dataset.....47

Figure 5- 12.Shape of dataset after being uploaded.....47

Figure 5- 13.Describe data.....47

Figure 5- 14.Describe ordinal..... 48

Figure 5- 15.Pandas profile selection screen..... 48

Figure 5- 16.Pandas profile quick analysis.....

Figure 5- 17.Pandas profile each variable info.....49

Figure 5- 18.Pandas profile correlation presented in heat map with different metrics.....50

Figure 5- 19.Pandas profile further analysis.....50

Figure 5- 20.Group by and present in bar chart.....51

Figure 5- 21.Bar Chart generated.....51

Figure 5- 22.Preprocessing.....52

Figure 5- 23.Machine Learning algorithms.....52

Figure 5- 24.Code of the algorithm.....53

Figure 5- 25.Link of algorithm and tips can be added.....53

Figure 5- 26.Code Logs.....54

Figure 5- 27.Download file page.....54

Figure 5- 28.Python editor.....54

Figure 5- 29.R editor.....55

Figure 5- 30.Text editor.....55

Figure 5- 31.Code steps.....56

Figure 5- 32.Together js for session sharing.....56

Figure 5- 33.Calendar.....57

Figure 5- 34.Customer support.....57

Use Case diagrams

Figure 5- 35.Login/Register.....61

Figure 5- 36.User home options.....

Figure 5- 37.Machine Learning Algorithms..... 63

Figure 5- 38.Preprocessing..... 64

Figure 5- 39.Download Files..... 65

Sequence diagrams

Figure 5- 40.Login/Register..... 66

Figure 5- 41.Upload dataset..... 66

Figure 5- 42.Group by represented in bar chart..... 67

Figure 5- 43.Describe/Describe ordinal.....68

Figure 5- 44.Search user & add user..... 68

Figure 5- 45.Code Logs..... 69

Figure 5- 46.Code Steps.....70

Figure 5- 47.Algorithm code, Tips and links..... 71

Figure 5- 48.Download Files..... 72

Figure 5- 49.Python, R, Text editors..... 72

Figure 5- 50.Machine learning Algorithms..... 73

Figure 5- 51.Pandas Profilling..... 74

Figure 5- 52.Preprocessing..... 75

Figure 5- 53.Whole website/user to user chat.....76

List of Test tables

Table 6- 1. Register.....

Table 6- 2.Login..... 82

Table 6- 3.Search_profile..... 83

Table 6- 4.Add_user.....84

Table 6- 5.Chat_with_friend.....85

Table 6- 6.Chat_with_entire_website.....86

Table 6- 7.Upload_file..... 86

Table 6- 8.Add_attribute.....87

Table 6- 9.Remove_attribute..... 88

Table 6- 10.Fill_nan..... 89

Table 6- 11.Label_encoder.....91

Table 6- 12.Get_dummies.....92

Table 6- 13.Replace_cells_values.....93

Table 6- 14.Add_equation_to_calculate_an_attribute.....95

Table 6- 15.Groupby_in_bar_chart.....96

Table 6- 16.Machine_learning_algorithms..... 97

Chapter 1: Introduction

1.1 Introduction

The sudden increase in interest in machine learning made a lot of students and non-programmers seek to learn more about it and be able to understand and integrate it in their systems or be able to create custom machine learning model for research purposes, however even so there is a lack of tutorials or easy methods to do machine learning without having python or R first and the only company known to provide this sort of services is MATLAB which is a pay to use model, which is not ideal for learning specially for students, however this website Mengen aims to be a website where the user can do all things related to machine learning on it, from data pre-processing, data visualization and implementing machine learning algorithm to having live chat with entire website users through live chat and forums, and see algorithms and tips on using them.

1.2 Purpose

The purpose of this document is to present a detailed description about purpose of Mengen, its goals, target, functionality, most likely business model, the growth potential of the website, how each of them has that been achieved or can be achieved and how to improve upon each aspect of the website.

1.3 Domain Background

Pre processing

Data pre-processing is a technique in machine learning that revolves around the transforming raw data into usable and understandable format, real life data are usually not understandable in their row format as they are usually inconsistent, incomplete, lucks patterns and most likely to contain many errors.

Pre-processing consists on many steps:

Data Cleaning is the processing of finding and removing inaccurate, corrupt, incomplete rows of the dataset then doing the operation of modifying or deleting the data through the use of data wrangling tools, or using batch processing scripts.

Data editing is defined as a process that includes reviewing and adjustment of collected survey data. The purpose of it is to manage the quality of the collected data. Data editing can be done manually using a computer or a combination of both.

Data reduction is conversion of numerical or alphabetic data into a modified, ordered, and simplified form. The basic concept is to reduce large amounts of data into meaningful parts. The main reason for doing such process is called Curse of dimensionality, when data has many attributes the machine learning model has high probability to miss identify the important attributes from the useless ones which will lead to a low accuracy machine learning model, data reduction aims at targeting those useless data and removing them which makes machine learning algorithm run faster and have better accuracy.

Data wrangling sometimes referred to as data modification, it transforms and maps data from one “raw” data format to another for the purpose of making the data more relevant and valuable for various purposes such as analysis Process.

Data Visualization

Data visualization means getting data and placing it in a visual context such as a map or graph, it makes it easier for the human brain to understand large and small data, and visualization makes it easy to detect patterns, trends, and outliers in a data group. By properly visualizing the data, complex data sets should be meaningful and messages should be clear. Some examples are Column Chart, Bar Graph, Line Graph, Dual Axis Chart, Area Chart, Stacked Bar Graph, Mekko Chart, Pie Chart, Scatter Plot Chart, Bubble Chart, Waterfall Chart, Funnel Chart, Bullet Chart, Heat Map.

Machine Learning

Supervised learning is a branch in machine learning. It is the task of learning through the data provided. In supervised learning the data provided are always paired with the desired output value. In order to train a model, the data is divided into two parts (Training, Testing) sets, then model is trained with the training set, attributes in both sets are usually vectors. The model then analyses training data and produces a function which can be used to predict new examples. The testing set is then used to test how well is that function is doing by using one of the measures.

Web Service

A web service is a software service used to communicate between two devices on a network, one of these devices is used to access functions that other device

provides, the connection between the two devices is usually done through HTTP using technologies such as XML, SOAP, WSDL, and UDDI.

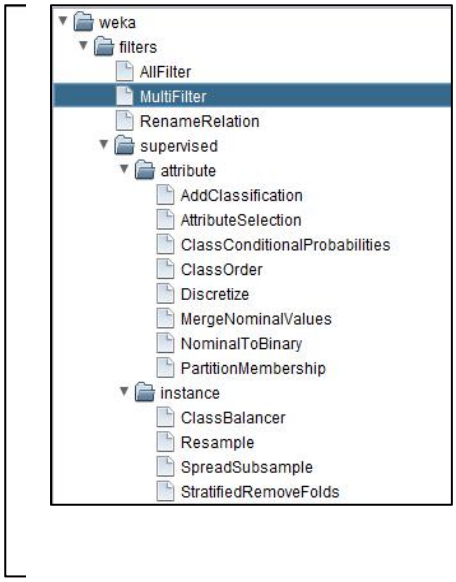
Cloud Computing

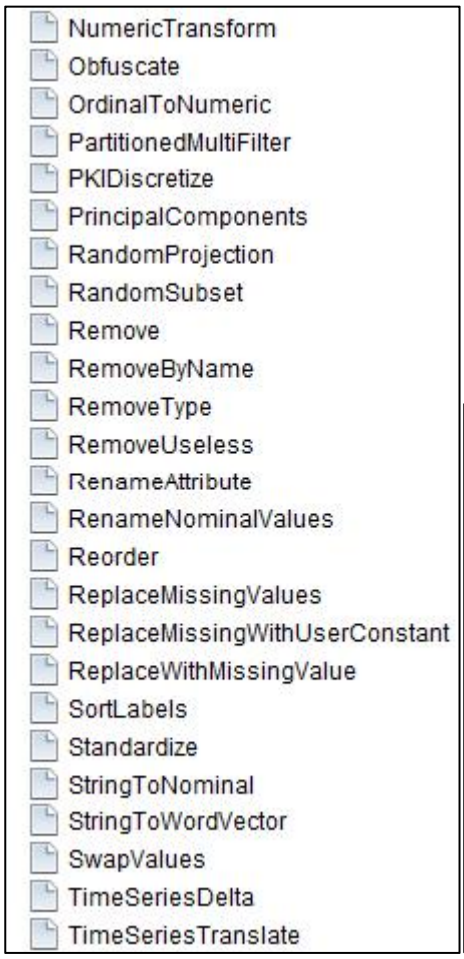
Cloud computing is the on-demand availability of computer system resources, especially data storage and computing power, without direct active management by the user. This term is generally used to describe a data center that is available to many users over the Internet. The large clouds that are mainstream today often have functions distributed from a central server to multiple locations. If the connection to the user is relatively close, you can designate it as an edge server.

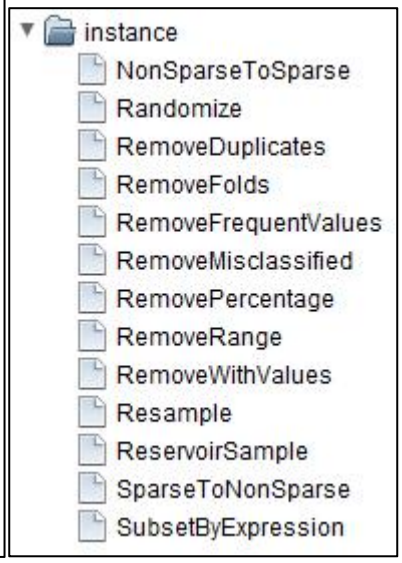
1.4 Problem Statement

This Website is meant to address more than one problem and they are as following: **Pre processing**

The problem in pre-processing is that there are many algorithms that are used in that process and for beginner and semi-experienced there are many algorithms that may not be known to them, also pre-processing can be very annoying for routine work, an application that worked to fix this issue was Weka but the amount of pre processing Weka can do is very limited and uncontrollable to some degree as the users to can only select algorithm name the hit apply, Mengen aims towards fixing this issue by having more algorithms to be able to use on click, the algorithms used have certain amount of flexibility to them which will allow users to have better experience when doing data cleaning than in Weka. Examples of pre-processing algorithms that are currently available in Weka:


weka
filters
AllFilter
MultiFilter
RenameRelation
supervised
attribute
AddClassification
AttributeSelection
ClassConditionalProbabilities
ClassOrder
Discretize
MergeNominalValues
NominalToBinary
PartitionMembership
instance
ClassBalancer
Resample
SpreadSubsample
StratifiedRemoveFolds


NumericTransform
Obfuscate
OrdinalToNumeric
PartitionedMultiFilter
PKIDiscretize
PrincipalComponents
RandomProjection
RandomSubset
Remove
RemoveByName
RemoveType
RemoveUseless
RenameAttribute
RenameNominalValues
Reorder
ReplaceMissingValues
ReplaceMissingWithUserConstant
ReplaceWithMissingValue
SortLabels
Standardize
StringToNominal
StringToWordVector
SwapValues
TimeSeriesDelta
TimeSeriesTranslate


instance
NonSparseToSparse
Randomize
RemoveDuplicates
RemoveFolds
RemoveFrequentValues
RemoveMisclassified
RemovePercentage
RemoveRange
RemoveWithValues
Resample
ReservoirSample
SparseToNonSparse
SubsetByExpression

1.4.1 **Data Visualization**

The problem in data visualization that it takes some one with some sort of coding experience to be able to create data representations for inexperienced users that causes a problem, even though it is relatively easy to make the main plots with the libraries available the more unique the plot becomes the much harder it becomes to make a plot and only experienced users are the ones who are able to make this sort of plots.

1.4.2 **Machine Learning**

For inexperienced users implementing machine learning algorithms can be quite hard, Mengen can help user implement a machine learning algorithm just by selecting algorithm name and hit run. For experienced users having the ability to quickly implement a machine learning algorithm without having to fine tune it is used a lot in order for the Data Scientist to know if this machine learning algorithm is convenient for his model, also to determine minimum threshold he should work with.

1.4.3 **Computing power**

Usually machine learning algorithms takes huge time to train this time can even span to weeks of to train a model, The reason for that is because of the computational complexity of the machine learning algorithms, where every data is changed to vectors

to be calculated, however the computing power used in relation to data is exponential, this can cause huge problems for users who want to train a machine learning model but lack computing power or memory space.

1.4.4 **Lack of machine learning references**

The sudden increase in machine learning leaded to the rapid increase in machine learning algorithms and they keep increasing day by day and the old algorithms keep getting updated and variants of old algorithms keep appearing, this sudden expansion lead to many people even the experts to be ignorant about many of those algorithms, thus people have to keep searching through the internet for a certain algorithms if they know it even exists, and usually for python users if this algorithms exists outside SKlearn library they won't know about it.

1.4.5 **Money wall**

Weka is an application that is an open source software used for quick data cleaning, analysis, and machine learning, however the uses of Weka are limited due the unability to select parameters in machine learning algorithms or running pre-processing algorithms that requires custom input from the user such as deleting certain rows, The only other application known similar to my project is Matlab online which requires users to be students with students code from the university or pay for it, for student not every single university take the trouble to make contact with Matlab to get student codes which limits the students.

1.5 Aim

The aim of this project is to make a single free platform that can allow new and experienced users to learn about the new machine learning algorithms and how to implement them in a very simple way, it also provides brief information about the algorithm and tips of when to use it also the link the to the algorithm if users wants to read more, the platform also provides a way for the user to upload his data and be able to do data pre processing just by selecting the function then selecting the attribute and clicking submit, the same process can be done on machine learning algorithms, the data pre processing and running machine learning algorithms should not be run using the computing power of the user, platform will also be able to provide data visualization with different types of interactive charts, use platform should also provide tools for users to learn in an interactive way.

1.6 Approach

After considering many possibilities it is found that the best solution to the problems proposed above is by creating a service-oriented web-based application.

Chapter 2: Literature review

2.1 Introduction

Web sites nowadays are not just limited to sharing information and communications between people, but it has risen to be used in educational institutions as a teaching tool, in business and factories as a management as tracking tool to manage and improve the business, also as a way to communicate between the users of the website.

Thus the rise of web services and cloud computing, web services over the years it evolved since the emergence of the first web service which was the Remote Procedure Call mechanism of Distributed Computing Environment, a software development framework from the early 1990s. Cloud computing has become the main stream for several years, for example statistics show that 90% of the companies are on the cloud and by the end of year 2019 the 60% of workloads will be running on cloud, considering that the cloud hosted 45% in the year 2018 this shows the amazing increase in the use of cloud. (Hosting Tribunal, 2020)

2.2 How computer scientists learn

Usually in most majors the type of education is based on formal institutions whether they are universities, colleges or schools, for those major the information they get in the university is enough for them to start their career and be good at their work, but for computer science its different, what computer science major offers for

the student is a way of how to think, how to find information the student needs and provide information for student to understand most basic information and concept in every field.

For computer science field in general most of the the persons knowledge come from learning using the sources provided from the internet, this sources vary from being online courses to groups specialized in a field, to reading papers and thesis.

2.3 Similar System

2.3.1 **Matlab**

MATLAB is a numerical computing environment developed by MathWorks and a proprietary language. MATLAB allows you to manipulate matrices, plot functions and data, implement algorithms, create user interfaces, and interface with programs written in other languages such as C, C ++, C #, Java, Fortran and Python.

2.3.2 **GNU Octave**

GNU Octave is the best known replacement for MATLAB. Actively developed for nearly 30 years, Octave works the same on Windows, Mac and Linux and is packaged for most major distributions.

Octave has a variety of options that can be used for non-default front-end interactions that ship with the new version. Some are more similar to the MATLAB interface than others.

2.3.3 **Scilab**

Scilab is an open source option for numerical calculations and runs on all major platforms including Windows, Mac and Linux. Scilab is probably the best-known alternative outside Octave, and like Octave, its implementation is very similar to MATLAB, but the exact compatibility of project developers is not the goal.

2.3.4 **Sage**

Sage is an open source math software system that may be a good option for anyone looking for an alternative to MATLAB. Built on top of various famous Python-based scientific computing libraries, its own language is syntactically similar to Python. There are many features such as a command line interface, browser-based notebook, tools for embedding formulas in other documents, and of course many math libraries.

2.3.5 **DataPine**

Datapine is a tool helps businesses transform data into valuable new insights and make data-based decisions in real time. An easy to use drag and drop interface enables

even non technical users to understand complex data sets and answer important business questions on the spot. Datapine offers a variety of innovative BI 4.0 features, including intelligent alerts, predictive analytics, and interactive dashboard capabilities.

2.4 Mengen

2.4.1 **Introduction**

Mengen is a huge and long term project with an enormous potential that will aim to be leading website for machine learning as it provides most major services a person that will need in one platform, while developing this website we took into consideration the following points:

- You learn by trying. The only way to improve programming is to actually try.
- Programming is not like studying for a test. Remembering things is not so important.
- Copying code is completely tolerated, most programmers use Google to solve most of their problems.
- Learning by yourself is really hard. Talking about code and working with other ambitious developers is a better learning method for most people.
- Trying to understand everything is a lost cause.
- Pair programming is one of the best ways to code.
- You can always as someone for help.
- You don’t need to be a math genius to be a data scientist or analyst.

- You might feel lost with the computer updates that keep appearing every day.
- Learning to code might not be easy.

2.4.2 **System features**

2.4.2.1 User Registration

Each user will have to create an account. The user will provide some credentials such as (user name or email address, password, etc.) to the system to prove their identity.

2.4.2.2 Email verification

Upon registration of the users, they will be required to activate their account by clicking on the link on the email they registered with.

2.4.2.3 Management of activities

The user is provided with tools that will help him to brain storm his ideas and be able to manage and be more organized while doing his tasks.

2.4.2.4 Dataset Upload

The user is be able to upload data to the website so he can work on it, the website will be able to support most of the versions of structured data. Such as (CSV, JSON, XML, etc…)

2.4.2.5 Data Pre processing

The website provides the users very simple way to do data pre processing and the result will be implemented on the uploaded data set

2.4.2.6 Data analysis

The website provides the users with the ability to see the data uploaded in a table with extra options such as search and sort, it will also provide data about the uploaded file

2.4.2.7 Data visualization

The website allows users to place uploaded data information it in a visual context such as a map or graph. Data visualization makes it easier for the human brain to understand large and small data, and visualization makes it easier to detect patterns, trends, and outliers in groups of data.

2.4.2.8 Machine learning

The website allows user to implement a machine learning algorithm just be selecting an algorithm then click submit.

2.4.2.9 Download dataset

Each user has the ability to download the data he worked on in any sort of format he wants of structured data such as (CSV, JSON, XML, etc…)

2.4.2.10 Download machine learning model

Each user has the ability to download the latest machine learning model he worked on.

2.4.2.11 Download python snippet

Every single operation the user did on the dataset will be written in form of python code then, the user can download it. After download user has to only read the file in python and run, and it give same result as it was on website.

2.4.2.12 Real time collaboration

The website allows users to have text chat, audio chat and screen sharing so that the users can have interactive learning and business sessions

2.4.2.13 Profile page

Each user has a profile page where he can add his information, skills and other links if he wants.

2.4.2.14 Add friend

Each user has the ability to add any friend he wants by going to his profile and clicking add friend.

2.4.2.15 Messaging system

The website has a messing system that will allow users to live chat with the people they added to their friend list.

2.4.2.16 Whole website messaging system

The website has a chat where every single user in the website is automatically part of, this will make the website lively and people can be part of interesting topics and help each other out.

2.4.2.17 Customer service

Website has a customer service provided by professional websites in order to track any complains the users might have.

2.4.2.18 Free

The website when deployed will try its best to make this platform free of charge and try relying on donations and a payment model for heavy users and companies as the operation of such website is expensive.

2.4.2.19 Forum

After a lot of thinking instead of developing a blog for the website from scratch the creation of sub-reddit for the website was the best option, Its free, better GUI, more options and better management.

2.5 Intended Audience and Reading suggestions

This website is directed to data scientists in general. It also aims towards students and people who want to generally want to learn and understand machine learning with algorithms without having to search for hours and not understanding the sophisticated algorithm description.

	Mengen	Matlab online	Weka	Datapine
Social platform			**X**	X
Run ML algorithms			X	X
Server computing power			X	
Free		X		X
Session sharing		X	**X**	X

Chapter 3: Project Specification and Planning

3.1 Programs and libraries

The website is developed using HTML, Java Script, PHP, MYSQL, python.

python is an object-oriented high-level programming language:

1. It is one of the easiest programming languages, it is easy to learn and emphasizes readability which makes it very easy to maintain.
2. Python offers the least amount of codes to write amongst the other programming languages, it can reach 1/5 the size of other languages.
3. It has pre-built libraries for data analysis and machine learning like NumPy, matplotlib and sklearn.
4. Python is platform independent and it can be used across multiple platforms with minimal amount of ode changing.

NumPy is a library that is made of multidimensional array objects and a number of routines to process those arrays. NumPy is used for mathematical and logical operations on arrays.

Pandas is a powerful package for data science on python and has the following uses:

1. It can easily represent data in a way that is suitable for using for data analysis using its series or data frame structures.

2. It contains multiple methods for all sorts of data filtering.
3. It has the ability to read data from variety of formats
4. Writing input output code is very easy, for example reading a csv file requires only one line of code while in java it can take up to 15 lines

Sklearn is a machine learning library for python that provides a lot of supervised and unsupervised learning algorithms, it is built based on NumPy, pandas and matplotlib the type of algorithms it provides are:

1. Regression
2. Classification
3. Clustering
4. Model selection
5. Preprocessing

HTML stands for hypertext markup language it describes the visual appearance of the document that is displayed on browser

.

PHP is a server side scripting language that.

Java Script .JavaScript is a web programming language that can modify both HTML and CSS

MYSQL is one of the most popular open source SQL database management systems that is developed, distributed, and supported by Oracle Corporation.

AdminLTE is a popular open source WebApp template for management dashboards and control panels. Responsive HTML template based on the CSS framework Bootstrap 3.

Chapter 4:

4.1 System Functional Requirements – FRs

4.1.1 Registration

System shall allow users to create new accounts. The username shall be the name of the user inside the website.

4.1.2 Login

Systems shall provide users with means to access the account. They register do it on the website. News

4.1.3 System homepage

System shall provide news for the users about new articles and interesting topics

4.1.4 Add friends

System should allow users to add friends using their username.

4.1.5 Chat

Systems shall allow users to have a chat between the user which are added on friend list.

4.1.6 Full Website chat

System provides a room for all users where they can all chat on to provide live chat and feedback in any problems that are presented by the users.

Forums

System shall provide a sub reddit page on the reddit platform where the users can chat share interesting topics and resources.

4.1.7 **Upload dataset files**

System provided means to upload any data any structured data set. that python supports including Json,csv and txt.

4.1.8 **Groupby function presented in a Bar chart**

System provided means to group by any attribute provided such as groupby count, sum, mean, mad, median, min, max, mode, absolute, prod, standard deviation, var, sem, skew, Kurt, quantile, cumulative sum, cumulative product, cumulative max, cumulative min. as it is created as a prototype to prove that the website can host. Provide many types of visualization

4.1.9 **Linear regression**

System provides means of using the machine learning algorithm “linear regression” and selecting the parameters from a drop-down menu and then writing parameters values and adding them.

4.1.10 **Logistic regression**

System provides means of using the machine learning algorithm “Logistic regression” and selecting the parameters from a drop-down menu and then writing parameters values and adding them.

4.1.11 **Decision tree classifier**

System provides means of using the machine learning algorithm “decision tree” and selecting the parameters from a drop-down menu and then writing parameters values and adding them.

4.1.12 **Support Vector machine**

System provides a means of the machine learning algorithm “support vector machine” and selecting the parameters from a drop-down menu and then writing parameters values and adding them.

4.1.13 **Add attribute**

System provides means of adding an attribute to a data set and choosing its name in a very simple and efficient way. The values inside the added attributes is null values.

4.1.14 **Remove attribute**

System provides means of removing an attribute from the data set just by selecting the attribute.

4.1.15 **Fill null attributes**

System provides very simple and efficient way to fill all the null values in the attributes, the null values can be replaced by the mean, median, most frequent values of the attribute, It can also remove all rows that contains null value, the null values can be also replaced by a custom value the user can enter.

4.1.16 **Label encoder**

System provides very simple and efficient way to implement label encoder to an attribute just by selecting the attribute, Label encoder converts all unique categorical values in an attribute to discrete numbers.

4.1.17 **Get dummies**

System provides very efficient and simple way to apply get dummies algorithm on an attribute. Just by selecting the attribute, Get dummies converts all categorical values in an attribute to a group of attributes which consists of 0 and 1.

4.1.18 **Change cells value**

System shall allow users to change the value of any cell within the data set. Just by writing the old value and the new value, if the new cell value is NaN then the cell will be replaced by null value.

4.1.19 **Removing rows that contain certain cell value**

System shall allow users to remove rows that contain certain values just by writing old value and leaving new value cell empty.

4.1.20 **Calculating an attribute**

System shall allow the users to calculate a new attribute or an old attribute and in a form of a mathematical equation, the mathematical equation has to be written in Python format, if the equation requires values of certain attributes, writing the attribute name in the middle of the equation will be enough for the equation to work.

4.1.21 **Show table**

System shall allow user to see their data set he uploaded and edited in an organized and graphically clean way.

4.1.22 **The table Sorting values**

The system shall allow users to sort any attribute of the table just by clicking the arrow next to the attribute name, the sorting order is done automatically.

4.1.23 **Search for certain value**

The user can search for a certain value using the search bar up above the table, if more than one value is found within the table, all the relevant rules will be shown to the user.

4.1.24 **Describe info**

The system automatically calculates some equations about the numeric attributes and gets us some statistical information such as 25%, 50%, 75%, count, Max, Min, Min and standard deviation for each and single attribute

4.1.25 **Describe ordinal**

The system automatically calculates some equations about Categorical attributes and gets us some statistical information such as count, frequency most frequent, count of unique values

4.1.26 **Data profile**

The system provides statistical information about the selected attributes and it provides an analysis like type, unique values, missing values, quantile, mean, mode, median, standard deviation, some skewness, frequent values, histograms, correlation between variables, count, heat map visualization, etc'

4.1.27 **Code Logs**

The system provides a log that includes every single step that the user has done on his data including custom values that the user has entered. With the order The user use them in.

4.1.28 **Session-based Website**

System shall allow each user to has his own session with his own data and won't interfere with other users.

4.1.29 **Sharing Screen and holding teaching sessions**

System provides means for users to have voice chat communication and sharing the screen. Using Togetherjs The screen is not shared but the HTML info within it which makes the data transfer rate much faster which makes it more responsive.

4.1.30 **Advanced text editor**

The system provides an advanced text editor on the website, Which then can be copied and printed in a normal Word document and work just fine or can view the source code write it in HTML format and put it in an HTML document.

4.1.31 **Python editor**

The system provides a python editor and online python editor. That supports the most important libraries in machine learning such as pandas, sklearn, matplotlib and Seaborn, the users can write their own code and run it and see the result instantly on the other half of the screen.

4.1.32 **R editor**

The system provides an online or editor compiler Which supports the most important machine learning libraries that are used, the users can write their own code and run it and see the result instantly on the other half of the screen.

4.1.33 **Calendar**

The system has a calendar page which has the most important events the website provides to the users marked on the calendar And it can be used to schedule for an event or a session by a famous Professor. Or just to give a notice for maintenance on website.

4.1.34 **Download Data Set**

The system provides the users means to download dataset.

4.1.35 **Download code logs**

The system provides means for the users to download the code logs to download of the steps that have done so far on the uploaded dataset.

4.1.36 **Create and download pandas profile**

The system provides a means to download the HTML page of the pandas profile which is useful that it has many charts and many. Useful information many useful statistical information such as correlations presented in heat maps that follow different metrics. Information presented in bar charts like the count of values and their value.

4.1.37 **Download Machine learning model**

The system will allow users to download the last machine learning model. They have applied. With this set of parameters that they set and ran the machine learning model on.

4.1.38 **Session and dataset code**

The system provides a code for the user every time he uploads a data set to the website. And every step that that the user of does on the data set. is recorded and saved to a set of steps that shows as buttons, then when the user uploads at the set again he can Search for the code. We can enter the code in the search bar and history of steps with show to him. We can click on any step of them and the Algorithm will run automatically. With the set of custom values that the user wrote last time.

4.1.39 **Algorithms and functions code**

The system shall provide the code for every function and algorithm provided on the website in a simple and efficient way with the user has to click on the algorithm

then select the code tab. The code is written in a tidy clean and easy-to-understand way So that new users can understand its usability and its functionality.

4.1.40 **Links and tips**

The system shall provide users with Links and the original page of each function and website and will provide tips about When to use each function and algorithm and their significance The tips will be written and in an easy to understand simple way.

4.1.41 **Customer support**

System shall provide an easy way to provide customer support for users, which can be also used to get feedback from the users.

Chapter 5: External Design

5.1 System user interface

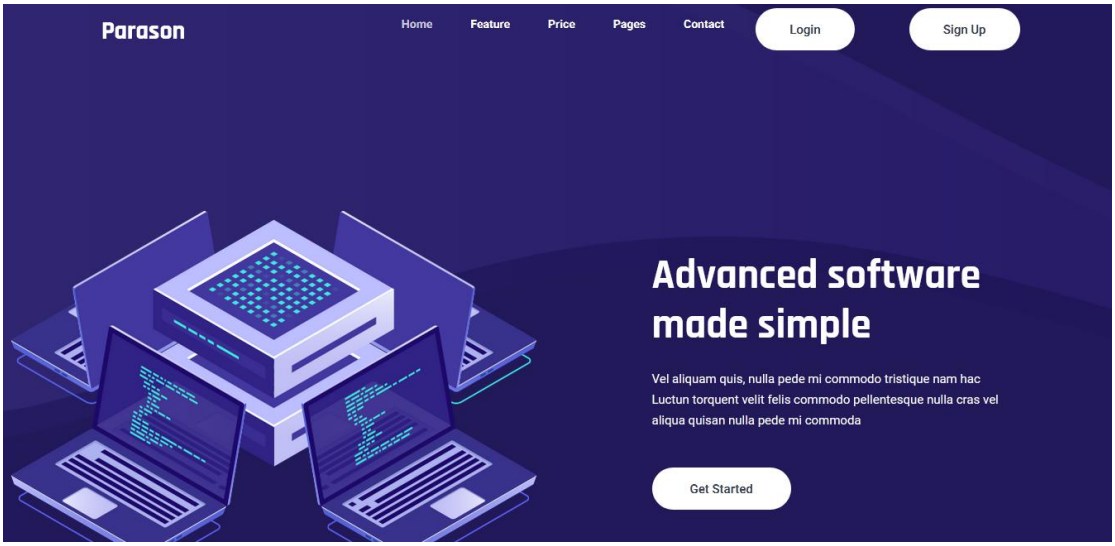


Figure 5- 1. Home page

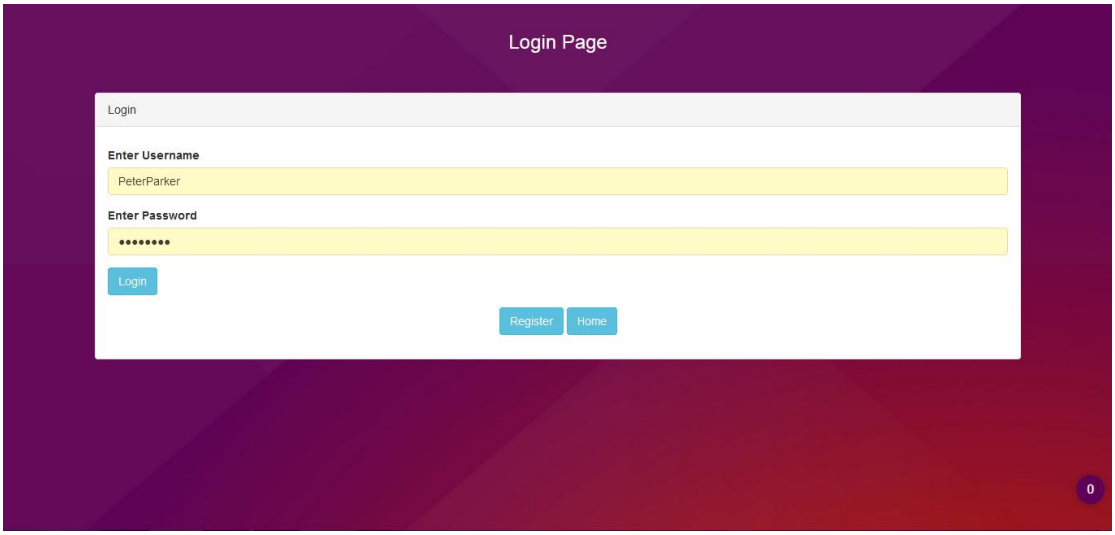


Figure 5- 2. Login interface

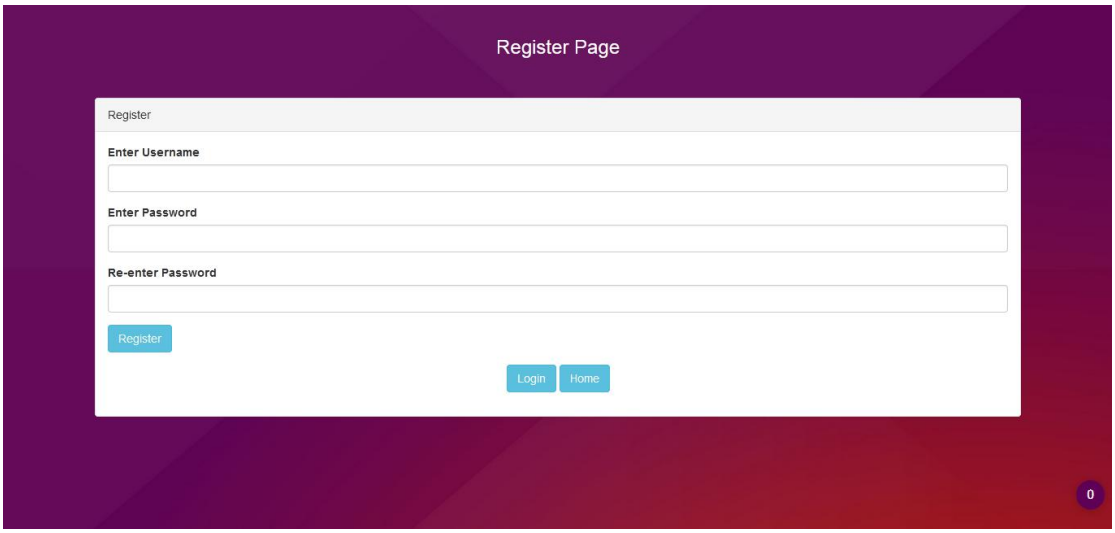


Figure 5- 3.Register interface

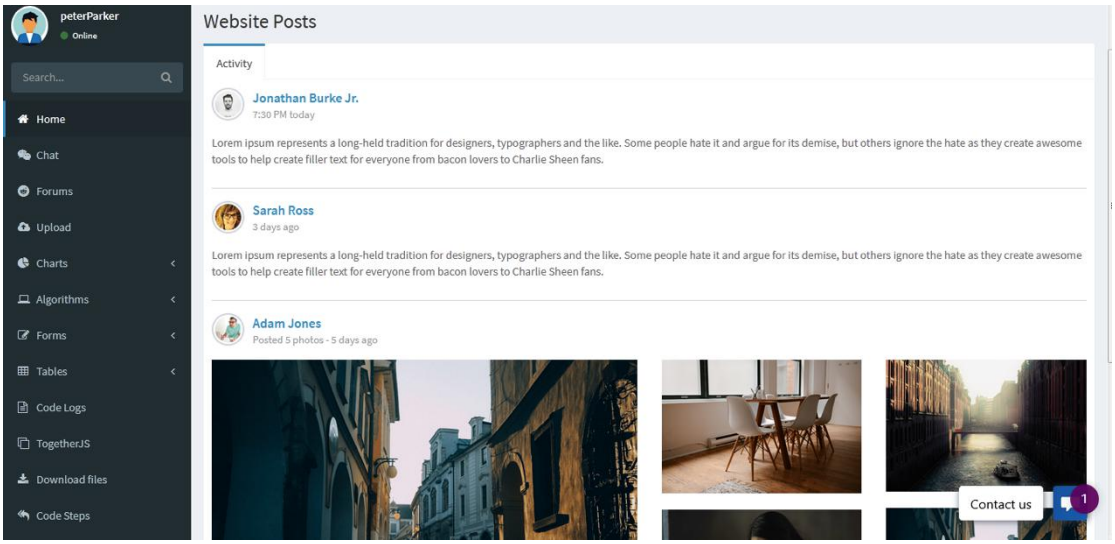


Figure 5- 4.User home page

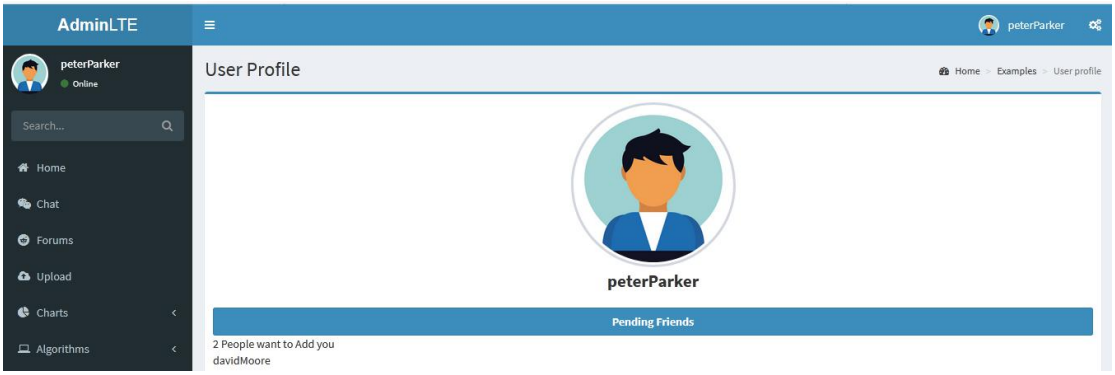


Figure 5- 5.User Profile

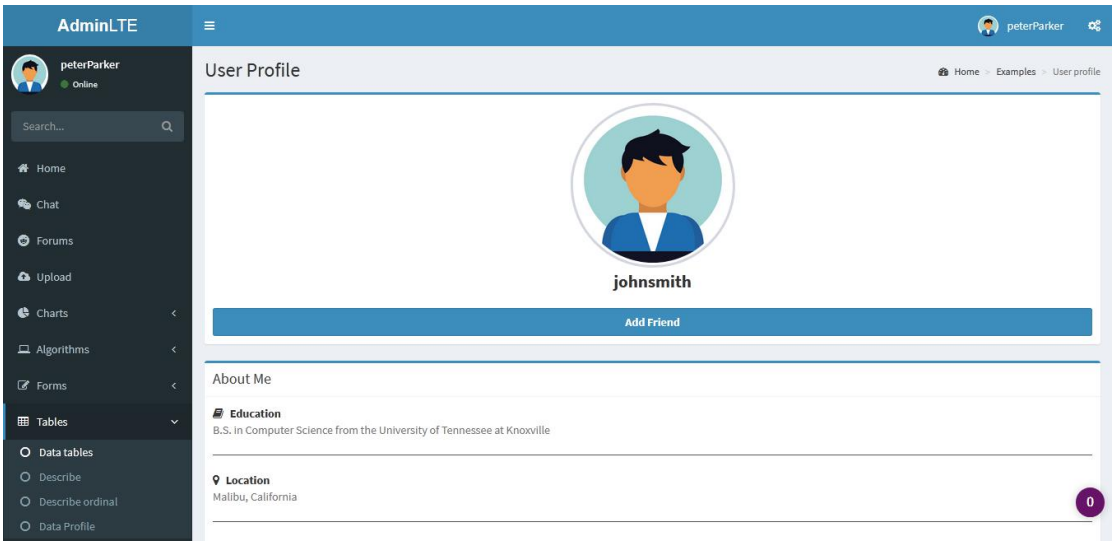


Figure 5- 6.Another user profile

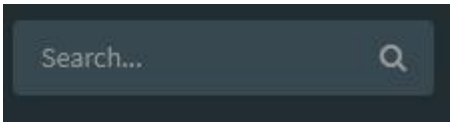


Figure 5- 7.User search bar

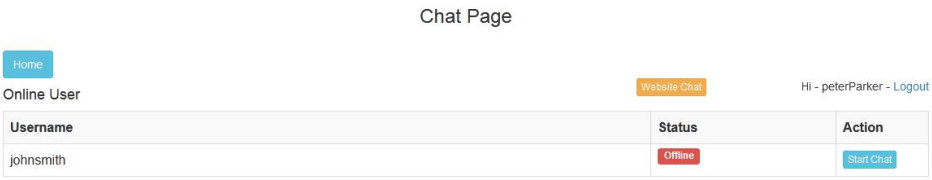


Figure 5- 8.Chat page

Figure 5- 9.User to user chat box

Figure 5- 10.Website chat box

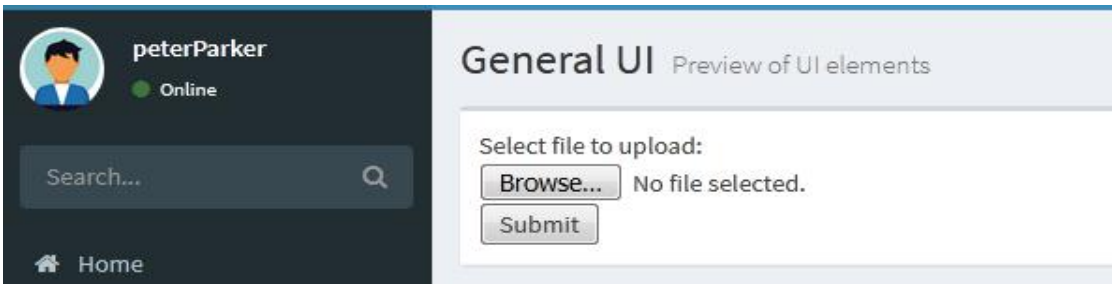


Figure 5- 11.Upload dataset

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	null	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925	null	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05	null	S
6	0	3	Moran, Mr. James	male	null	0	0	330877	8.4583	null	Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075	null	S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	null	S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	null	C

Showing 1 to 10 of 891 entries

Previous 1 2 3 4 Contact us

Figure 5- 12.Shape of dataset after being uploaded

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
25%	223.5	0	2	20.125	0	0	7.9104
50%	446	0	3	28	0	0	14.4542
75%	668.5	1	3	38	1	0	31
Count	891	891	891	714	891	891	891
Max	891	1	3	80	8	6	512.3292
Mean	446	0.3838383838	2.3086419753	29.6991176471	0.5230078563	0.3815937149	32.2042079686
Min	1	0	1	0.42	0	0	0
Std	257.3538420152	0.4865924543	0.836071241	14.5264973323	1.1027434323	0.8060572211	49.6934285972

Showing 1 to 8 of 8 entries

Previous 1 Next

Figure 5- 13.Describe data

	Name	Sex	Ticket	Cabin	Embarked
Count	891	891	891	204	889
Frequency	1	577	7	4	644
Top	Pain, Dr. Alfred	male	1601	G6	S
Unique	891	2	681	147	3

Showing 1 to 4 of 4 entries

Previous 1

Figure 5- 14.Describe ordinal

Figure 5- 15.Pandas profile selection screen

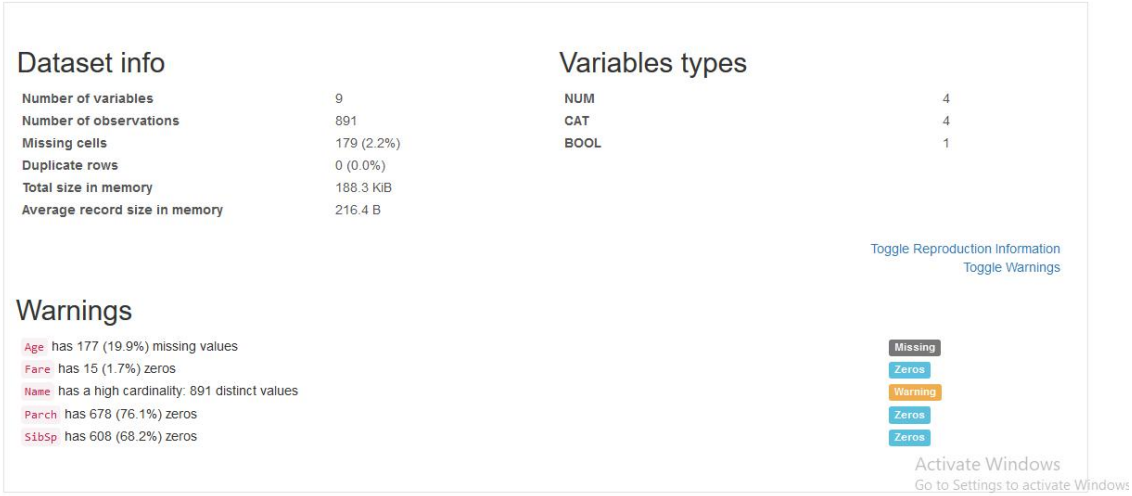


Figure 5- 16.Pandas profile quick analysis

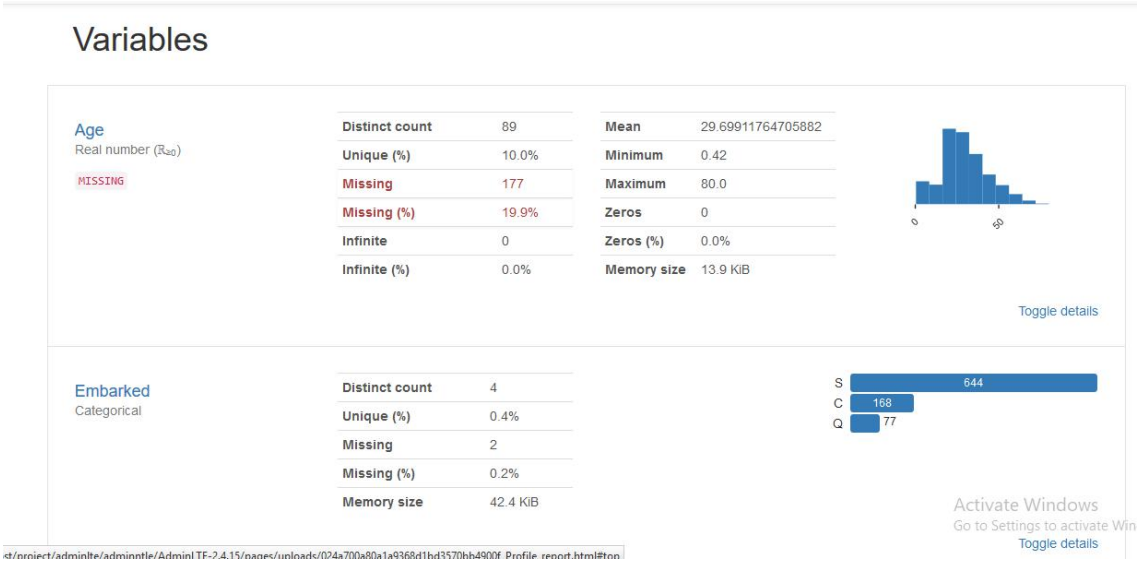


Figure 5- 17.Pandas profile each variable info

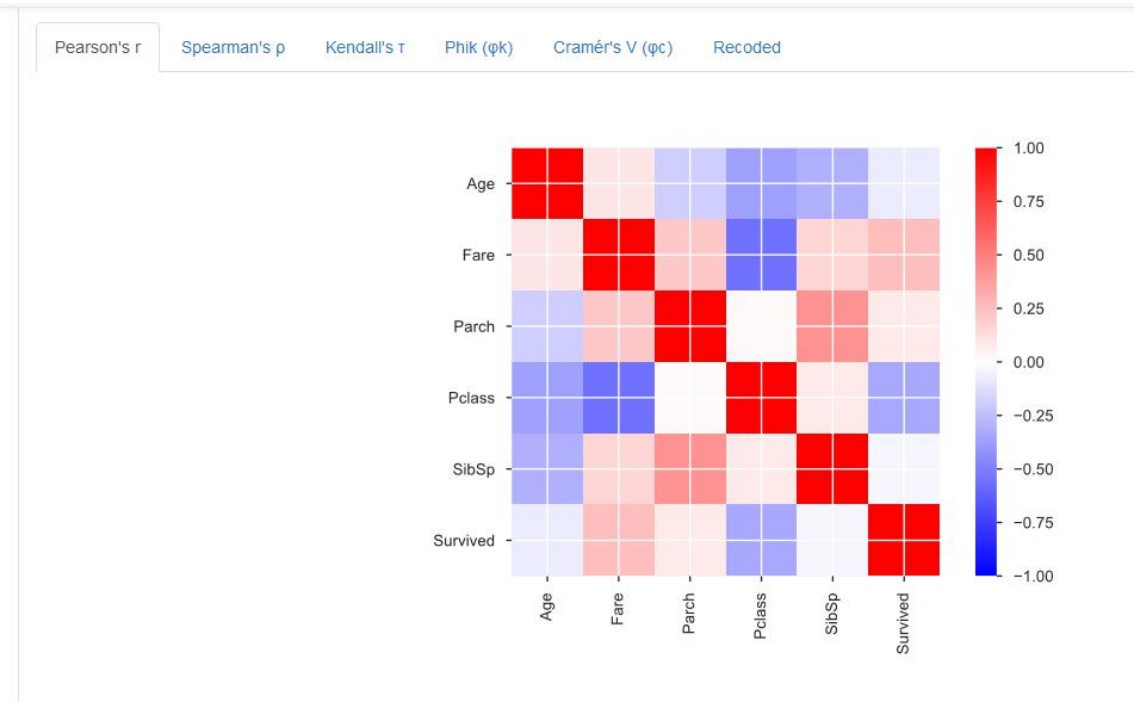


Figure 5- 18.Pandas profile correlation presented in heat map with different metrics

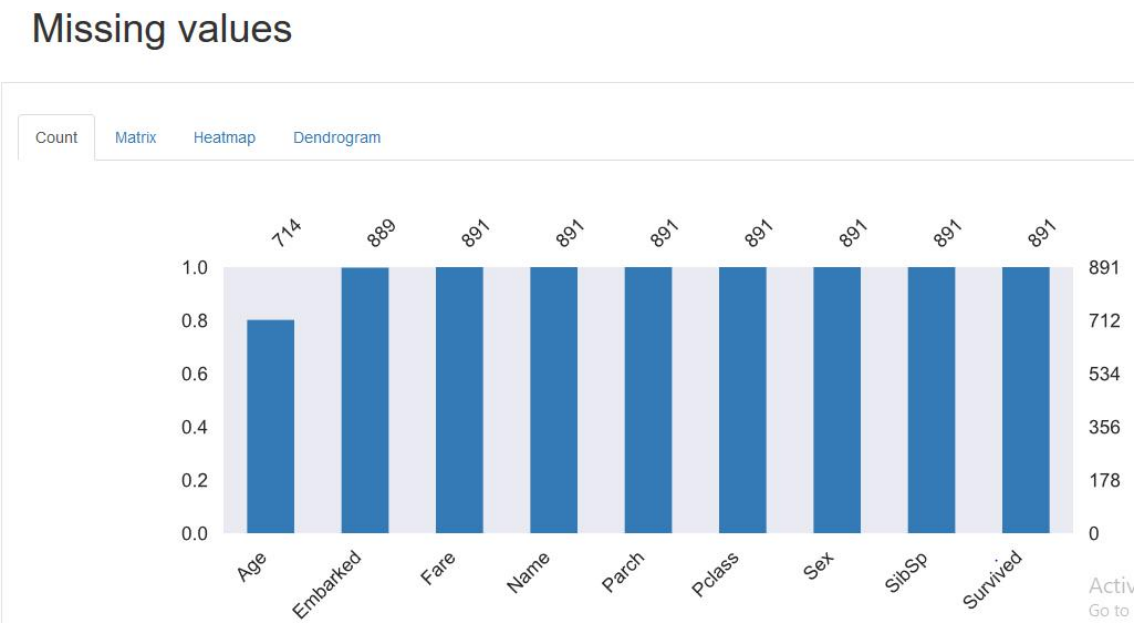


Figure 5- 19.Pandas profile further analysis

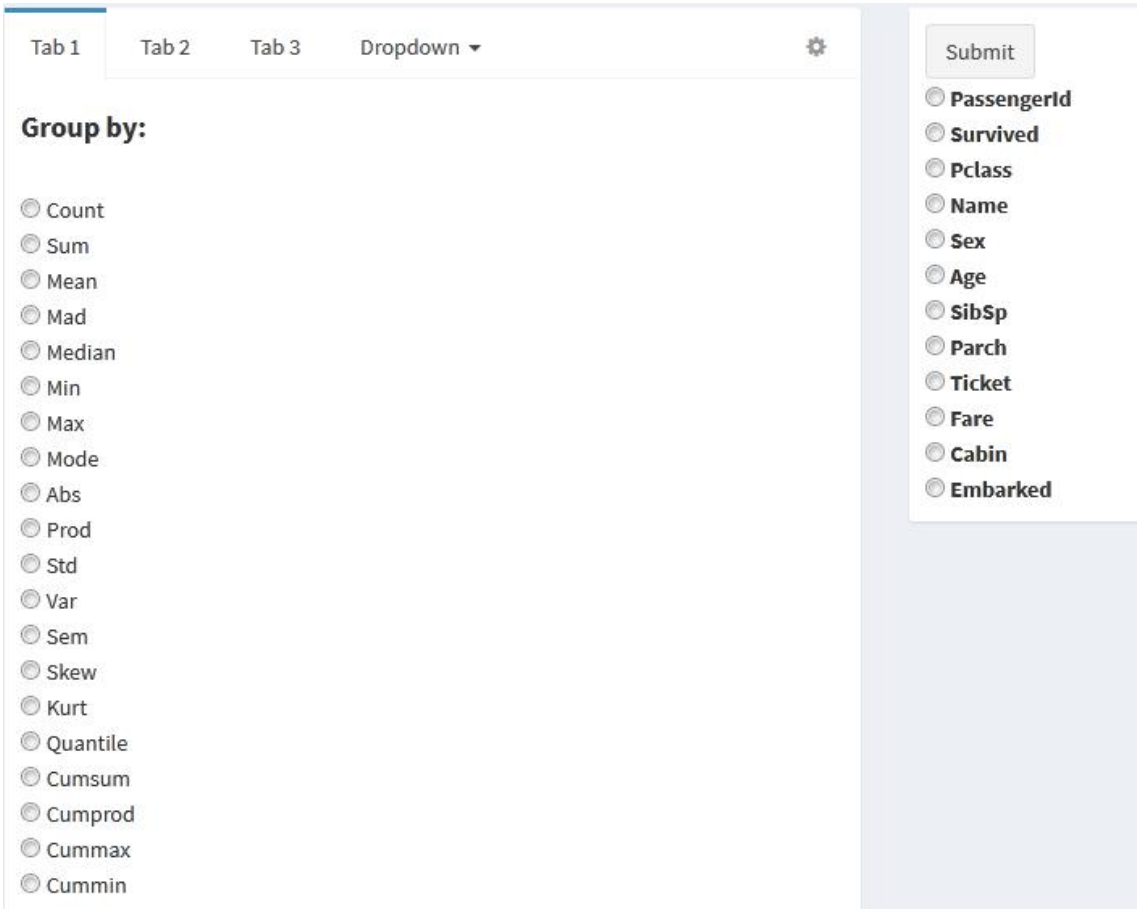


Figure 5- 20.Group by and present in bar chart

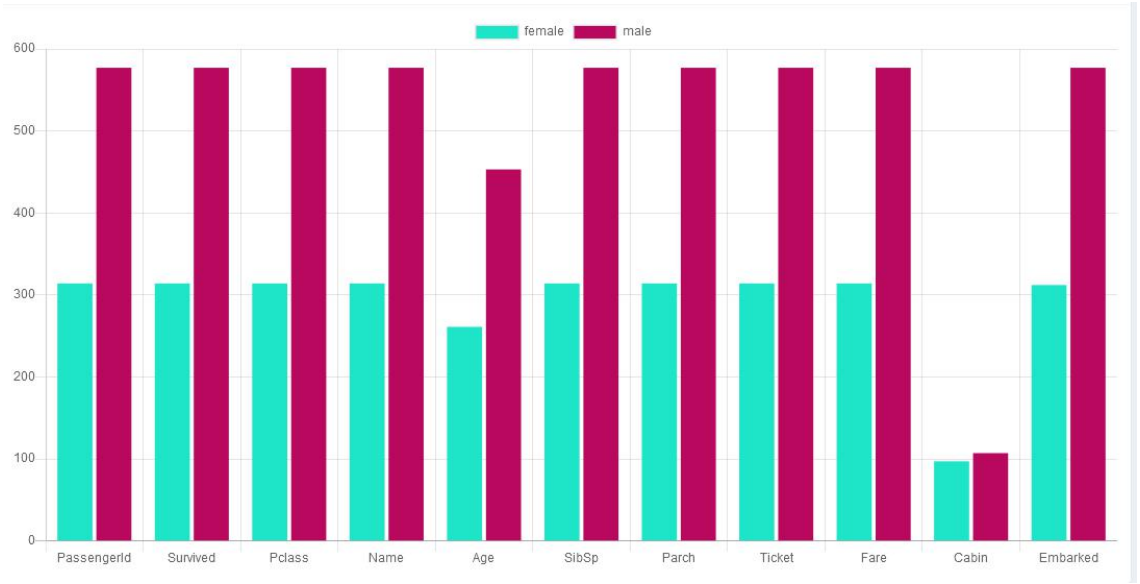


Figure 5- 21.Bar Chart generated

Figure 5- 22.Preprocessing

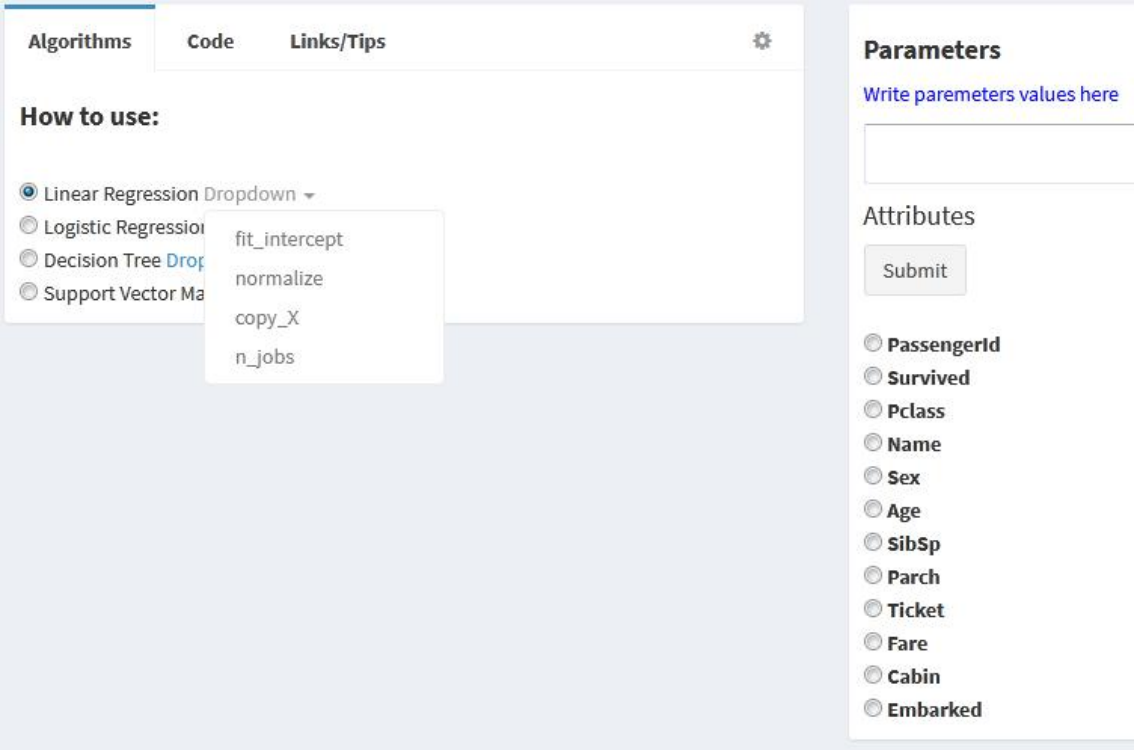


Figure 5- 23.Machine Learning algorithms

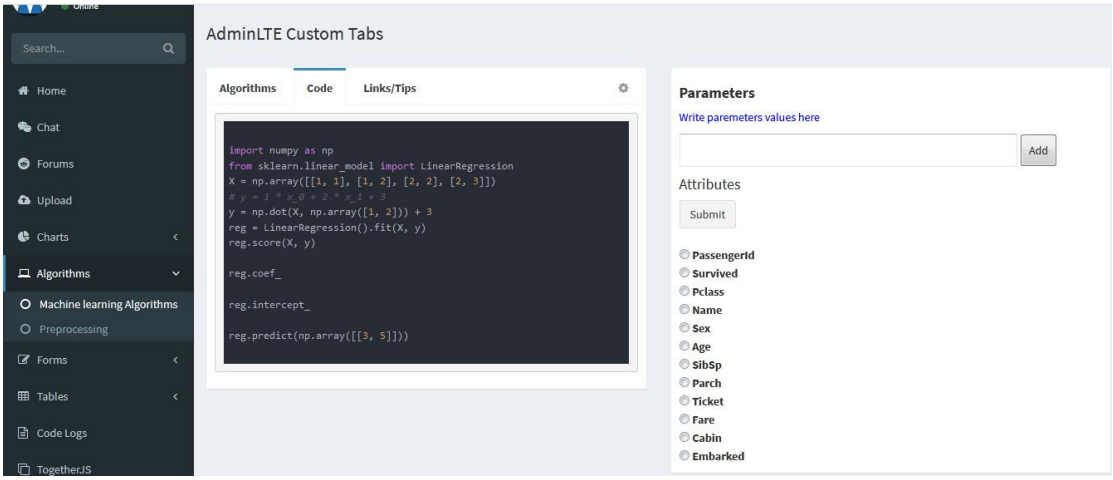


Figure 5- 24.Code of the algorithm

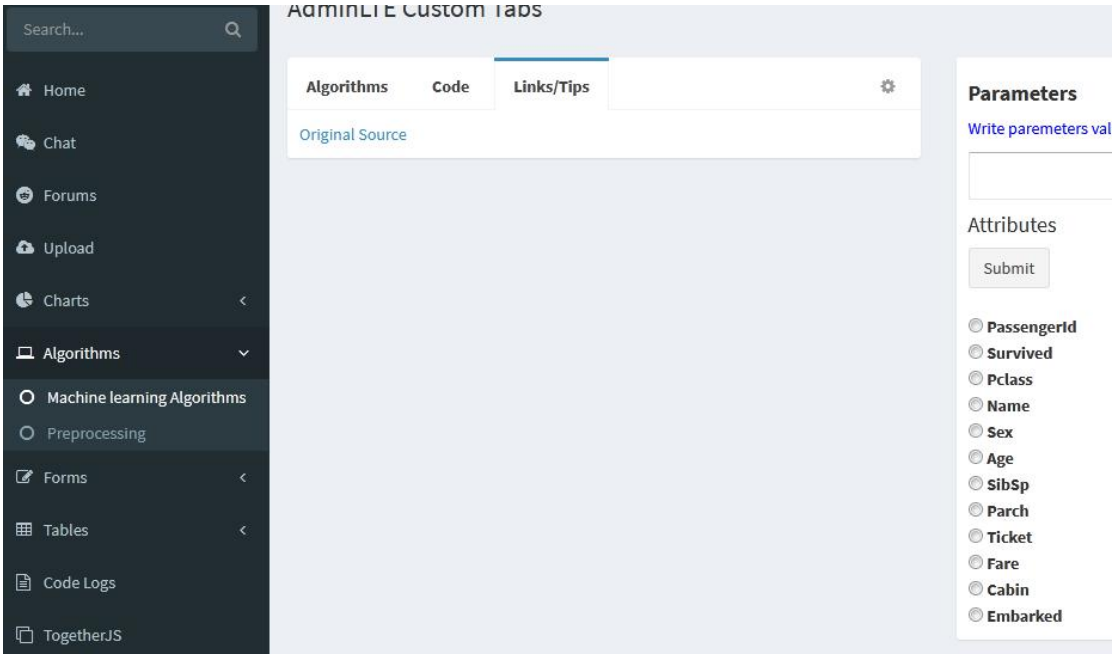


Figure 5- 25.Link of algorithm and tips can be added

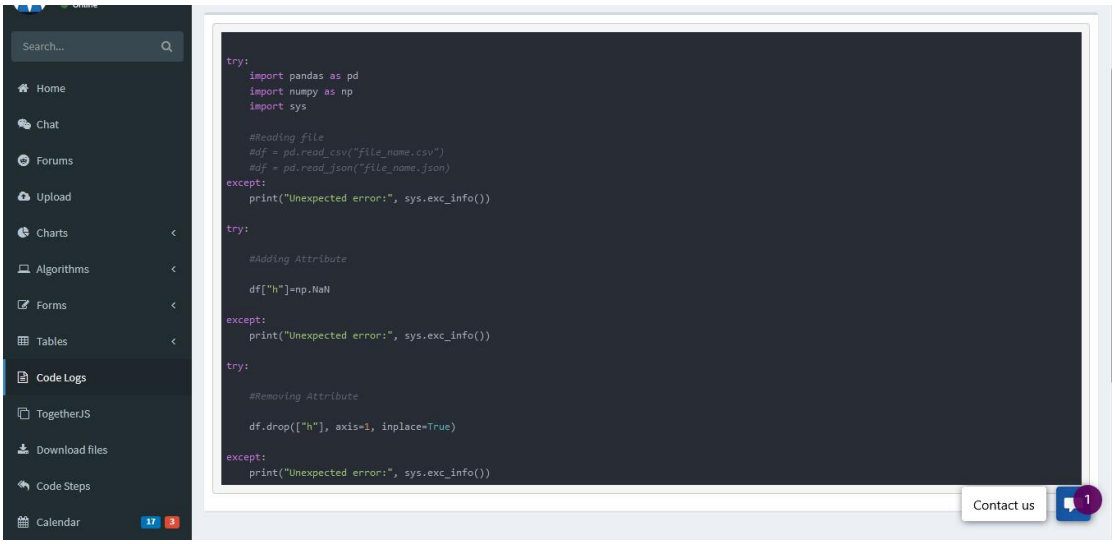


Figure 5- 26.Code Logs

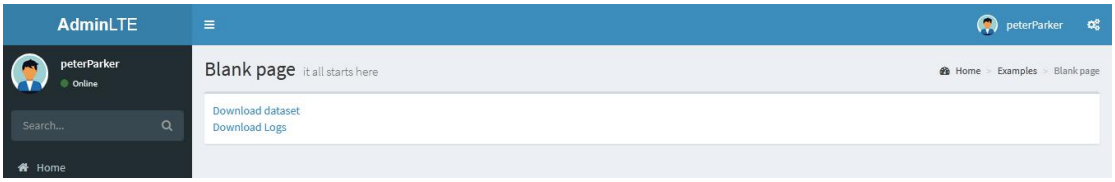


Figure 5- 27.Download file page

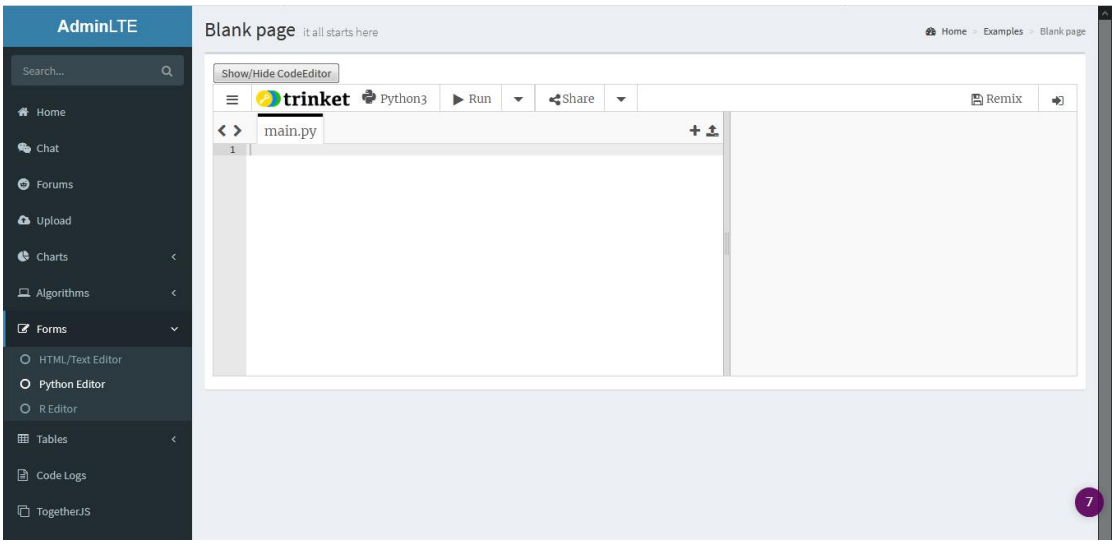


Figure 5- 28.Python editor

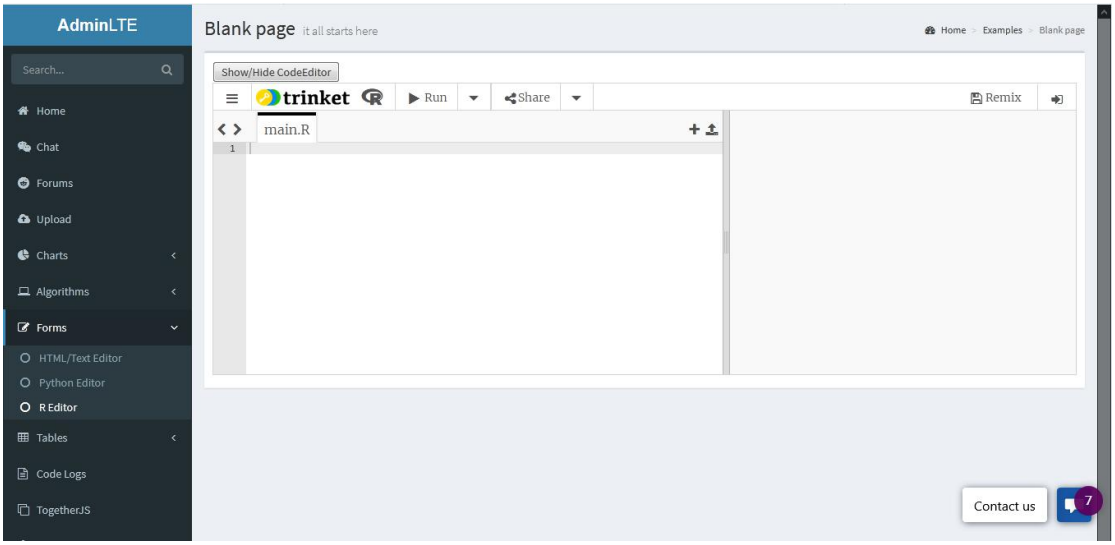


Figure 5- 29.R editor

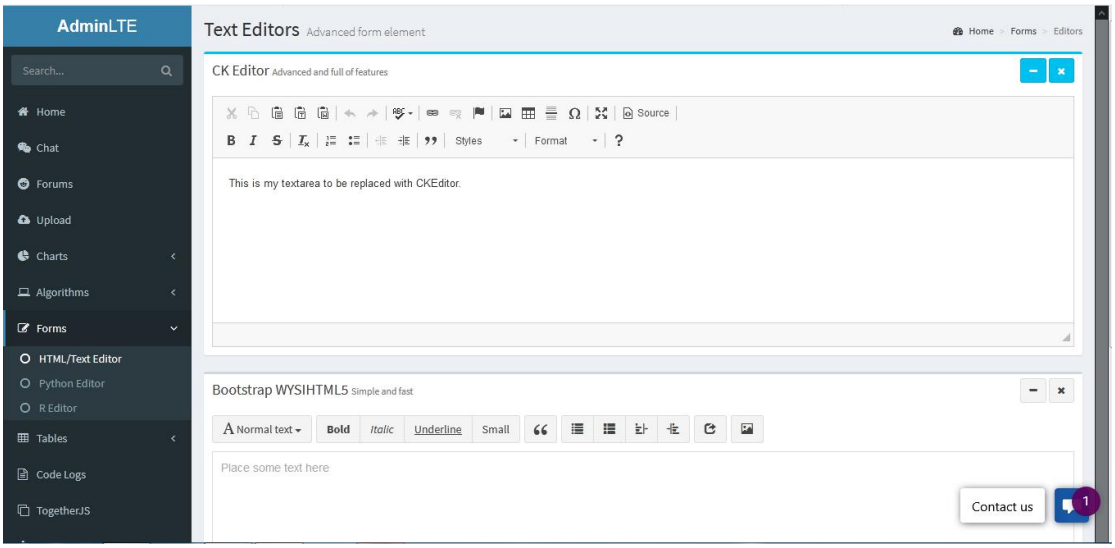


Figure 5- 30.Text editor

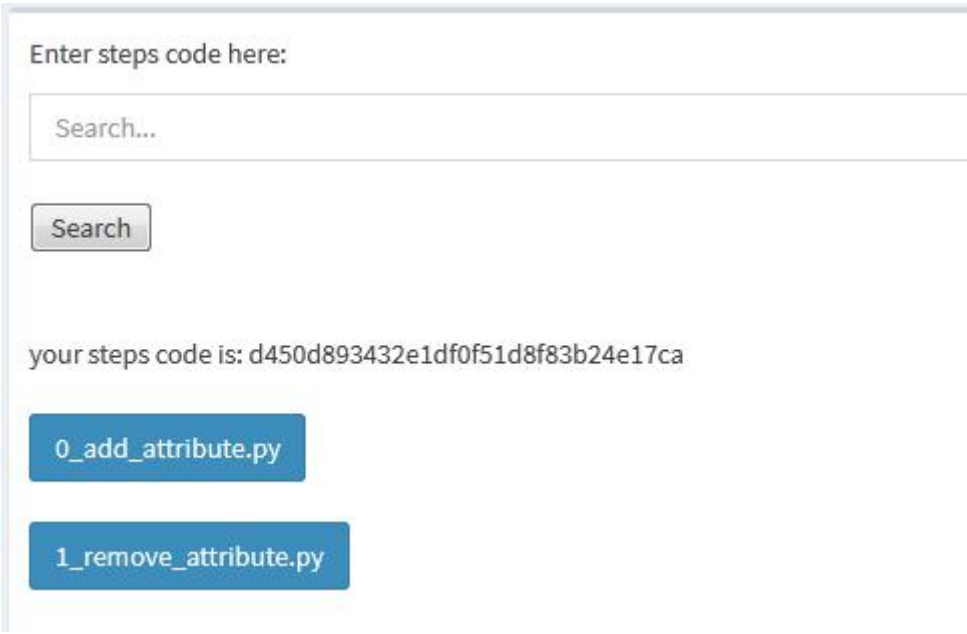


Figure 5- 31.Code steps

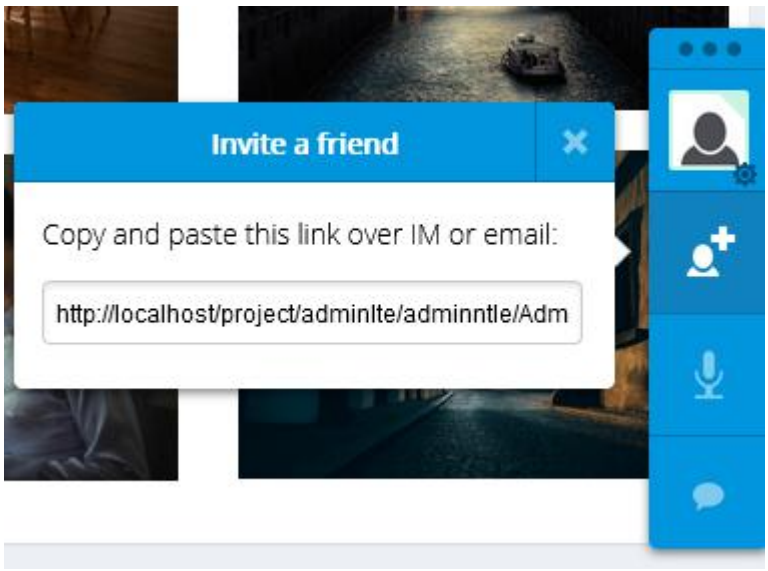


Figure 5- 32.Together js for session sharing

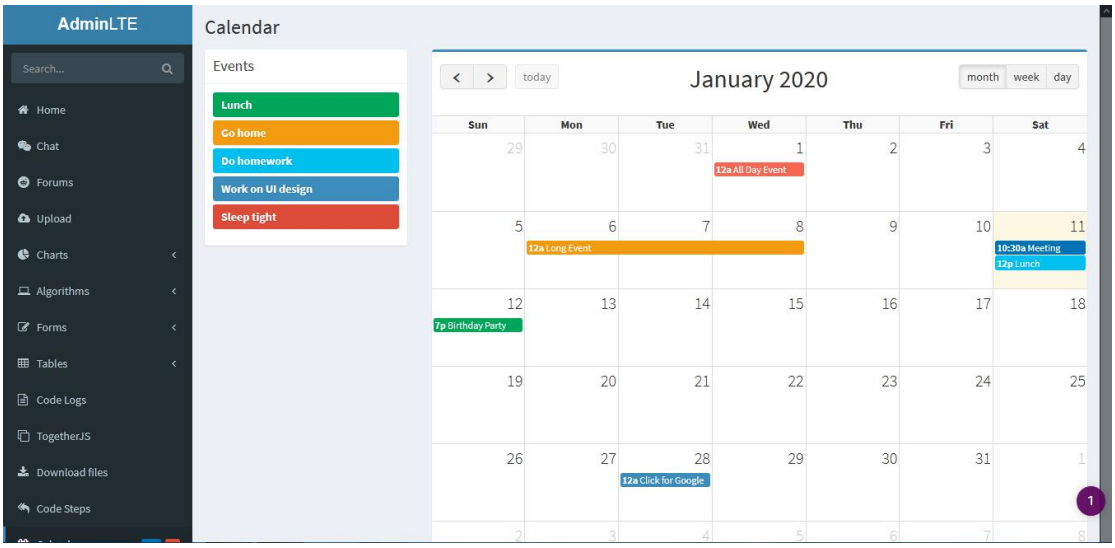


Figure 5- 33.Calendar

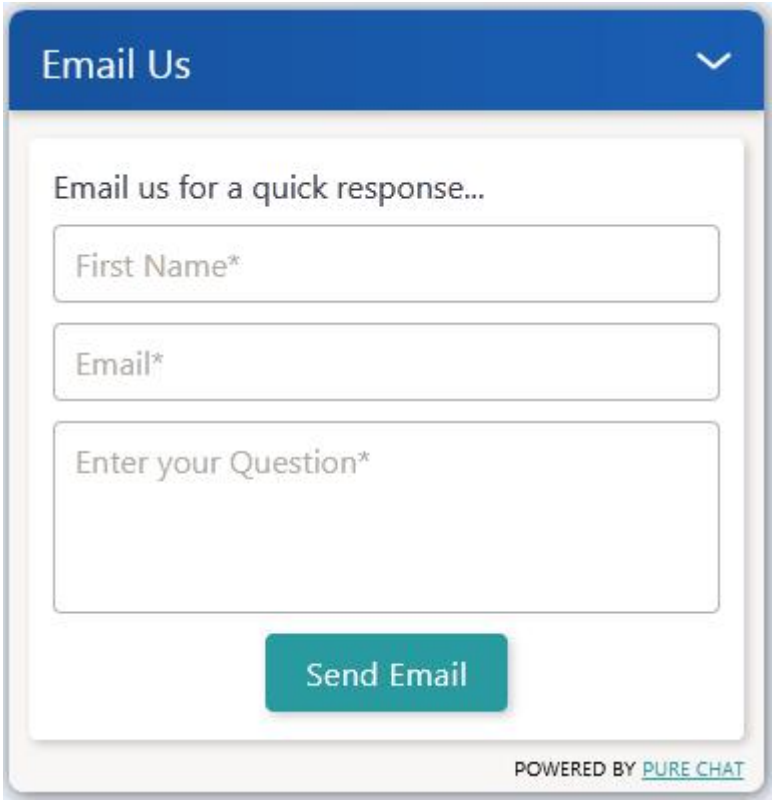
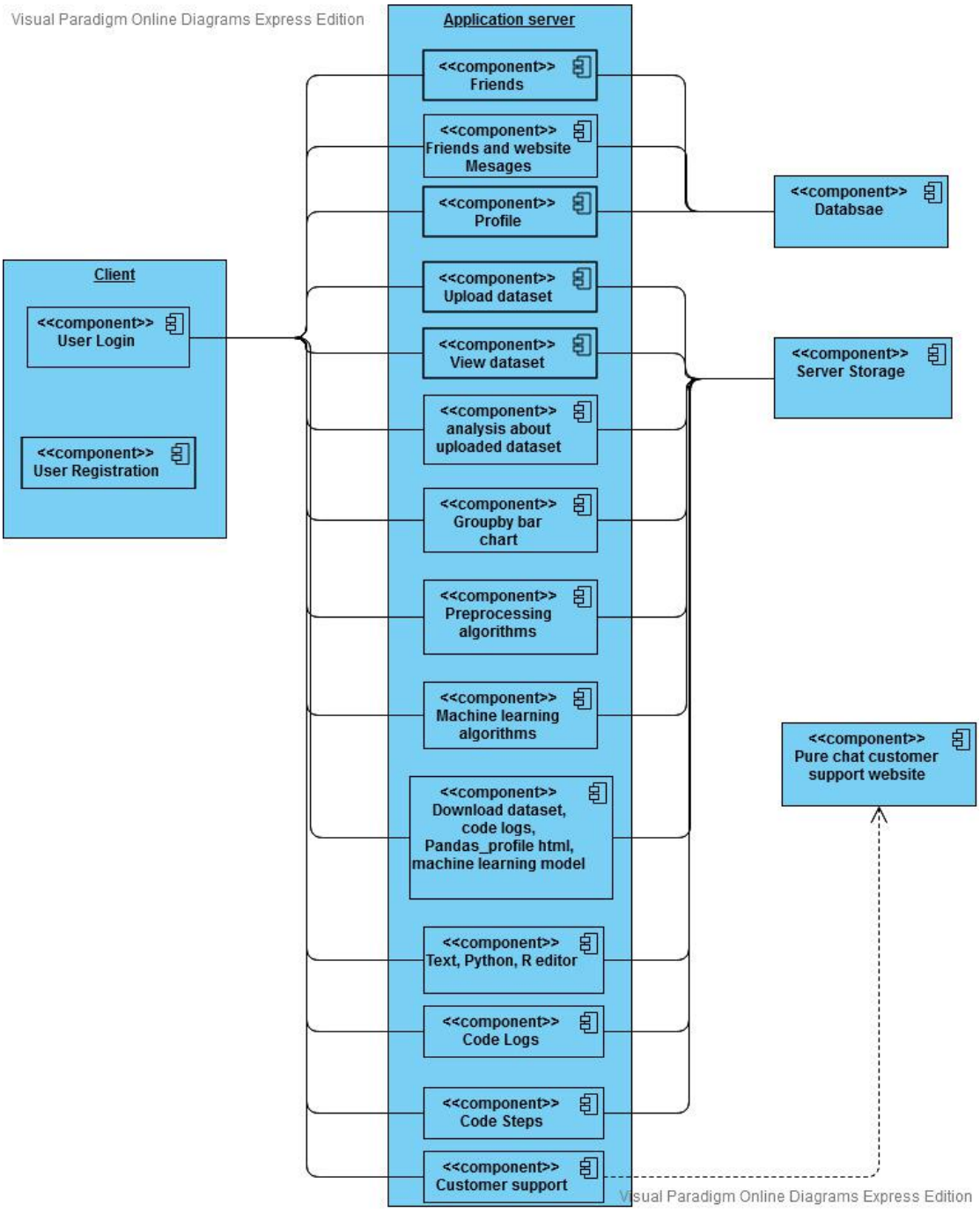


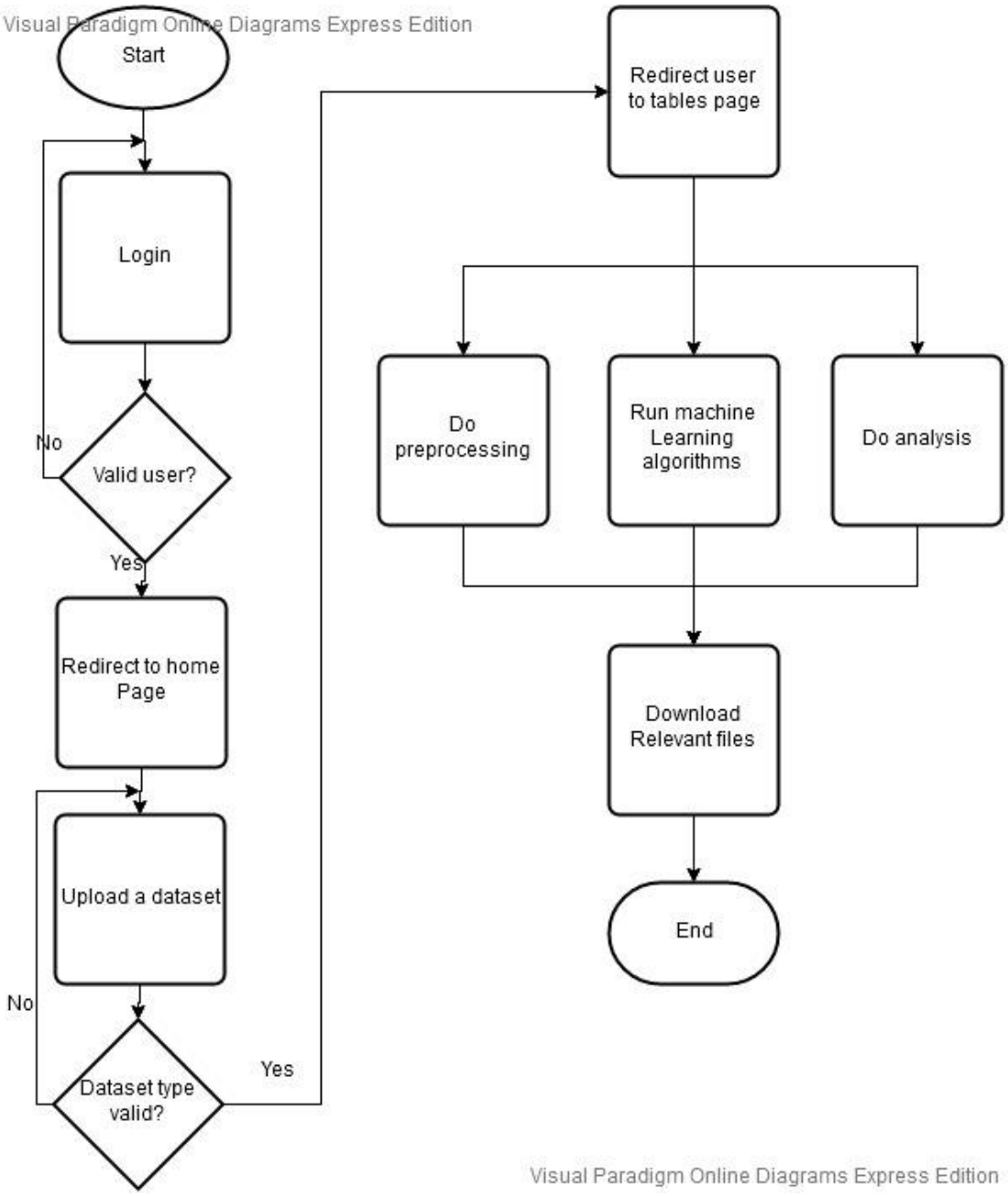
Figure 5- 34.Customer support

Chapter 6: Internal Design

6.1 Deployment diagram



6.2 Main Flow Diagram


Visual Paradigm Online Diagrams Express Edition
Start
Login
Valid user?
No
Yes
Redirect to home Page
Upload a dataset
Dataset type valid?
No
Yes
Redirect user to tables page
Do preprocessing
Run machine Learning algorithms
Do analysis
Download Relevant files
End
Visual Paradigm Online Diagrams Express Edition

6.3 Use case Diagrams

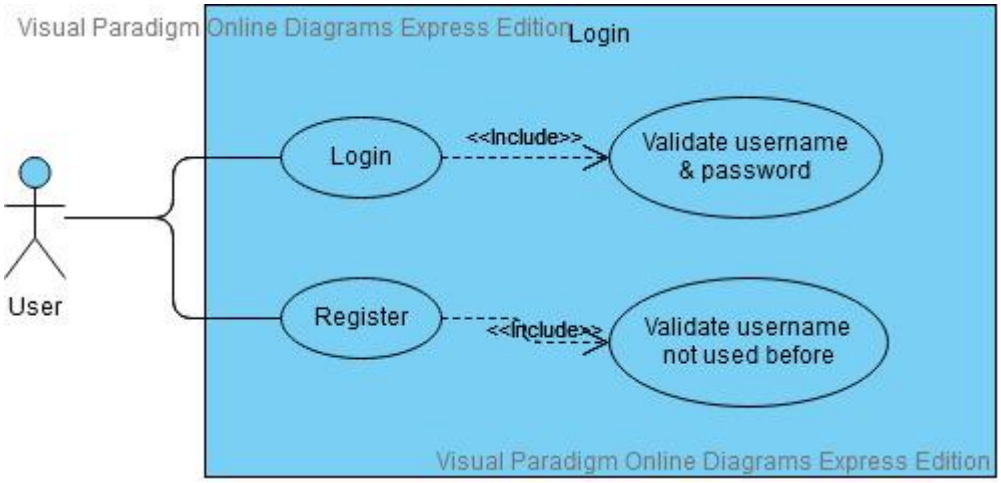


Figure 6- 1.Login/Register

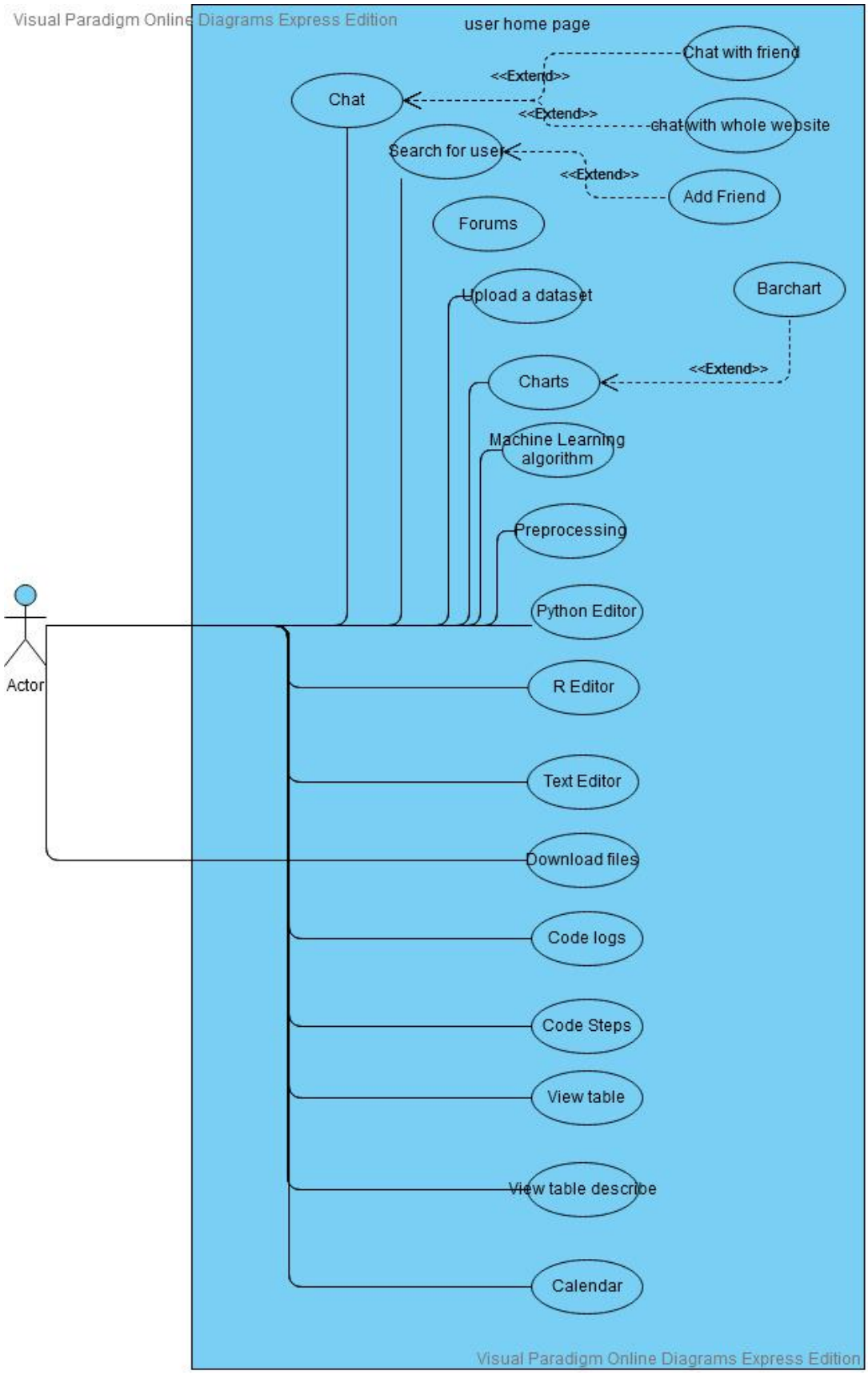


Figure 6- 2.User home options

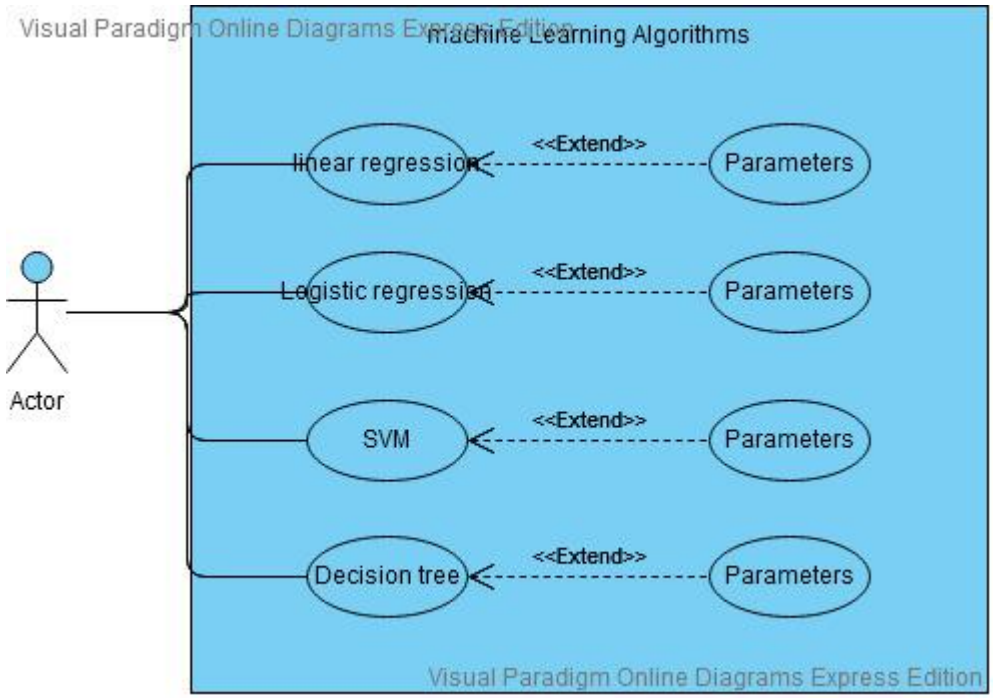


Figure 6- 3.Machine Learning Algorithms

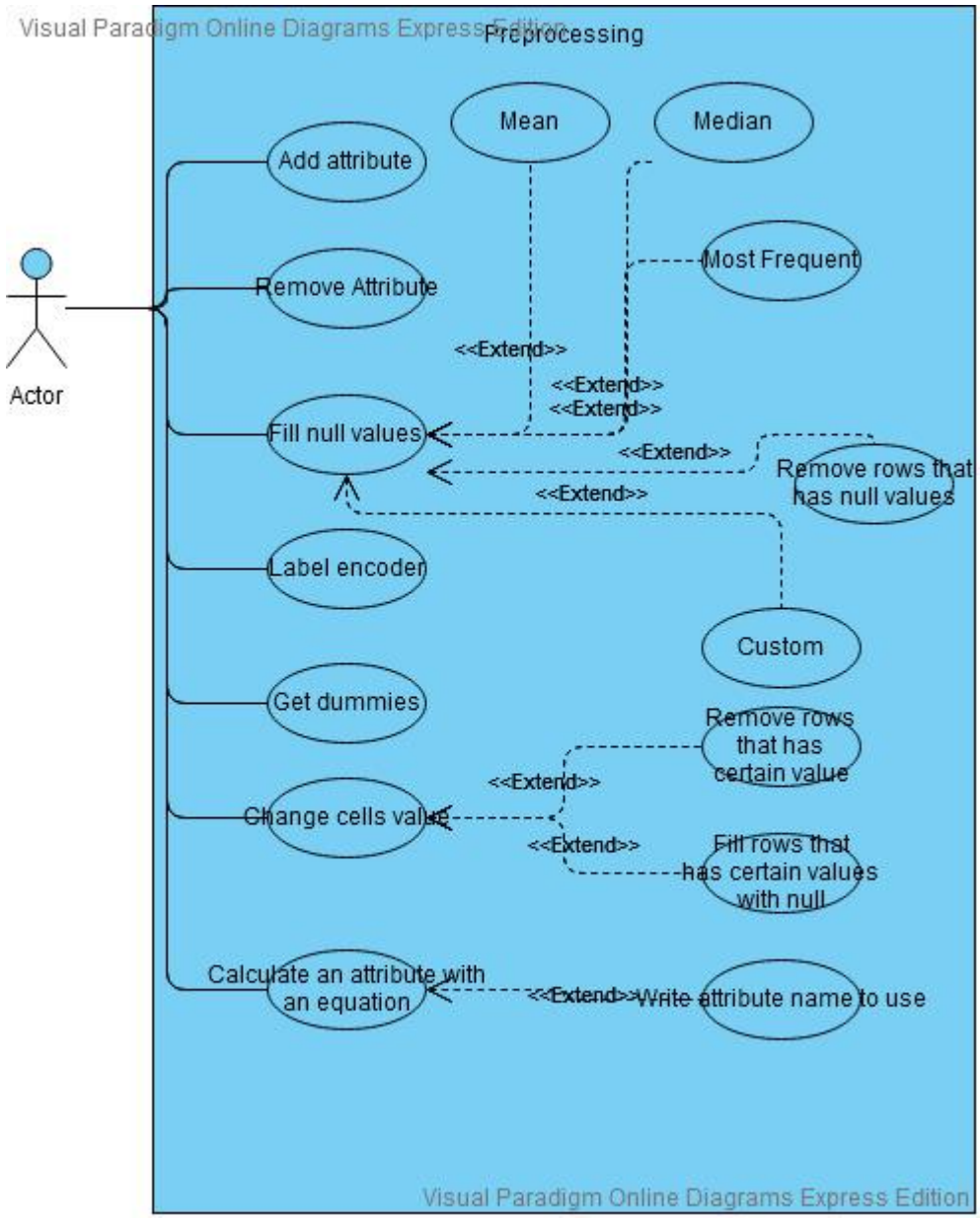


Figure 6- 4.Preprocessing

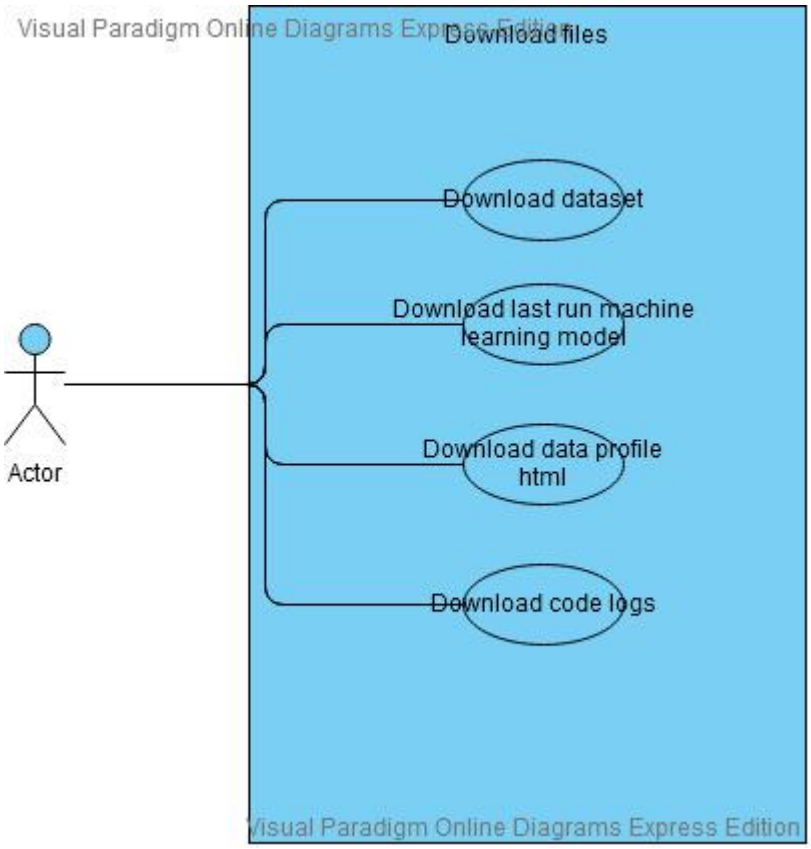


Figure 6- 5.Download Files

6.4 Sequence Diagram

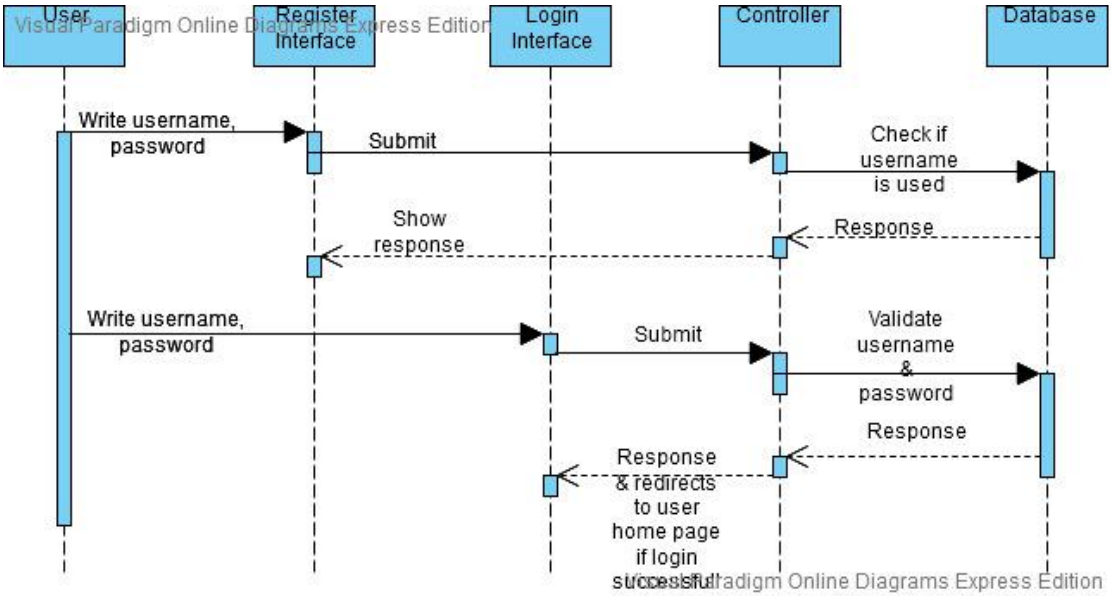


Figure 6- 6.Login/Register

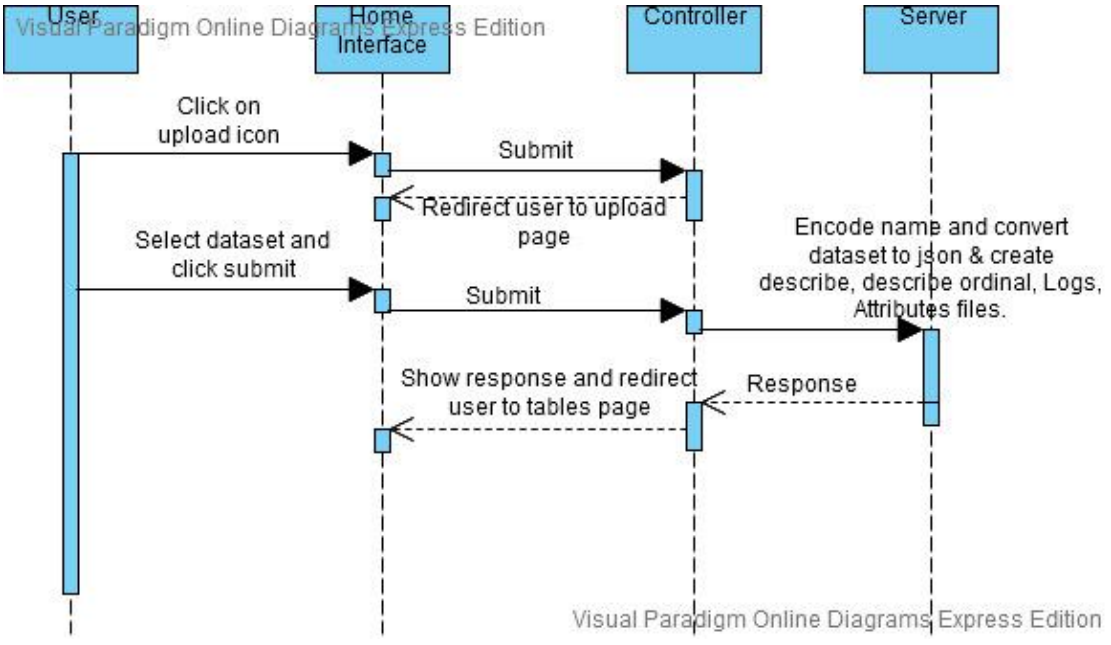


Figure 6- 7.Upload dataset

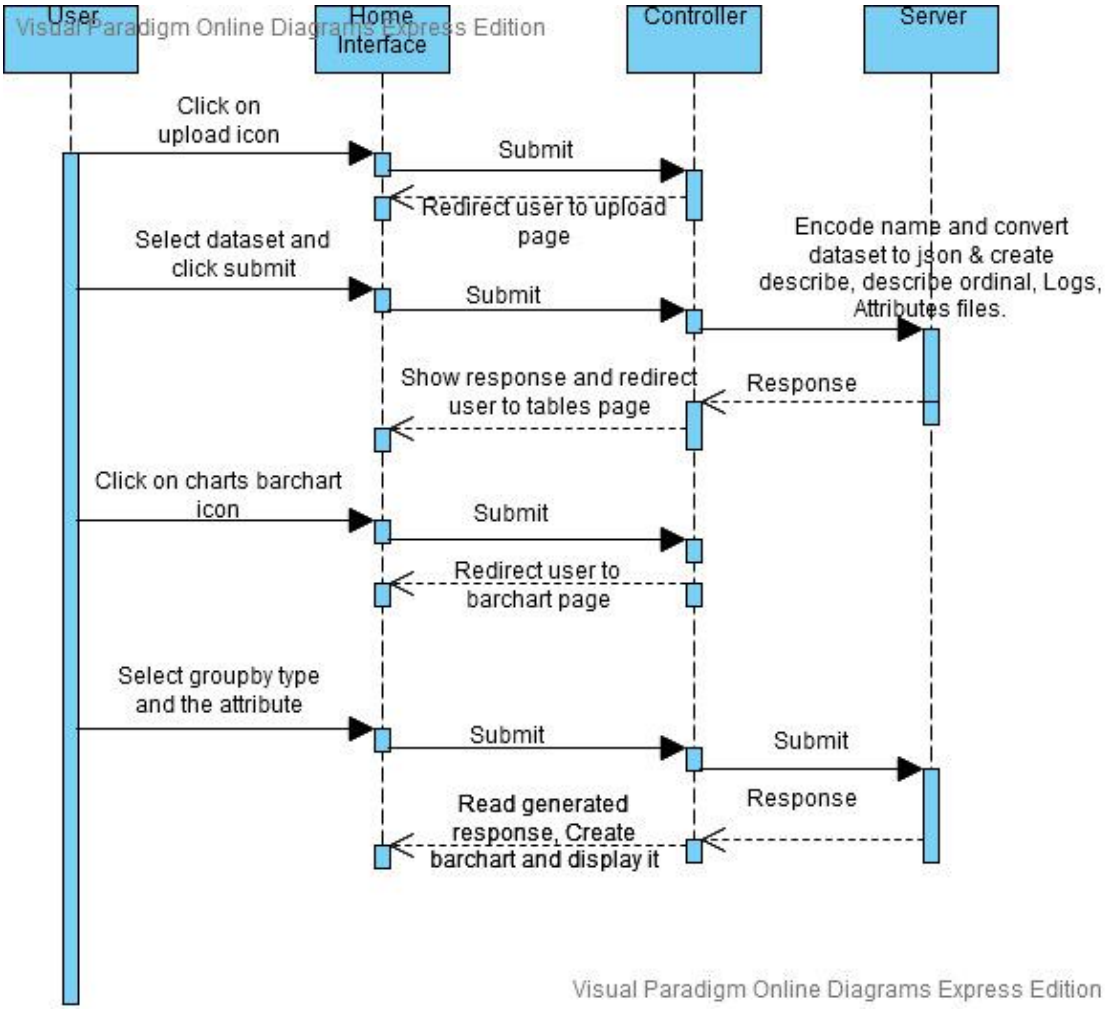


Figure 6- 8.Group by represented in bar chart

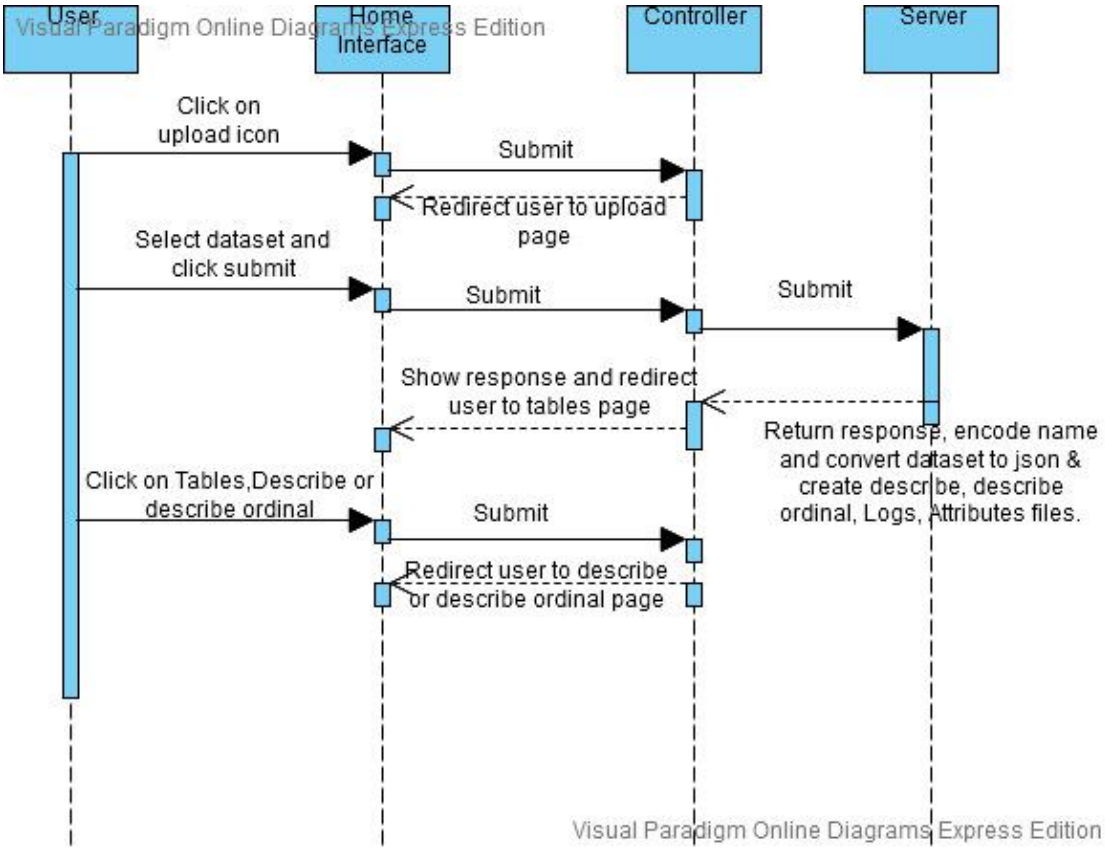


Figure 6- 9.Describe/Describe ordinal

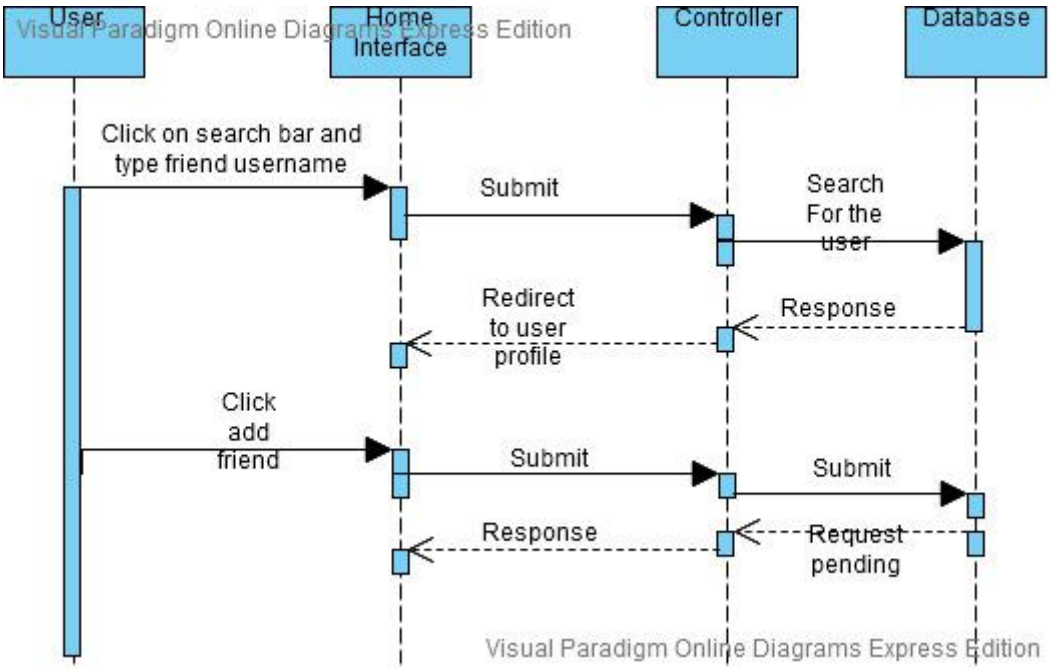


Figure 6- 10.Search user & add user

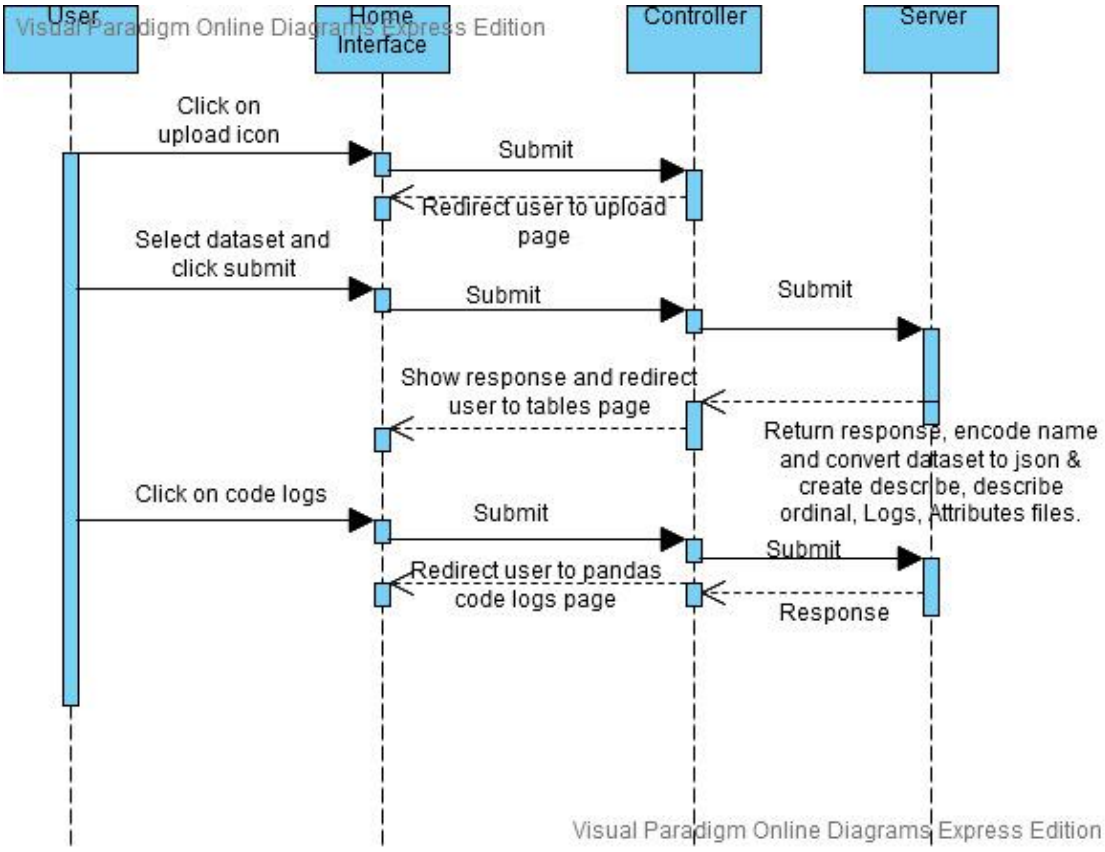


Figure 6- 11.Code Logs

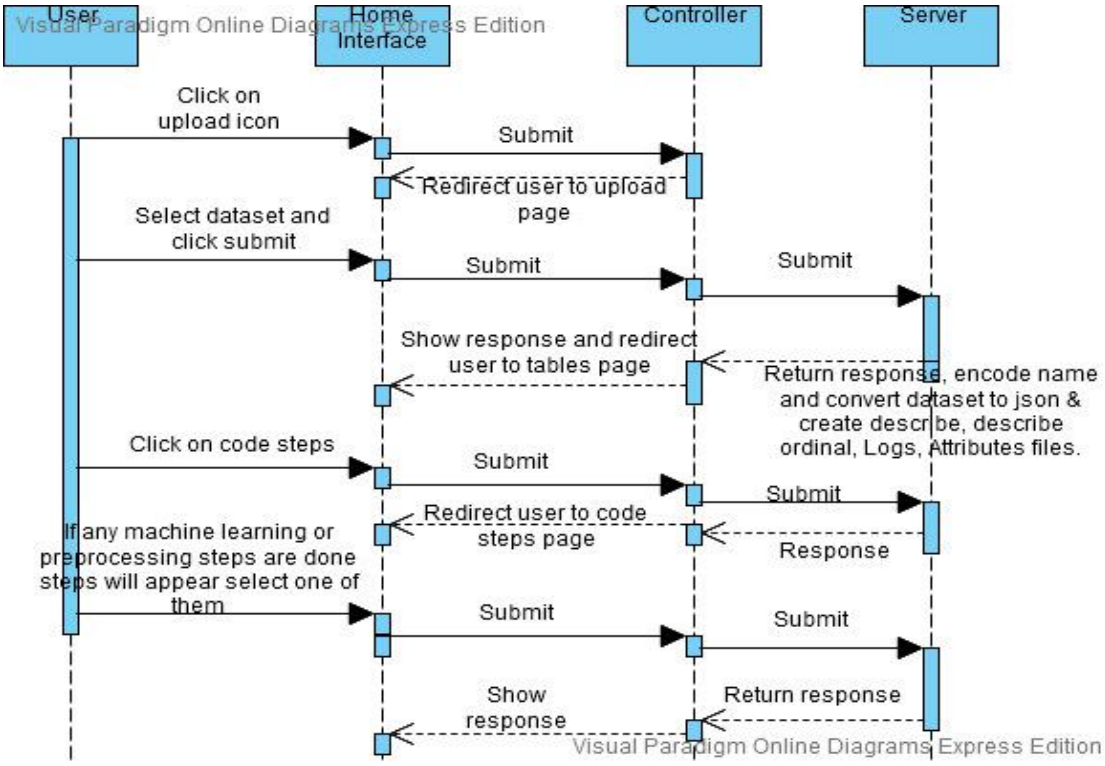


Figure 6- 12.Code Steps

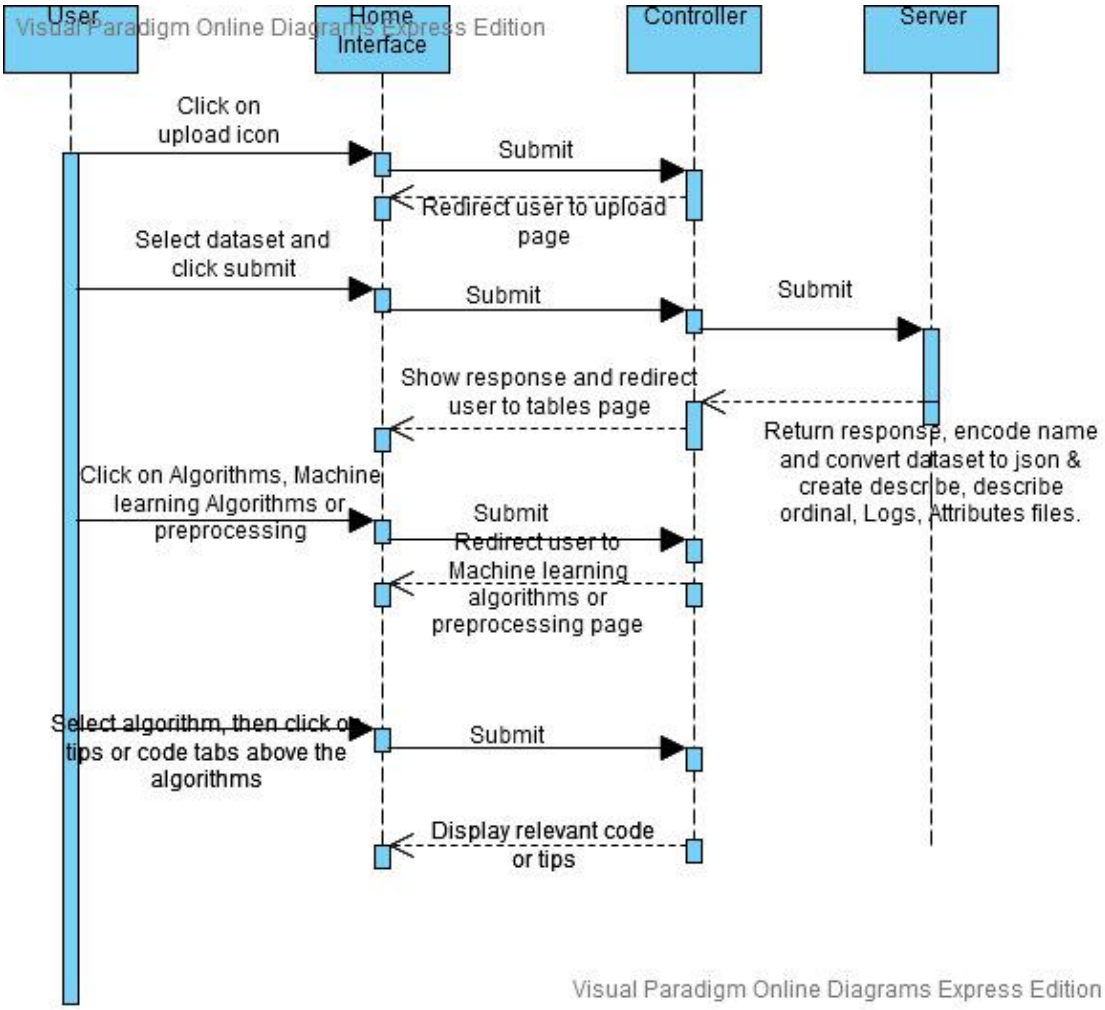


Figure 6- 13.Algorithm code, Tips and links

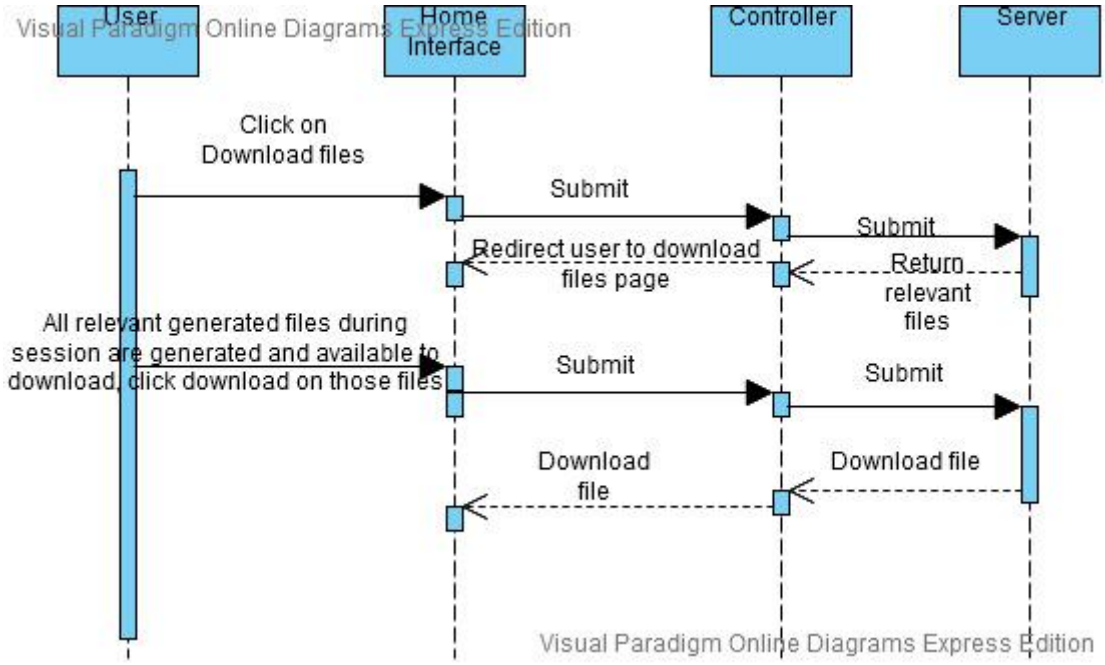


Figure 6- 14.Download Files

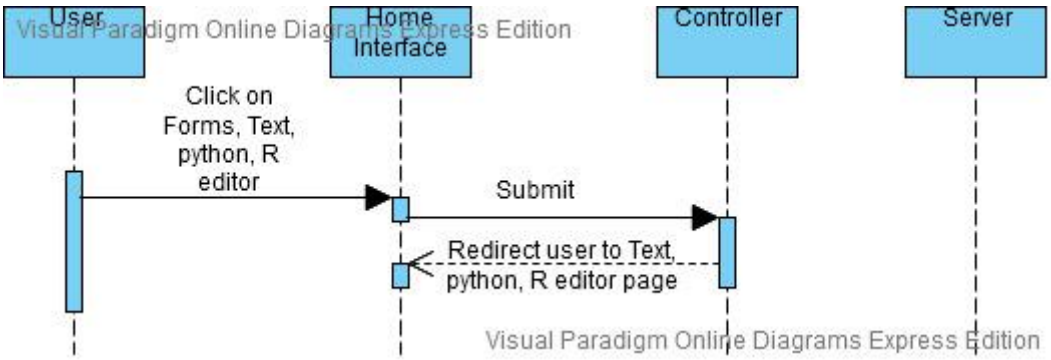


Figure 6- 15.Python, R, Text editors

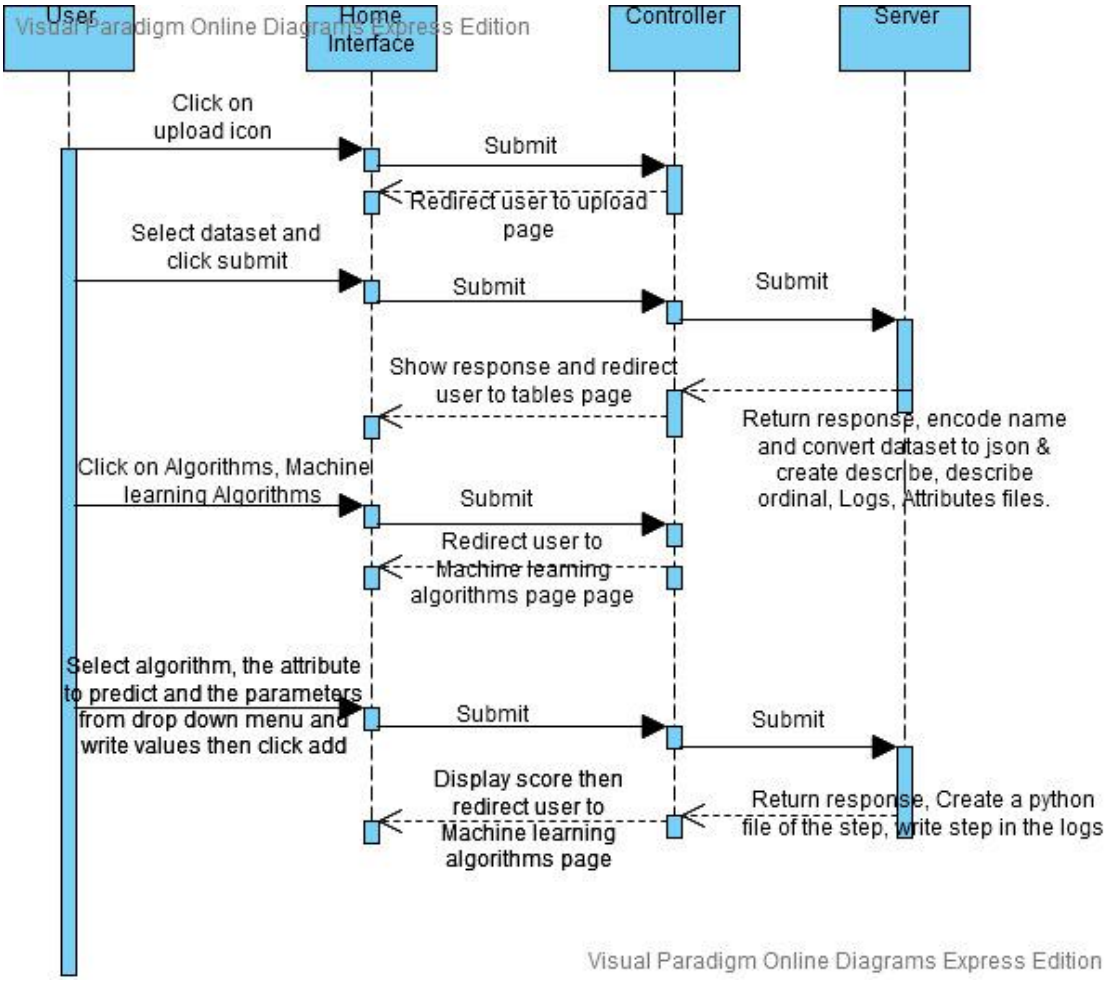


Figure 6- 16.Machine learning Algorithms

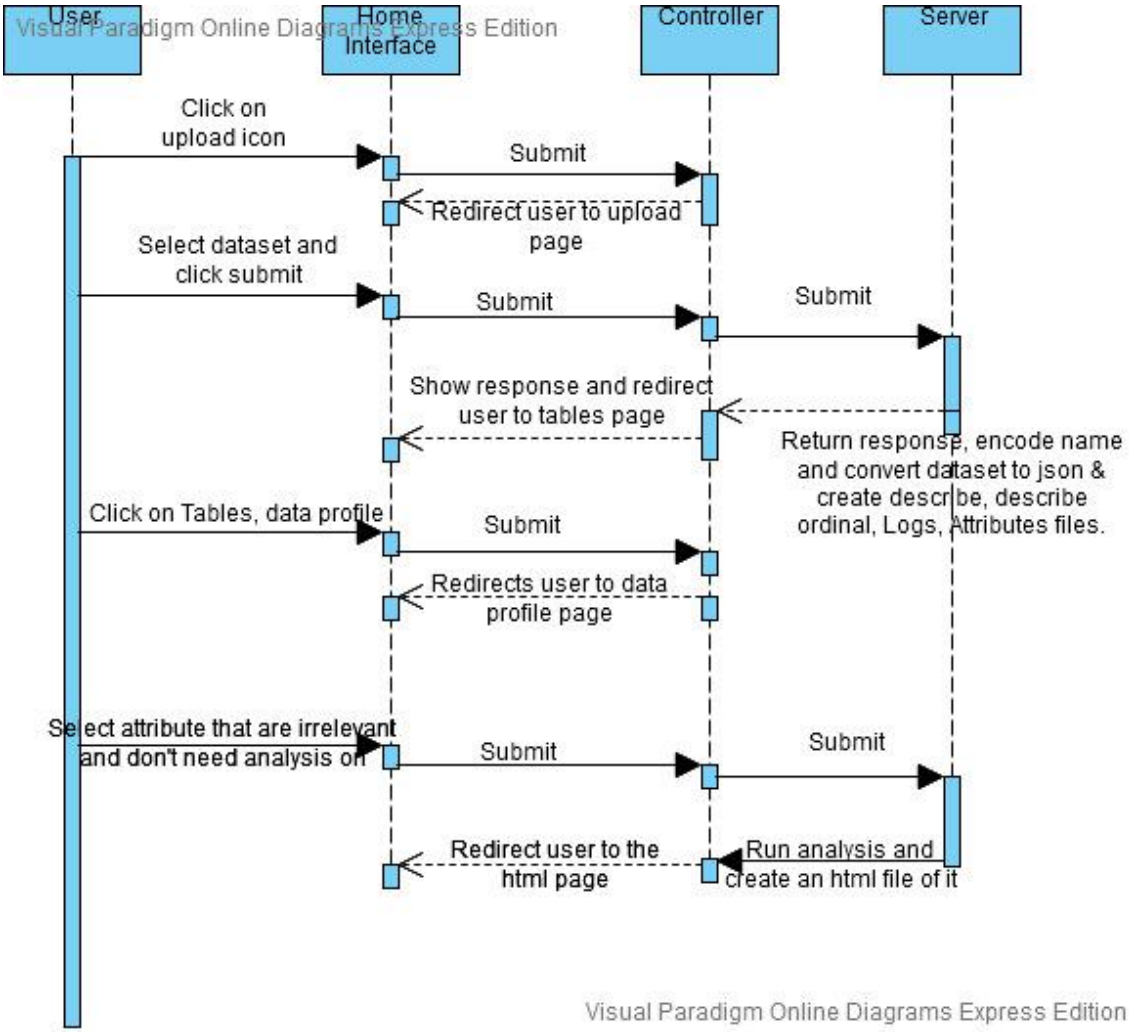


Figure 6- 17.Pandas Profilling

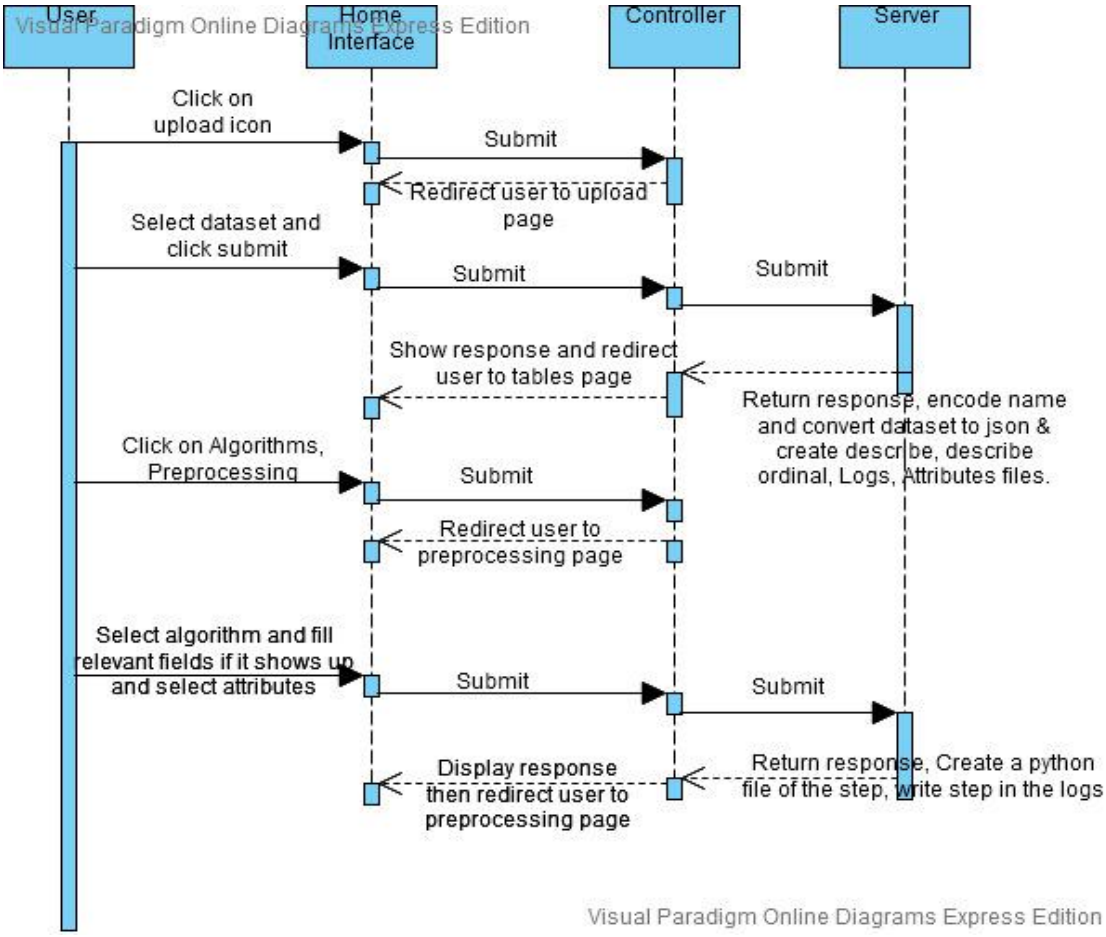


Figure 6- 18.Preprocessing

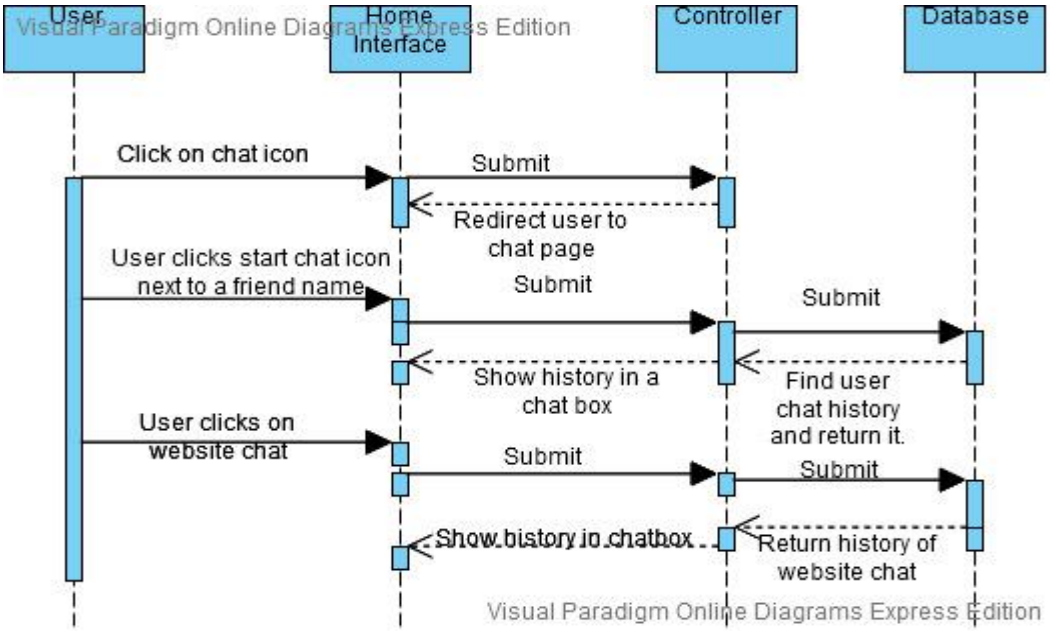
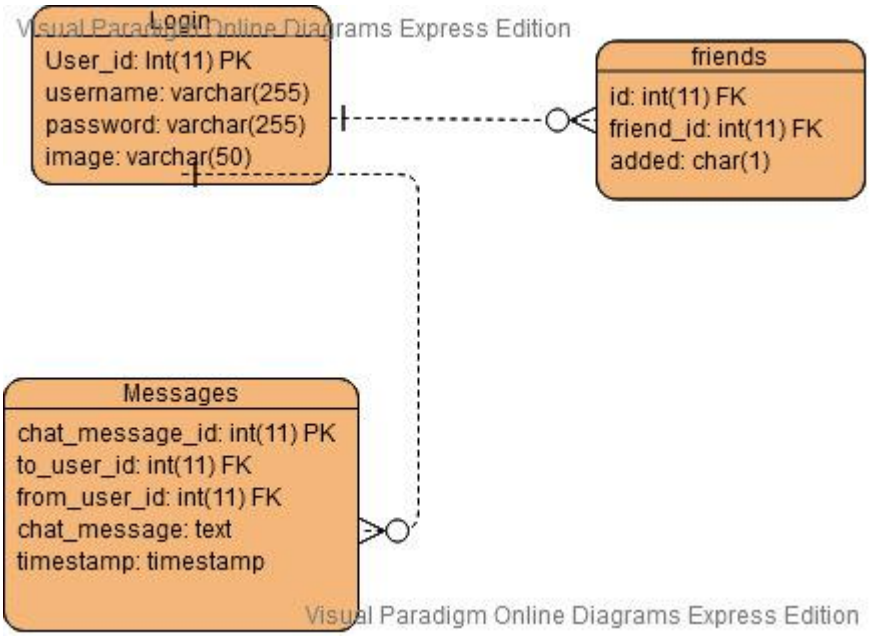


Figure 6- 19.Whole website/user to user chat

6.5 Database

6.5.1 **ERD Diagram**



	user_id	username	password	image
Edit Copy Delete	1	johnsmith	$2y$10$4REfvTZpxLgkAR/IKG9QiOkSdahOYIR3MeoGJAyiWmR...	default.jpg
Edit Copy Delete	2	peterParker	$2y$10$4REfvTZpxLgkAR/IKG9QiOkSdahOYIR3MeoGJAyiWmR...	default.jpg
Edit Copy Delete	3	davidMoore	$2y$10$4REfvTZpxLgkAR/IKG9QiOkSdahOYIR3MeoGJAyiWmR...	default.jpg
Edit Copy Delete	4	peterPark	$2y$10$T6nWApdtV/ZVrDcnmLaWXu4/tgb3CvUZNaqf4vx7BxZ...	default.jpg

Accounts

id	friend_id	added
2	1	y
1	2	y
1	3	y
3	1	y
3	2	n
2	2	n

Friends table

	login_details_id	user_id	last_activity	is_type
Edit Copy Delete	1	1	2019-10-14 01:46:48	no
Edit Copy Delete	2	2	2019-10-14 01:46:48	no
Edit Copy Delete	3	2	2019-10-14 08:51:58	no
Edit Copy Delete	4	2	2019-10-14 09:01:44	no
Edit Copy Delete	5	2	2019-10-14 13:28:08	no
Edit Copy Delete	6	1	2019-10-14 13:28:05	no
Edit Copy Delete	7	2	2019-10-14 13:30:26	no
Edit Copy Delete	8	2	2019-10-14 13:38:30	no
Edit Copy Delete	9	1	2019-10-14 13:38:29	no
Edit Copy Delete	10	3	2019-10-14 13:38:31	no
Edit Copy Delete	11	2	2019-10-14 13:40:01	no
Edit Copy Delete	12	1	2019-10-14 13:39:59	no
Edit Copy Delete	13	2	2019-10-14 13:47:15	no
Edit Copy Delete	14	1	2019-10-14 13:47:17	no
Edit Copy Delete	15	2	2019-10-18 20:46:15	no
Edit Copy Delete	16	1	2019-10-18 20:46:05	no
Edit Copy Delete	17	2	2019-10-21 12:54:50	no

You can also edit most values by double-clicking directly on them.

Login details table

	chat_message_id	to_user_id	from_user_id	chat_message	timestamp	status
Edit Copy Delete	1	0	1	hi 	2019-10-14 01:46:00	1
Edit Copy Delete	2	0	2	hello	2019-10-14 01:46:34	1
Edit Copy Delete	3	0	2	hello	2019-10-14 13:26:41	1
Edit Copy Delete	4	0	2	ok!	2019-10-14 13:26:51	1
Edit Copy Delete	5	2	1	hi	2019-10-14 13:31:45	0
Edit Copy Delete	6	1	2	hey man how are you?	2019-10-14 13:31:58	0
Edit Copy Delete	7	0	2	okaaay!	2019-10-14 13:32:22	2
Edit Copy Delete	8	0	1	okaaay!	2019-10-14 13:32:33	1
Edit Copy Delete	9	0	3	nice 	2019-10-14 13:35:00	1
Edit Copy Delete	10	0	3	no	2019-10-14 13:35:06	1
Edit Copy Delete	11	1	2	hello	2019-10-18 20:44:56	0
Edit Copy Delete	12	1	2	j	2019-12-29 15:48:02	1
Edit Copy Delete	13	1	2	no	2020-01-11 02:13:13	1
Edit Copy Delete	14	0	2	yo	2020-01-11 02:13:24	1

Chat message table

Chapter 7: Testing

7.1 Software testing

System software testing is the process of evaluating a system and its components to check whether it functions as it was intended or not.

Testing is process of executing the system and its functionality in order to find any errors, faults, any missing requirements or if the requirements with misunderstood and needs to be modified.

7.1.1 Testing process

The building and testing process of this system was done using bottom up development process.

7.1.2 Unit testing

All the machine learning algorithms were developed and tested on a where the data path and the attributes were hard coded to test if the algorithm does its required functions.

7.1.3 **Integration testing**

Table 7- 1. Register

Use case ID	Register
Use case name	Register
Description	This use case allows the user to make an account in the system to access the functions relevant to the website user role. To register the users have to insert their username, password and re-enter password the click register.
Primary actor	Website user
Precondition	None
Post condition	The system displays a message saying “Registration completed”
Main flow	1. User opens the website. 2. User clicks Sign Up on the top right corner. 3. User enters username, password and re-enters the password and clicks register. 4. System displays a message saying “Registration completed”. 5. Use case ends.
Alternative flaws	**Missing username** 1. The system displays a message saying “Username is required” 2. Use case resumes at main flow step 3 3. Use case ends. **Missing password** 1. The system displays a message saying “Password is required” 2. Use case resumes at main flow step 3 3. Use case ends.

	Missing Re-enter password or passwords don't match 1. The system displays a message saying "Password not match" 2. Use case resumes at main flow step 3 3. Use case ends. **Username is already being used** 1. The system displays a message saying "Username already taken" 2. Use case resumes at main flow step 3 3. Use case ends.

Table 7- 2.Login

Use case ID	Login
Use case name	Login
Description	This use case allows the user to make an account in the system to access the functions relevant to the website user role. To login the users have to insert their username, password and click login.
Primary actor	Website user
Precondition	1. Registered an account
Post condition	The system displays the home page
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page 5. Use case ends.
Alternative flaws	**Missing username or password** 1. The system prompts user to enter username or password 2. Use case resumes at main flow step 3

	3. Use case ends. **Wrong password** 1. The display a message saying “Wrong password” 2. Use case resumes at main flow step 3 3. Use case ends. **Wrong username** 1. The display a message saying “Wrong username” 2. Use case resumes at main flow step 3 3. Use case ends.

Table 7- 3.Search_profile

Use case ID	Search_profile
Use case name	Search_profile
Description	This use case allows the user to search for an account in the system to see the relevant user information. To search for users the user has to enter username of the user in the bar at top left corner after login then click on the search icon.
Primary actor	Website user
Precondition	Logged in with an account
Post condition	The system displays the user page
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page. 5. User enter the username he wants to search and clicks on search icon.

	6. System shows the user profile. 7. Use case ends.
Alternative flaws	**Wrong username** 1. The system shows a message saying “Wrong Username”. 2. The system redirects user to homepage. 3. Use case ends.

Table 7- 4.Add_user

Use case ID	Add_user
Use case name	Add_user
Description	This use case allows the user to search for an account in the system to see the relevant user information. To search for users the user has to enter username of the user in the bar at top left corner after login then click on the search icon, after profile appears users clicks add user.
Primary actor	Website user
Precondition	1. Logged in with an account
Post condition	The system displays a message saying “Request sent” or “Friend added “ if that person sent a request.
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page. 5. User enter the username he wants to search and clicks on search icon. 6. System shows the user profile. 7. User clicks add user icon. 8. System displays a message saying “Request sent” or “Friend added “ if that person sent a request.

	9. Use case ends.
Alternative flaws	**Wrong username** 1. The system shows a message saying “Wrong Username”. 2. The system redirects user to homepage. 3. Use case ends.

Table 7- 5.Chat_with_friend

Use case ID	Chat_with_friend
Use case name	Chat_with_friend
Description	This use case allows the user to chat with an another user in the system. To chat with friend the user has to click chat icon on the side bar after login, click on start chat next to friend name, a chat box appears user sends text to his friend.
Primary actor	Website user
Precondition	1. Logged in with an account 2. Has a friend added on friend list
Post condition	The system sends a message without any error.
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page. 5. User clicks chat icon on side bar 6. System opens chat window 7. User clicks on start chat next to his friends name. 8. System opens chat box 9. User types a message and clicks send 10. Message is sent 11. Use case ends.
Alternative flaws	

Table 7- 6.Chat_with_entire_website

Use case ID	Chat_with_entire_website
Use case name	Chat_with_entire_website
Description	This use case allows the user to chat with all the other users in the system. To chat with friend the user has to click chat icon on the side bar after login, click on website chat, a chat box appears user sends text to his friend.
Primary actor	Website user
Precondition	1. Logged in with an account.
Post condition	The system sends a message without any error.
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page. 5. User clicks chat icon on side bar 6. System opens chat window 7. User clicks on website chat. 8. System opens chat box. 9. User types a message and clicks send 10. Message is sent 11. Use case ends.
Alternative flaws	

Table 7- 7.Upload_file

Use case ID	Upload_file
Use case name	Upload_file

Description	This use case allows the user to upload structured dataset to the system. To upload the file the user has to click upload icon on the side bar after login, click on upload, then upload the dataset.
Primary actor	Website user
Precondition	Logged in with an account.
Post condition	The system displays a message saying “File uploaded successfully” then redirects user to tables page.
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page. 5. User clicks upload icon on side bar and uploads a file. 6. System displays a message saying “File uploaded successfully” and redirects user to tables page. 7. Use case ends.
Alternative flaws	**Uploaded a file with the extension not supported** 1. System displays a message saying “We don’t support that type of data” and send user to tables page. 2. Use case ends.

Table 7- 8.Add_attribute

Use case ID	Pre-processing_1
Use case name	Add_attribute
Description	This use case allows the user to add an attribute to structured dataset uploaded to the system. To add an attribute the user should click on pre-processing icon on the side bar after login, click on add column, write the name of the attribute and click submit.

Primary actor	Website user
Precondition	1. Logged in with an account. 2. Uploaded a dataset.
Post condition	The system displays a message saying “Attribute has been added” then redirects user to pre-processing page.
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page. 5. User clicks upload icon on side bar and uploads a file. 6. System displays a message saying “File uploaded successfully” and redirects user to tables page. 7. User clicks on algorithms, Pre-processing on side bar. 8. System shows Pre-processing page. 9. User click on add column and writes column name then clicks submit 10. System displays a message saying “Attribute has been added” then redirects user to pre-processing page. 11. Use case ends.
Alternative flaws	**User does not enter an attribute name** 1. System displays a message saying “Please write an attribute name” and redirects user to pre-processing page. 2. Use case ends.

Table 7- 9.Remove_attribute

Use case ID	Pre-processing_2
Use case name	Remove_attribute
Description	This use case allows the user to remove an attribute from structured dataset uploaded to the system. To remove an attribute the user should click on pre-processing icon on the side bar after login, click on remove column, select attributes from checkbox on

	the right then click submit.
Primary actor	Website user
Precondition	1. Logged in with an account. 2. Uploaded a dataset.
Post condition	The system displays a message saying “Attribute has been removed” then redirects user to pre-processing page.
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page. 5. User clicks upload icon on side bar and uploads a file. 6. System displays a message saying “File uploaded successfully” and redirects user to tables page. 7. User clicks on algorithms, Pre-processing on side bar. 8. System shows Pre-processing page. 9. User click on remove column and selects attribute from check box list on the right then clicks submit. 10. System displays a message saying “Attribute has been removed” then redirects user to pre-processing page. 11. Use case ends.
Alternative flaws	**User does not select an algorithm or attribute** **1.** System displays a message saying “Please choose command and attributes” and redirects user to pre-processing page. **2.** Use case ends.

Table 7- 10.Fill_nan

Use case ID	Pre-processing_3
Use case name	Fill_nan
Description	This use case allows the user to fill null values in an attribute from structured dataset uploaded to the system. To fill null values in an

	attribute the user should click on pre-processing icon on the side bar after login, click on Fill Nan, choose to fill null values from the drop down menu by either (mean, median, most frequent, remove rows that contain null values, custom value) select attributes from checkbox on the right then click submit.
Primary actor	Website user
Precondition	1. Logged in with an account. 2. Uploaded a dataset.
Post condition	The system displays a message saying “Null values has been filled” then redirects user to pre-processing page.
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page. 5. User clicks upload icon on side bar and uploads a file. 6. System displays a message saying “File uploaded successfully” and redirects user to tables page. 7. User clicks on algorithms, Pre-processing on side bar. 8. System shows Pre-processing page. 9. User clicks on Fill Nan, choose to fill null values from the drop down menu by either (mean, median, most frequent, remove rows that contain null values, custom value) ‘if custom values is selected the user types in the new value’ and selects attribute from check box list on the right then clicks submit. 10. System displays a message saying “Null values has been filled” then redirects user to pre-processing page. 11. Use case ends.
Alternative flaws	**User does not select an algorithm or attribute** 1. System displays a message saying “Please choose command and attributes” and redirects user to pre-processing page. 2. Use case ends **Custom value is empty** 1. System displays a message saying “please enter a value” then

	redirects user to pre-processing page. 2. Use case ends **Any unpredictable error** 3. System displays a message saying “Unexpected error: .. the reason of the error..” then redirects user to pre-processing page. 4. Use case ends

Table 7- 11.Label_encoder

Use case ID	Pre-processing_4
Use case name	Label_encoder
Description	This use case allows the user to label encode an attribute from structured dataset uploaded to the system. To label encode an attribute the user should click on pre-processing icon on the side bar after login, click on Label Encoder, select attributes from checkbox on the right then click submit.
Primary actor	Website user
Precondition	1. Logged in with an account. 2. Uploaded a dataset.
Post condition	The system displays a message saying “Attribute is now label encoded” then redirects user to pre-processing page.
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page. 5. User clicks upload icon on side bar and uploads a file. 6. System displays a message saying “File uploaded successfully” and redirects user to tables page.

	7. User clicks on algorithms, Pre-processing on side bar. 8. System shows Pre-processing page. 9. User click on Label Encoder the selects the attributes from check box list on the right then clicks submit. 10. System displays a message saying “Attribute has been Label Encoded” then redirects user to pre-processing page. 11. Use case ends.
Alternative flaws	**User does not select an algorithm or attribute** 1. System displays a message saying “Please choose command and attributes” and redirects user to pre-processing page. 2. Use case ends **Any unpredictable error** **1.** System displays a message saying “Unexpected error: .. the reason of the error..” then redirects user to pre-processing page. **2.** Use case ends

Table 7- 12.Get_dummies

Use case ID	Pre-processing_5
Use case name	Get_dummies
Description	This use case allows the user to one hot encode an attribute from structured dataset uploaded to the system. To one hot encode an attribute the user should click on pre-processing icon on the side bar after login, click on Get dummies, select attributes from checkbox on the right then click submit.
Primary actor	Website user
Precondition	1. Logged in with an account. 2. Uploaded a dataset.

Post condition	The system displays a message saying “Dummies has been created” then redirects user to pre-processing page.
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page. 5. User clicks upload icon on side bar and uploads a file. 6. System displays a message saying “File uploaded successfully” and redirects user to tables page. 7. User clicks on algorithms, Pre-processing on side bar. 8. System shows Pre-processing page. 9. User click on Get dummies then selects the attributes from check box list on the right then clicks submit. 10. System displays a message saying “Dummies has been created” then redirects user to pre-processing page. 11. Use case ends.
Alternative flaws	**User does not select an algorithm or attribute** 1. System displays a message saying “Please choose command and attributes” and redirects user to pre-processing page. 2. Use case ends **Any unpredictable error** **1.** System displays a message saying “Unexpected error: .. the reason of the error..” then redirects user to pre-processing page. **2.** Use case ends

Table 7- 13.Replace_cells_values

Use case ID	Pre-processing_5
Use case name	Replace_cells_values
Description	This use case allows the user to change cells value that has a certain values within an attribute from structured dataset

	uploaded to the system. To replace a cell value the user should click on pre-processing icon on the side bar after login, click on replace cell value, write the old and new values, select attributes from checkbox on the right then click submit.
Primary actor	Website user
Precondition	1. Logged in with an account. 2. Uploaded a dataset.
Post condition	The system displays a message saying “cells value has been replaced” then redirects user to pre-processing page.
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page. 5. User clicks upload icon on side bar and uploads a file. 6. System displays a message saying “File uploaded successfully” and redirects user to tables page. 7. User clicks on algorithms, Pre-processing on side bar. 8. System shows Pre-processing page. 9. User clicks on replace cells value,write the old and new cells value, the selects attributes from check box list on the right then clicks submit. 10. System displays a message saying “Cells value has been replaced” then redirects user to pre-processing page. 11. Use case ends.
Alternative flaws	**User does not select an algorithm or attribute** 1. System displays a message saying “Please choose command and attributes” and redirects user to pre-processing page. 2. Use case ends **Any unpredictable error** **1.** System displays a message saying “Unexpected error: .. the reason of the error..” then redirects user to pre-processing page. **2.** Use case ends

Table 7- 14.Add_equation_to_calculate_an_attribute

Use case ID	Pre-processing_6
Use case name	Add_equation_to_calculate_an_attribute
Description	This use case allows the user to make an equation and write the result in an attribute from structured dataset uploaded to the system. To calculate an equation the user should click on pre-processing icon on the side bar after login, click on replace cell value, write the equation, select attributes from checkbox on the right then click submit.
Primary actor	Website user
Precondition	1. Logged in with an account. 2. Uploaded a dataset.
Post condition	The system displays a message saying “equation is done” then redirects user to pre-processing page.
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page. 5. User clicks upload icon on side bar and uploads a file. 6. System displays a message saying “File uploaded successfully” and redirects user to tables page. 7. User clicks on algorithms, Pre-processing on side bar. 8. System shows Pre-processing page. 9. User clicks on add an equation to calculate an attribute,write the equation, the selects attributes from check box list on the right then clicks submit. 10. System displays a message saying “Equation is done” then redirects user to pre-processing page. 11. Use case ends.
Alternative flaws	**User does not select an algorithm or attribute**

	1. System displays a message saying “Please choose command and attributes” and redirects user to pre-processing page. 2. Use case ends **Any unpredictable error** **1.** System displays a message saying “Unexpected error: .. the reason of the error..” then redirects user to pre-processing page. **2.** Use case ends

Table 7- 15.Groupby_in_bar_chart

Use case ID	Charts_1
Use case name	Groupby_in_bar_chart
Description	This use case allows the user to groupby an attribute and present it in a bar chart from structured dataset uploaded to the system. To Groupby the attribute the user should click on bar chart icon after login, select the attribute which he will groupby and type of group by then clicks submit.
Primary actor	Website user
Precondition	1. Logged in with an account. 2. Uploaded a dataset.
Post condition	The system displays a message saying “Groupby is done” then redirects user to bar_chart page.
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page. 5. User clicks upload icon on side bar and uploads a file. 6. System displays a message saying “File uploaded successfully”

	and redirects user to tables page. 7. User clicks on charts, bar chart on side bar. 8. System shows bar chart page. 9. User selects an attribute then groupby type then clicks submit. 10. System displays a message saying “Equation is done” then redirects user to pre-processing page. 11. Use case ends.
Alternative flaws	**User does not select an algorithm or attribute** 1. System displays a message saying “Please choose groupby type and attributes” and redirects user to bar chart page. 2. Use case ends **Any unpredictable error** **1.** System displays a message saying “Unexpected error: .. the reason of the error..” then redirects user to pre-processing page. **2.** Use case ends

Table 7- 16.Machine_learning_algorithms

Use case ID	Machine_learning_algorithms
Use case name	Linear_regression,Decision_Tree,SVM,Logistic_regression
Description	This use case allows the user to use machine learning algorithms to predict an attribute from structured dataset uploaded to the system. To predict the attribute the user should click on Algorithms, machine learning algorithms icon after login, select the attribute which he will predict and type of algorithms and add any parameters if he wanted by selecting parameter from drop down menu then writing the value then clicks submit.

Primary actor	Website user
Precondition	1. Logged in with an account. 2. Uploaded a dataset. 3. Dataset ready to use machine learning algorithm on
Post condition	The system displays a message showing the score then redirecting user to scripts page.
Main flow	1. User opens the website. 2. User clicks Login on the top right corner. 3. User enters username, password and then clicks login. 4. System displays the home page. 5. User clicks upload icon on side bar and uploads a file. 6. System displays a message saying “File uploaded successfully” and redirects user to tables page. 7. User clicks on Algorithms, machine learning algorithms on side bar. 8. System shows machine learning algorithms page. 9. User selects an algorithm and optional (selects an attribute from drop down menu and then writes it value then clicks add) then selects attribute he wants to predict then clicks submit. 10. System displays a message showing the score of the algorithm then redirects user to machine learning algorithm page. 11. Use case ends.
Alternative flaws	**User does not select an algorithm or attribute** 1. System displays a message saying “Please select an algorithm and attribute to predict” and redirects user to machine learning algorithm page. 2. Use case ends **Any unpredictable error** **1.** System displays a message saying “Unexpected error: .. the reason of the error..” then redirects user to machine learning algorithm page. **2.** Use case ends

Chapter 8: Conclusion

Mengen is a new platform for new and experienced users that provides major services in machine learning, the most of machine learning users learn about machine learning and interact with other users through many links and there is not a single place where a user can learn all about machine learning, Mengen aims to be similar to W3Schools in machine learning where users can learn all about machine learning on a single platform which makes learning process much easier, also it aims to build a social and active environment by making a forum page on reddit and adding chat option to whole website and user to user so sharing information be much easier, The platform also allows to hold collaborative sessions where the user can hold a session and have a voice chat with other users while sharing content of his screen. And to facilitate the process the user can try and write code on the website compiler without needing to download anything on his pc.

Chapter 9: Future Work

This project has much potential as it is everlasting development whether keeping up with the new algorithms or updating the new ones, or adding additional functionality so the user can optimize his data more.

9.1 Adding containers

A container will be created for each time a user logs in, the containers will be added as a form of security as it protects our system and other users from security attacks.

9.2 Adding more Preprocessing functionalities

The amount of preprocessing functionalities at the moment is nowhere satisfactory to produce a good model, adding more functionalities will search much greater purpose and will result in much better model score.

9.3 Adding more machine learning algorithms

The current amount of machine learning algorithms is very low and it doesn’t include unstructured machine learning algorithms or neural networks which

are both doable on this website. Adding more algorithms serves the purpose of user having greater variety at choosing and implementing a machine learning model

9.4 Implementing a payment plan

The process of running machine learning model is very power consuming and renting a database is also expensive, Payment plan will help in keeping the project running while keeping a free trail with limits for new users and people who want to learn only.

9.5 Making a model to predict best steps

After many users use the website the plan is it make a machine learning model that helps the user by telling him best steps to use so he can get best result out of his model.

9.6 Adding unsupervised, NLP algorithms and ability to build neural networks

After So far the website covers only supervised algorithms which is a very small portion of the data provided being able to implement unsupervised and NLP will be a huge step to make people learn more. It will also provide feature to be able to build neural networks.

9.7 Using D3.js library

This library includes over 130 types of charts which can be used in future to do complex visualisations.

9.8 Using Qualas

This library was just implemented in apache spark new version which allows user to utilize several machines using the code same as pandas library, this will allow website to be implemented on several machines.

9.9 Implementing a wizard

A wizard will be implemented to guide beginner level users through the process to be able to utilize the algorithms and make them learn.

9.10 Implementing dataflow/graphical representation.

Instead of selecting algorithms users can go to graphical interface, drag and drop box that has algorithms names in it and connect them with a line, then press run. This will make all algorithms run in succession and show end result.

9.11 Adding persistent.

Persistent storage will be added to users who has a payment plan.

Chapter 10: References

[1] Afsar, J. (2019). *Univariate, Bivariate And Multivariate Data | Explanation | Engineering Intro*. [online] Engineering Intro. Available at: http://www.engineeringintro.com/statistics/introduction-statistics/univariate-bivariate-and-multivariate-data/ [Accessed 7 Jun. 2019].

[2] Santarcangelo, J. (2019). *how does sklearn compute the Accuracy score step by step?*. [online] Stack Overflow. Available at: https://stackoverflow.com/questions/37665680/how-does-sklearn-compute-the-accuracy-score-step-by-step [Accessed 7 Jun. 2019].

[3] En.wikipedia.org. (2019). *Sensitivity and specificity*. [online] Available at: https://en.wikipedia.org/wiki/Sensitivity_and_specificity [Accessed 7 Jun. 2019].

[4] Afsar, J. (2019). *Univariate, Bivariate And Multivariate Data | Explanation | Engineering Intro*. [online] Engineering Intro. Available at: http://www.engineeringintro.com/statistics/introduction-statistics/univariate-bivariate-and-multivariate-data/ [Accessed 7 Jun. 2019].

[5] En.wikipedia.org. (2019). *Naive Bayes classifier*. [online] Available at: https://en.wikipedia.org/wiki/Naive_Bayes_classifier#Introduction [Accessed 7 Jun. 2019].

[6] Towards Data Science. (2019). *Naive Bayes Classifier*. [online] Available at: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c [Accessed 7 Jun. 2019].

[7] Towards Data Science. (2019). *Naive Bayes Classifier*. [online] Available at: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c [Accessed 7 Jun. 2019].

[8] DataCamp Community. (2019). *Decision Tree Classification in Python*. [online] Available at: https://www.datacamp.com/community/tutorials/decision-tree-classification-python [Accessed 7 Jun. 2019].

[9] En.wikipedia.org. (2019). *Gradient boosting*. [online] Available at: https://en.wikipedia.org/wiki/Gradient_boosting [Accessed 7 Jun. 2019].

[10] Towards Data Science. (2019). *Support Vector Machine — Introduction to Machine Learning Algorithms*. [online] Available at: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 [Accessed 7 Jun. 2019].

[11] Python.org. (2019). *What is Python? Executive Summary*. [online] Available at: https://www.python.org/doc/essays/blurb/ [Accessed 7 Jun. 2019].

[12] Cuelogic Technologies Pvt. Ltd. (2019). *Role of Python in Artificial Intelligence (AI) | Python Machine Learning*. [online] Available at:

https://www.cuelogic.com/blog/role-of-python-in-artificial-intelligence [Accessed 7 Jun. 2019].

[13] Anaconda. (2019). *What is Anaconda - Anaconda*. [online] Available at: https://www.anaconda.com/what-is-anaconda/ [Accessed 7 Jun. 2019].

[14] Jupyter.org. (2019). *Project Jupyter*. [online] Available at: https://jupyter.org/ [Accessed 7 Jun. 2019].

[15] DataCamp Community. (2019). *Pandas Tutorial: DataFrames in Python*. [online] Available at: https://www.datacamp.com/community/tutorials/pandas-tutorial-dataframe-python [Accessed 7 Jun. 2019].

[16] www.tutorialspoint.com. (2019). *NumPy Tutorial*. [online] Available at: https://www.tutorialspoint.com/numpy [Accessed 7 Jun. 2019].

[17] Stack Abuse. (2019). *Beginner's Tutorial on the Pandas Python Library*. [online] Available at: https://stackabuse.com/beginners-tutorial-on-the-pandas-python-library/ [Accessed 7 Jun. 2019].

[18] Codecademy. (2019). *What is Scikit-Learn?* | *Codecademy*. [online] Available at: https://www.codecademy.com/articles/scikit-learn [Accessed 7 Jun. 2019].

[19] Hosting Tribunal. (2020). *25 Must-Know Cloud Computing Statistics in 2020*. [online] Available at: https://hostingtribunal.com/blog/cloud-computing-statistics