ÇANKAYA UNIVERSITY FACULTY OF ENGINEERING COMPUTER ENGINEERING DEPARTMENT

Project Report

Version 1

CENG 407

Innovative System Design and Development I

<Web-based Multifunctional Graphical User Interface for Data Analysis>

Alim GEYİK
201411023
Oğulcan BAŞARAN
201411205
Büşra KUTLUER
201511040
Nurseli BAL
201611656

Advisor: *Assist. Prof. Dr. Gül TOKDEMİR*

Table of Contents

Abstract

The data that people create in the virtual world (social media posts, news, shopping bills, emails, etc.) is increasing and accumulating day by day. These data may be unnecessary and meaningless to most of us, but they are vital to an analyst, a company or a business. These data are used in the promotion of a new product, in the design of a store, in an application update or in determining the products to be discounted. Every footprint in the virtual world is truly a treasure for someone interested in the point we step on. Of course, not every data is meaningful and usable. At this point data analysis programs and platforms come into play. It is precisely these applications and platforms that will reveal the meaning of the available data and determine the value of this treasure. WADAV is a web platform that allows users to review theri data and access the information they need. Our goal in developing WADAV was to create a simple and understandable interface that would enable all types of users to evaluate the data they had. In this report, we explained why we plan to create WADAV software requirements and how we plan to design our software.

Key words:

WADAV, data, data analysis,

1. Introduction

1.1. Motivation

We are a group of senior students from the computer engineering department interested in data science. Our group members have taken the data mining course being taught at our school and are working in this project because of their interest in this field. Unlike many data analysis applications currently available on the market, we aim to interactive interface to examine data in more detail. For this purpose, we want to enable users to reach a faster experience from anywhere with WADAV platform. In order to further develop ourselves in the field of data analysis and data visualization, we enrolled in courses via the internet. We plan to complete this project by developing ourselves with the help of these courses.

1.2. Problem Statement

Our main problem in this project is that the data has many missing elements or unnecessary information. In order to achieve more consistent and understandable results, we have to clean the data for analysis by passing through the correct preprocessing stages.

1.3. Related Work

We wanted to design and build an accessible platform that would not only allow everybody to uncover the hidden predictive power of data with ease, but also would make the whole experience “enjoyable”.[1] (BigML)

Exploratory’s Simple UI experience makes it possible for anyone to use Data Science to Explore data quickly, Discover deeper insights, and Communicate effectively.[2]

1.4. Solution Statement

As a result, there will be a data analysis and data visualization system for WADAV users, which has an interactive interface, works easily and quickly, and can perform the preprocessing stages of data.

2. Literature Review

Nowadays, the use of simulations and serious games in learning and assessment is widespread. Serious games used for purposes other than mere entertainment. The starting point is the concept of the serious game itself, and what the means. Serious games are allowed learners to experience some situations that are impossible in the real world for different reasons like safety, cost, time, etc. However, they are also claimed to have positive impacts on the players' development of several different skills [3]. Although there is much theoretical support for the benefits of digital games in learning and education, there is mixed empirical support. In this report, we searched about how to improve the children's mental capabilities using games.

2.1. Introduction

Nowadays the biggest treasure is considered information. Companies, businesses, and all people want to have knowledge. About their own businesses, about their customers, about their friends, about everyone else. Social media is the primary proof of this situation. At this point, if we need to address this issue, it is necessary to organize all this information chaos. Data analysis and visualization methods facilitate our work in this field. They enable us to have more readable, predictable, computable and processable data.

In this project, we are trying to analyze and visualize our data in line with our needs with an interactive interface. We will design a website that contains many different analysis methods, tables, and graphs. Our main goal is to provide the user with a user-friendly interface in the fastest way possible. In this process, we will examine many different analysis tools and applications and try to reach the most stable and reliable results. We will try to present the diversity of analysis and visualization at the highest level possible.

2.2. Data Analysis

Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusion and supporting decision-making. [4] Data analysis consists of several steps. If we follow

these steps in order and correctly, we will have an analysis process in which we can get the most accurate results. These steps are determining the requirements, data collection, data cleaning, data analysis, data interpretation. And after completing these steps correctly, we will continue with data visualization.

First, we need to determine why we need data analysis. For example, do we own a hotel business where our customers will have a better experience? Do we want to turn our hotel into a more sophisticated complex? If we ask feedback from our customers after determining our aim, we may have collected the necessary data for the roadmap to follow. For our example, we can ask our customers ideas about room comfort, cleanliness, hotel facilities, etc.

Immediately after this stage, we have to clean the data set. Our data set may have incomplete data, or if we don't have the data set ourselves, it may have data that we don't need. The cleaning phase is one of the most important steps for data analysis because incomplete or unnecessary information can remove us from our purpose. After collecting, clearing and processing the data, we are ready for analysis.

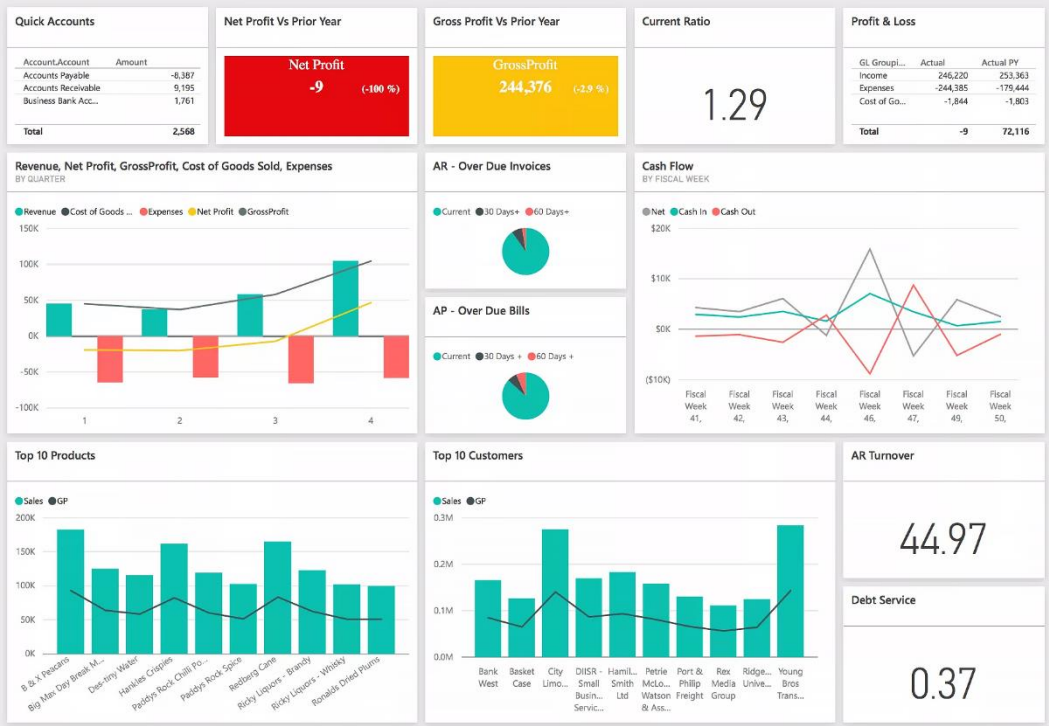
After all these steps, we may have the information we need or we may need to collect more data. At this stage, we can use data analysis tools and software to help us understand, interpret, and achieve results based on requirements.[5] And it's time to interpret our results. We can choose the way to express or transmit our data analysis result and come to the end of a successful analysis process.

2.3. Data Visualization

Data visualization is a graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.[6] Data visualization is very common in your day to day life; they often appear in the form of charts and graphs. In other words, data shown graphically so that it will be easier for the human brain to understand and process it. We can use data visualization to put our data, which we have completed the latest data analysis process, into a form that is understandable for everyone. In this way, our results can be better examined and evaluated and the most accurate roadmap can be determined.

If we talk about a few simple data visualization methods, these scatter plots, line graphs, pie charts, and bar graphs. Line Charts are generally used to depict a trend of data over time. One of the most ubiquitous charts in use is the pizza chart or the pie chart. It is a visualization element for comparing the percentages of a kind of data. However, there are several disadvantages. We need many labels to identify and

understand chart elements. The leaner and much more elegant alternative to the pie chart is the bar chart, which is able to express all that the pie says and much more, without making a mess of labels and legends. And finally Word-cloud. With the advent of social media and the different platforms where people can write out their feelings, reviews and opinion, one data visualization tool that has become quite common is the word-cloud. Word clouds help us find out what people are talking about.[7]



2.4. Data Analysis and Visualization Tools

Today, data analysis and visualization tools are really popular. Many companies and businesses (Google, Facebook, Instagram, Apple, etc.) use these tools. In our literature study, we examined a few of these which are popular. We have examined web-based systems such as BigML, Kaggle and desktop programs such as Exploratory and Orange. These applications and tools help us to analyze and visualize the data we have. They include many visualization methods and analysis methods. They support the programming languages (Python, R, JavaScript, etc.) that we need during analysis and visualization, helping us to get more detailed results. BigML is a machine-based data analysis and visualization web-interface. You can upload your own data as well as open-source data sets. It has many methods of data visualization while facilitating data

cleaning, editing, and analysis. Kaggle is a website that allows Google's data analysts and data scientists to review and write articles. It contains many data sets and analyzes. Supports many programming languages. It also creates discussion environments for data analysts, such as a blog page. Exp is a desktop application developed for data analysis and visualization. They cooperate with many large companies to provide support. They have R language support and SQL support for dataset editing.

2.5. Conclusion

The general concept of data analysis and visualization is to bring a library layout to this dispersed pool of information, to reveal key points and necessary information. In this growing data complex, companies and individual employees find it hard to find their way. Our Project serves them as a GPS in this data chaos. It's a system that shows users how to access the actual information they need and search through all that chaos.

3. Software Requirements Specification

3.1. Introduction

This document is the software requirement specification document for the Web-based Multifunctional Graphical User Interface for Data Analysis and Visualization-WMGUIDAV Project, briefly Web App Data Analysis and Visualization - WADAV. This system allows users to analyze their data without using any desktop application.

3.1.1. Purpose

The purpose of this document is to outline requirements for the WADAV. This document describes the general concept and explains the required features of the system. It describes the interfaces, user characteristics, functionalities, and constraints of the system to make the specifications confirmed and to refer to the development phase of the system.

3.1.2. Scope of Project

In this project, we aimed to present statistical and variable information units in a different and creative way by isolating them from classical schematic forms. One of

the most important goals of data visualization is to make complex and confusing data normally presented in the classic format easy to understand with easy-to-understand graphical interfaces. We needed to present the data in a comprehensible way and get the right design in line with the overall objectives of the project. The content of this design covers many fields such as tools, IDEs, protocols, back-end, front-end, etc. parts used in the project.

Firstly, we have decided to use Phyton as the language. Python is a powerful, dynamic programming language used in a wide range of fields. The Pandas library offers high-performance, easy-to-use data structures, and data analysis features. Thus, without the need for statistical programs such as R or Stata, we can do data analysis and modeling. Combined with tools and libraries such as IPython, statemodels, and sci-kit-learn, a powerful data analysis environment can be achieved in terms of performance and productivity.

Secondly, the protocol can be used as Web Distributed Authoring and Versioning (WebDAV). İs a hypertext transfer (HTTP) protocol that allows you to manage files and documents stored on web servers.

With WebDAV, you can connect to the cloud environment like Google Drive, OneDrive, OwnCloud, and do a lot of things like deleting, writing, copying, moving, and you can do it using your internet browser without installing any applications.

As a storage system, we have thought of a platform where large amounts of data can be securely held, quick access to information, the integrity of information, and access to multiple users simultaneously.

Instead of stacking all the data in one place, we thought of using MySQL, which stores it in different tables and databases. The control, optimization, and repair of the tables are done quickly.

Finally, we thought of using HTML5 and CSS3 to visualize the project (website). It has advantages such as working together using the same database common, making it easy to use for everyone.

3.1.3. Overview

WADAV is a web-based development environment that enables users of all technical skill levels to create interactive visualizations via a graphical user interface. Users upload data and decide how to visualize it on a chart or graph. As a user builds a project, he or she interacts with the visualization via a live preview. Once finished, download the graphs and report of data analysis.

3.1.4. Definitions, Acronyms, and Abbreviations

- Analysis – Detailed examination of the elements or structure of something.
- Data Set - A data source or data subset.
- Data Source - A collection of data, such as the contents of a CSV or JSON file.
- Event - A user interaction (click, hover, etc.) that triggers an action.
- EU - End-User
- Visualization - Any technique for creating images, diagrams to communicate a message.

3.2. Overall Description

The following section and subsections present an overall description of the WADAV. In particular, the product has been put into perspective through a detailed assessment of the system, user, software and hardware interfaces.

3.2.1. Product Perspective

WADAV is an interface that processes data to enable people, companies or businesses to develop or Express themselves better. It is an interface where they can simply load data, analyze and visualize the loaded data. It does not require membership, it provides fast and free use. Just load the data and select the process you want to do.

3.2.2. User Interface

This section describes the path a user will take after connecting to the website. After logging in to the web site, you reach the welcome page. And you can see the interface diagram of the web site below.

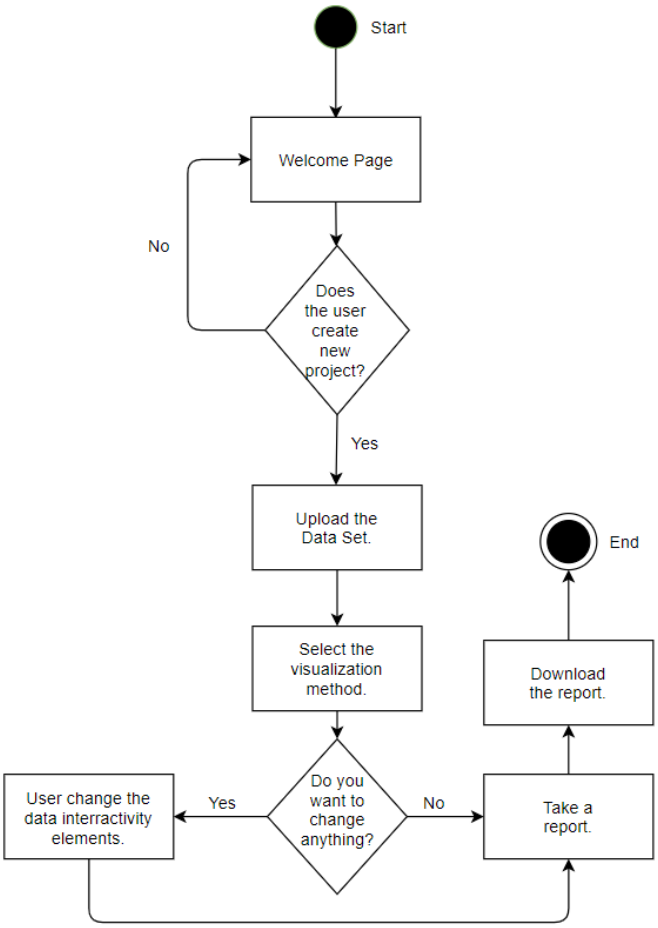


Fig1. User Interface Mapping

The welcome page will Show you instructions. According to these guidelines, you must first create a Project. After you name the Project, you have to upload the dataset that you want to examine. The system then prepares the data set for analysis and you will only have to select the visualization method you want to see. If you want to make changes in visualization methods, you can change the values you want with the interactive interface and see the results of the changes from the graphs and diagrams.

After the process is completed, you can download the report and graphics for your project.

3.2.3. Software Interface

The loaded dataset is prepared for visualization. At this stage, the system applies certain analysis methods to data. Classification, Regression, Analyze and Clustering methods are prepared for data visualization.

3.2.4. Hardware Interface

It can be accessed from all devices with CPU and operating system with Internet access.

3.3. Requirements Specification

3.3.1. Use-Case

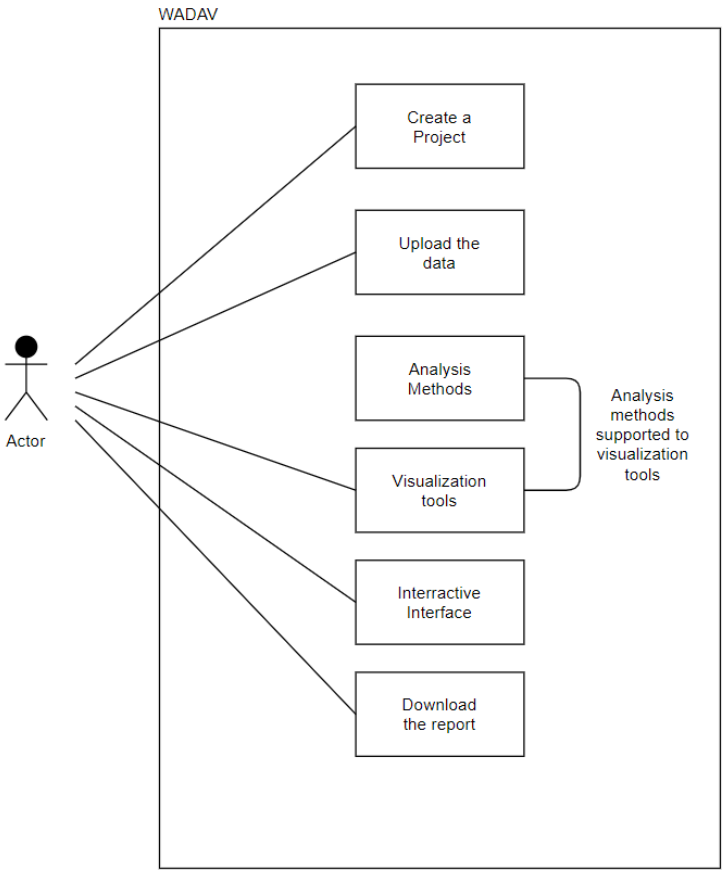


Fig2. Use-Case Diagram

- Create Project
- Upload the data
- Analysis method
- Visualization Tools
- Interactive Interface
- Download the Report

3.3.2. Brief Description

The figure shows the interaction between the actor and the system. The users can see the visualized form of data by uploading their data. Interactive Interface provides users to visualize the graphics associated with the chosen attribute. Users also can change the type of graphics. After all the operations are done, they can download the visualization and analysis report.

4. Software Design Document

4.1. Introduction

This document is the Software Design Document for the Web-based Multifunctional Graphical User Interface for Data Analysis and Visualization-WMGUIDAV Project, briefly Web App Data Analysis and Visualization - WADAV. This system allows users to analyze their data without using any desktop application. In this document, the project will be detailed.

4.1.1. Purpose

The purpose of this Software Design Document is providing the details of project titled as Web Based Multifunctional Graphical User Interface for Data Analysis. The target audience is people who would like to visualize, interpret their data. Analysts,

scientists, and engineers commonly employ this visual method of data analysis. In this project, we will create a Web-based application that is used for visualization data. Data visualization refers to techniques used to communicate insights from data through visual representation. Its main goal is to distill large datasets into visual graphics to allow for easy understanding of complex relationships within the data. By using visual elements like charts, graphs, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. However, before we start to visualize the given data, since real world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors, we firstly preprocess the data. Therefore, preprocessing data is one of the important aim of our project.

4.2. Scope of Project

In this project, we aimed to present statistical and variable information units in a different and creative way by isolating them from classical schematic forms. One of the most important goals of data visualization is to make complex and confusing data normally presented in the classic format easy to understand with easy-to-understand graphical interfaces. We needed to present the data in a comprehensible way and get the right design in line with the overall objectives of the project. The content of this design covers many fields such as tools,IDEs, protocols, back-end, front-end etc. parts used in the project. Firstly, we have decided to use Phyton as the language . Python is a powerful, dynamic programming language used in a wide range of fields. The Pandas library offers high-performance, easy-to-use data structures and data analysis features. Thus, without the need for statistical programs such as R or stata, we can do data analysis and modeling. Combined with tools and libraries such as IPython, statsmodels, and scikit-learn, a powerful data analysis environment can be achieved in terms of performance and productivity. Secondly, The protocol can be used as Web Distributed Authoring and Versioning (WebDAV). İs a hyper text transfer (HTTP) protocol that allows you to manage files and documents stored on web servers. With WebDAV, you can connect to the cloud environment like Google Drive, OneDrive, OwnCloud, and do a lot of things like deleting, writing, copying, moving, and you can do it using your internet browser without installing any applications. As a storage system, we have thought of a platform where large amounts of data can be securely held, quick access to

information, integrity of information, and access to multiple users simultaneously. Instead of stacking all the data in one place, we thought of using MySQL, which stores it in different tables and databases. The control, optimization and repair of the tables are done quickly. Finally, we thought of using HTML5 and CSS3 to visualize the project (website). It has the advantages such as working together using the same database common, making it easy to use for everyone.

4.2.1. Development Responsibility

- Different data import option
- Mapping capability
- Visualizations can make complex topics easy to understand.
- Compare data by time intervals
- Ability to obtain snapshots of data
- Highly customizable and extensible
- Embedding graphics into websites and applications should be easy.
- Stacked graph charts are an effective way to compare and contrast data.
- Scatter plots are a simple way to represent data trends.
- Line charts are effective at showing trends and comparisons.
- Scatter plots can show a multitude of data, especially when color-coded to show more points
- Chord diagrams show relationships between groups of entries.
- Dashboards can showcase numerous data visualizations side by side.
- Bubble charts can showcase numerous data points simultaneously.
- Multi-axis line charts are better when they're annotated (this one uses tooltips when hovering over points on the lines).
- Stacked area line charts are visually striking visualizations.

4.3. Architecture Design

In this process, during which the project report was prepared, we examined many simlar applications and platforms to determine our development methods and the way we will take. We have examined the methods of data preprocessing, data cleaning data analysis, and data visualization of these applications and platforms. As a result of these investigations, we have determined the paths we want to follow.

Sütun1	Big ML	Exploratory.io	Orange
Classification	Desigion Tree - Deepnets	Desigion Tree - Random Forest	Desigion Tree - Random Forest - kNN
Regression	Linear Regression - Logistic Regression - Deepnets	Linear Regression Logistic Regression Random Forest	Linean Regression - Logistic Regression - SVM
Analysis	Cluster Analysis - Time Series - Anomaly Detection	Correlation - Distance - Market Basket - Principal Component	Distance - ROC Analysis - Time Series - Correspondence Analysis
Clustering	K-Mean & G-Mean	K-Mean	K-Mean - Hierarchical Clustering - Louvain Clustering
Visualization	Scatter & Histogram Plot - Pie Chart - Cluster & Tree Visualization	Pivot - Table - Bar - Line - Area - Density - Scatter - Histogram - Word Cloud - Heatmap	Scatter - Histogram - Line - Box Plot - Mosaic Sieve - Venn Diagram - Pythagorean Tree
Interractivity	Create a tables by selected element for graphics.	Select the part of data, analysis & visualize the selected part.	Instruct visualization to send out a data subset that corresponds to the selected part of visualization
User-Interface	Web - Based	Desktop Application	Desktop Application
Data Type	CSV - TSV - ZIP	CSV - Excel - JSON - Log File - SAS/SPSS/STATA	
Extras		Wilconox Test - Kruskal-Wallis Test - Chi-Square Test - Normality Test	

Table of similar applications and methods

4.3.1. Classification

4.3.1.1. Decision Tree

Decision tree is one of the most popular machine learning algorithms. This algorithm is used for classification and regression. It is a part of Supervised Learning. A decision tree is a tree where each node represents a feature(attribute), each link(branch) represents a decision(rule) and each leaf represents an outcome(categorical or continues value).

In this Project, the path of the data will be visualized on the tree.

Advantages:

- Model Interpretability is high which means it is easy to understand.
- Model maintenability is low which means modifying the model is easy.
- Classification for classifying an unseen record is low.

4.3.1.2. Random Forest

This algorithm is another algorithm of classification. This algorithm can also use both for classification and the regression kind of problems and also a supervised classification.

Advantages:

- The same random forest algorithm or the random forest classifier can use for both classification and the regression task.
- Random forest classifier will handle the missing values.
- When we have more trees in the forest, random forest classifier won't overfit the model.
- Model can do classification for categorical values .

4.3.1.3. KNN

This is another classifiction algorithm that we are going to use.

The aim of this study is to examine the closeness of the new individual to the k from the previous individuals.

Advantages:

- Modifying the model is easy.
- Model training cost is low.
- It is easy to understand.

4.3.2. Clustering

4.3.2.1. K-Means Clustering

This machine learning algorithm is one of the simplest and popular unsupervised algorithm. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided.

Advantages:

- The algorithm is easy to implement.
- It scales to large data sets.
- Adabtivity is high to new data.

4.3.2.2. Hierarchical Clustering

This is an another clustering algorithm. It groups similar objects.

Hierarchical clustering can be performed with either a distance matrix or raw data.

4.3.3. Regression

4.3.3.1. Linear Regression

Linear regression is a basic and commonly used type of predictive analysis. attempts to model the relationship between two variables by fitting a linear equation to observed data. we predict scores on one variable from the scores on a second variable. When there is only one predictor variable, the prediction method is called simple regression. Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line.

- The principal advantage of linear regression is its simplicity, interpretability, scientific acceptance, and widespread availability.
- Linear regression is the first method to use for many problems.
- Analysts can use linear regression together with techniques such as variable recoding, transformation, or segmentation.

4.3.3.2. Logistic Regression

Logistic regression is another technique borrowed by machine learning from the field of statistics. It models the probability of the default class. Logistic regression is a linear method, but the predictions are transformed using the logistic function

Advantages:

- Logistic regression is a linear method, but the predictions are transformed using the logistic function. Logistic regression is less prone to over-fitting but it can overfit in high dimensional datasets.
- Logistic regression is easier to implement, interpret and very efficient to train.

4.3.3.3. Support Vector Machine (SVM)

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Advantages :

- SVM works relatively well when there is clear margin of separation between classes.
- It is more effective in high dimensional spaces.
- It is effective in cases where number of dimensions is greater than the number of samples.

- It is relatively memory efficient.

4.3.4. Analysis

4.3.4.1. Cluster

Cluster analysis is a class of techniques that are used to classify objects or cases into relative groups called clusters. Cluster analysis is also called classification analysis or numerical taxonomy. In cluster analysis, there is no prior information about the group or cluster membership for any of the objects. Cluster Analysis has been used in marketing for various purposes. Cluster analysis involves formulating a problem, selecting a distance measure, selecting a clustering procedure, deciding the number of clusters, interpreting the profile clusters and finally, assessing the validity of clustering.

4.3.4.2. Time Series

Time series analysis is a statistical technique that deals with time series data, or trend analysis. Time series data means that data is in a series of particular time periods or intervals. The data is considered in three types:

- Time series data: A set of observations on the values that a variable takes at different times.
- Cross-sectional data: Data of one or more variables, collected at the same point in time.
- Pooled data: A combination of time series data and cross-sectional data.

4.3.4.3. Principal Conponent Analysis (PCA)

Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to make data easy to explore and visualize and to reduce a large set of variables to a small set that still contains most of the information in the large set. Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components.

4.3.5. Visualization

The table shows the methods that used the most stable and functional applications and platforms (BigML, Exploratory and Orage). We aim to adapt these basic methods to our Project and use them.

Visualization Methods	Intteractivity Method	Extra
Scatter Plot	Creates a table showing the information of the hovered point.	When preparing data for analysis, this graph is created for each attribute.
Histogram Chart	Have a scrolling interface for the appropriate data types.	When preparing data for analysis, this graph is created for each attribute.
Pie Chart	Shows the percentage values of the pie chart by hovering.	
Area Chart	A graph of a single variable or a graph showing the interaction of multiple variables.	
Heat Map	Presents correlation between attributes in a table supported by colors. Values can be viewed by hover.	
Desigion Tree	After hover shows the path of the point.	
Box Plot	Value tables of selected points are displayed.	

Table of Visualization Methods

5. Conclusions

This document has discussed the software requirement specification document for the Web-based Multifunctional Graphical User Interface for Data Analysis and Visualization-WMGUIDAV Project, briefly Web App Data Analysis and Visualization - WADAV. This system allows users to analyze their data without using any desktop application.

The data that people create in the virtual world (social media posts, news, shopping bills, emails, etc.) is increasing and accumulating day by day. These data may

be unnecessary and meaningless to most of us, but they are vital to an analyst, a company or a business. Of course, not every data is meaningful and usable. At this point data analysis programs and platforms come into play. It is precisely these applications and platforms that will reveal the meaning of the available data and determine the value of this treasure. WADAV is a web platform that allows users to review theri data and access the information they need. Data Visualization has been used in many areas to see our data and to interpret it easily. In this report we have represented a Project that is used to visualize our data and the algorithms that we will be using. The purpose of this document is to outline requirements for the WADAV. This document describes the general concept and explains the required features of the system. One of the most important issue here is dealing with the data. We have also discussed Data Preprocessing that is very strong problem in dealing with the data. We have understand it explore it, and make it clear dealing with outliers, noisy data etc.

Referances

[1]"About BigML.com", *BigML.com - Machine Learning made easy*, 2019. [Online]. Available: https://bigml.com/about/. [Accessed: 30- Dec- 2019]

[2]"Exploratory", *Exploratory.io*, 2020. [Online]. Available: https://exploratory.io/. [Accessed: 30- Dec- 2019]

[3]"Data analysis", En.wikipedia.org, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Data_analysis. [Accessed: 04- Nov- 2019]

[4]Guru99.com, 2019. [Online]. Available: https://www.guru99.com/what-is-data-analysis.html. [Accessed: 05- Nov- 2019]

[5]"Data visualization beginner's guide: a definition, examples, and learning resources", Tableau Software, 2019. [Online]. Available: https://www.tableau.com/learn/articles/data-visualization. [Accessed: 11- Nov- 2019]

[6]"What is Data Visualization | Basic Concept with Charts And Graphs", EDUCBA, 2019. [Online]. Available: https://www.educba.com/what-is-data-visualization/. [Accessed: 12- Nov- 2019]

[7]Dataaspirant. (2020). How the random forest algorithm works in machine learning. [online] Available at: https://dataaspirant.com/2017/05/22/random-forest- algorithm-machine-learing/ [Accessed 10 Jan. 2020].

[8]Trevino, A. (2020). Introduction to K-means Clustering. [online] Blogs.oracle.com. Available at: https://blogs.oracle.com/datascience/introduction-to-k- means-clustering [Accessed 10 Jan. 2020].

[9] Onlinestatbook.com. (2020). Introduction to Linear Regression. [online] Available at: http://onlinestatbook.com/2/regression/intro.html [Accessed 10 Jan. 2020].

[10] Brownlee, J. (2020). Logistic Regression for Machine Learning. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/logistic-regression-for-machine-learning/ [Accessed 10 Jan. 2020].

[11]Medium. (2020). Support Vector Machine — Introduction to Machine Learning Algorithms. [online] Available at:https://towardsdatascience.com/support-

vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 [Accessed 10 Jan. 2020].

[12] Explained Visually. (2020). Principal Component Analysis explained

visually. [online] Available at: http://setosa.io/ev/principal-component-analysis/ [Accessed 10 Jan. 2020].

[13] Bioinformatics Laboratory, U. (2020). Orange Data Mining - Documentation. [online] Orange.biolab.si. Available at: https://orange.biolab.si/docs/ [Accessed 10 Jan. 2020].