YAKUP KOÇ
IMPLEMENTATION OF REAL-TIME GEOGRAPHICAL DATA
IN A BUSINESS INTELLIGENCE PLATFORM USING
MACHINE LEARNING METHODS
NEU
2021

IMPLEMENTATION OF REAL-TIME GEOGRAPHICAL DATA IN A BUSINESS INTELLIGENCE PLATFORM USING MACHINE LEARNING METHODS

A THESIS SUBMITTED TO THE INSTITUTE OF GRADUATE STUDIES
OF
NEAR EAST UNIVERSITY

By
YAKUP KOÇ

In Partial Fulfillment Of The Requirements For
The Degree Of Master Of Science
In
Information Systems Engineering

NICOSIA, 2021

IMPLEMENTATION OF REAL-TIME GEOGRAPHICAL DATA IN A BUSINESS INTELLIGENCE PLATFORM USING MACHINE LEARNING METHODS

**A THESIS SUBMITTED TO THE INSTITUTE OF GRADUATE STUDIES
OF
NEAR EAST UNIVERSITY**

**By
YAKUP KOÇ**

**In Partial Fulfillment Of The Requirements
ForThe Degree Of Master Of Science
In
Information Systems Engineering**

NICOSIA, 2021

Yakup Koç: IMPLEMENTATION OF REAL-TIME GEOGRAPHICAL DATA IN A BUSINESS INTELLIGENCE PLATFORM USING MACHINE LEARNING METHODS

Approval of Director of
Institute of Graduate Studies

Prof.Dr. K. Hüsnü Can Başer

We certify this thesis is satisfactory for the award of the degree of Master of Science in
Information Systems
Engineering

Examining Committee in Charge:

Doç. Dr. Boran Şekeroğlu	Department of Information Systems Engineering, NEU
Doç. Dr. Dilber Uzun Özşahin	Supervisor, Department of Biomedical Engineering, NEU
Doç. Dr. Kamil Dimililer Kâmil	Department of Electrical & Electronic Engineering, NEU

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Yakup Koç

26.02.2021

ACKNOWLEDGEMENTS

First I would like to express my gratitude to the Father Almighty, maker of heaven and earth; who has given me health, dedication, ability to cope with the problems that I encounter and the people who add value to my life; to help me achieve all the success I wish to reach throughout my life. I am deeply grateful to my father, whom I lost four years ago, for all of the life lessons he taught me and for being my biggest supporter. I would like to express my deepest appreciation to my mother, my wife and our children for their continuous support and encouragement from when I was accepted to the Near East University Information Systems Engineering Master's Program in 2017 until my successful completion. I sincerely thank the Near East University family for providing scholarships for my entire master's program. I would like to thank Assoc. Prof. Dr. Boran Şekeroğlu, Head of Information Systems Engineering Department, for their interest and academic support during my master's program. I am particularly grateful to Assoc. Prof. Dr. Dilber Uzun Özşahin for their support, guidance and corrections towards my success throughout my thesis.

I would like to express my gratitude to my Professors İbrahim Doğan, Besime Erin and Erdi Şafak for their support and precious contribution, helping me successfully complete the program. And endless thanks to my advisors, professors, colleagues and friends whose names I cannot count, who added value to my work that I completed with great self-confidence.

ABSTRACT

When the question “Where?” is asked in the newest scientific research areas such as machine learning, internet of things, deep learning, big data, data minings and artificial intelligence around the world; geographical data gives us the answer. Studies that don't have the answer to this question have great errors of estimation. The applications that are used the most around the world in any field ask for your location information as their first action.

Today, with developments in technology, the environments where data is stored are increasing. In addition, researchers have started to conduct analyses and make predictions about the data in order to process these data and to make future data meaningful. Machine learning algorithms have started to be used process these analyses and predictions with large data sets.

The evolution of technology and the industry has been continuing to ignore all authorities. While these developments were taking place, the research was focused on business intelligence. Business intelligence is defined as analyzing and processing the raw version of big data and the processes that convert the data to significant and productive information. In this study, a mobile and web-based business intelligence platform was developed. The real-time data stacks produced in this application were put in relevant tables within the PostgreSQL database in accordance with business logic. The geographical data within this data stack, latitude and longitude, were classified using the k-means clustering algorithm in data mining. With this classification, it was tried to make the business intelligence platform management significant and productive. Besides, as the users were going through their daily routines, their maps were made using their GPS codes. These maps helped increase the productivity of business processes by implementing machine learning methods to business intelligence.

Keywords: Geographical data; machine learning; business intelligence; mobile applications; K-means algorithm; web service; big data

ÖZET

Küresel anlamda en yeni bilimsel araştırmalar olan makine öğrenmesi, nesnelerin interneti, derin öğrenme, büyük veri, veri madenciliği ve yapay zeka gibi konularda "nerede?" sorusunun cevabını coğrafi veriler vermektedir. Bu sorunun cevabının olmadığı araştırmalarda büyük yanılgı tahminleri oluşmaktadır. Dünyada herhangi bir alanda en fazla kullanılan uygulamaların ilk aksiyonu sizden konum bilginizi paylaşmanızı istemektedir.
Günümüzde teknolojideki gelişmelerle birlikte verilerin depolandığı ortamlar artmaktadır. Ayrıca araştırmacılar, bu verileri işlemek ve gelecekteki verileri anlamlı hale getirmek için verilerle ilgili analizler yapmaya ve tahminler yapmaya başlamıştır. Bu analiz ve tahminleri büyük veri setleriyle işlemek için makine öğrenme algoritmaları kullanılmaya başlanmıştır. Teknoloji ve endüstrinin çok hızlı gelişen evrimi tüm otoriteleri yok saymaya devam etmektedir. Bu gelişmeler yaşanırken araştırmalar iş zekası üzerinde yoğunlaşmıştır. İş zekası büyük verinin ham halinin analizi yapılarak işlenmesi, anlamlı ve verimli bir bilgiye dönüştürülmesi sağlayan süreçler olarak tanımlanır. Bu tez çalışması ile birlikte mobil ve web tabanlı iş zekası platformu geliştirilmiştir. Geliştirilen bu uygulama üzerinde üretilen gerçek zamanlı veri yığınlarının Postgesql veritabanı içerisinde iş mantığına göre ilgili tablolarla konumlandırılmıştır. Konumlandırılan bu veri yığının içerisindeki coğrafik verileri olan enlen ve boylam verisini veri madenciliği içerisinde bulunan k-means kümeleme algoritması kullanılarak sınıflandırılma yapılmıştır. Yapılan sınıflandırma ile iş zekası platformun yönetimi anlamlı ve verimli hale getirilmeye çalışmıştır. Ayrıca çalışmamızda kullanıcıların gündelik rutinlerini gerçekleştirirken GPS kodlarıyla haritaları oluşturulmuştur. Bu haritalar makine öğrenmesi metotlarının iş zekasına uygulanmasıyla iş süreçlerinin verimliliğin arttırılmasını sağlanmıştır.

Anahtar Kelimeler: Coğrafi veri; makine öğrenmesi; iş zekası; mobil uygulamalar; k-means algoritma; web servis; büyük veri

TABLE OF CONTENTS

CHAPTER 4: ALGORITHMS

CHAPTER 5: RESULT AND APPLICATIONS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI:	Artificial Intelligence
AWS:	Amazon Web Services
CDR:	Call Detail Record
CSV:	A Comma-Separated Values
DOM:	Document Object Model
GPS:	Global Positioning System
ISO:	International Organization for Standardization
JDBC:	Java Database Connectivity
JSON:	JavaScript Object Notation
MIT:	Massachusetts Institute of Technology
MVC:	Model View Controller
SQL:	Structured Query Language
XML:	Extensible Markup Language

CHAPTER 1

INTRODUCTION

1.1 Background of Study

It is seen that there is a rivalry over the use of internet traffic in the world between desktop and mobile devices. They are in serious competition from the constant development of technology to the usability of the devices. As the devices get smaller and the functionality of these devices increase, it is seen that this rivalry is progressing in mobile devices' favor. In the report published in January 2020 by StatCounter, a web traffic analyst company, it is stated that while mobile devices use 53.3% of the total traffic with an increase of 8.6%, desktop devices use 44% of the total traffic with a decrease of 6.8%. Mobile devices, which are increasingly taking place in human life, also use internet traffic in different areas. Widespread traffic is seen in various categories such as social media, video streaming, shopping, eating and business solution applications. Changing lifestyles add new trends into our habits. In our study, we analyze the applications to figure out what kind of a problem that they solve.

The aim of our study is to help people make a profit while carrying out their daily errands or help them get their errands provided saving on time and resources, via machine learning methods based on our GPS coordinates and location movements (Aarts, et al., 2003). I analyzed many new and old applications for sector research, and as a result of these, I ended up with results that couldn't go beyond old-fashioned client-server algorithms with specific limitations. Applications such as Bounty, FieldAgent and Armut can be given as examples. One of these applications offers certain missions under stiff conditions, without letting the users create mission requests themselves. The other one claims to complete missions with limited human resources under certain limitations and asks for high prices. Our study comes up with an innovative and more functional solution that provides everything that these applications provide and also lets users create missions.

1.2 Statement of Problem

Our study aims to come up with new solutions, implementing new technologies such as machine learning to business intelligence. Especially during the pandemic; travel restrictions, treatments or quarantine processes affected many sectors negatively. For instance, transportation and logistics companies have been experiencing great difficulties in carrying out their work (Alanis, et al., 2018). Large suppliers are forced to set up their own cargo systems. Many sectors are experiencing serious difficulties.

None of the existing mobile applications can provide a method that can be a solution to these issues. This is where our application comes into play. As you go through your daily routines, our application is mapping your most visited routes in the background using your GPS information via machine learning (Figure 1.1).

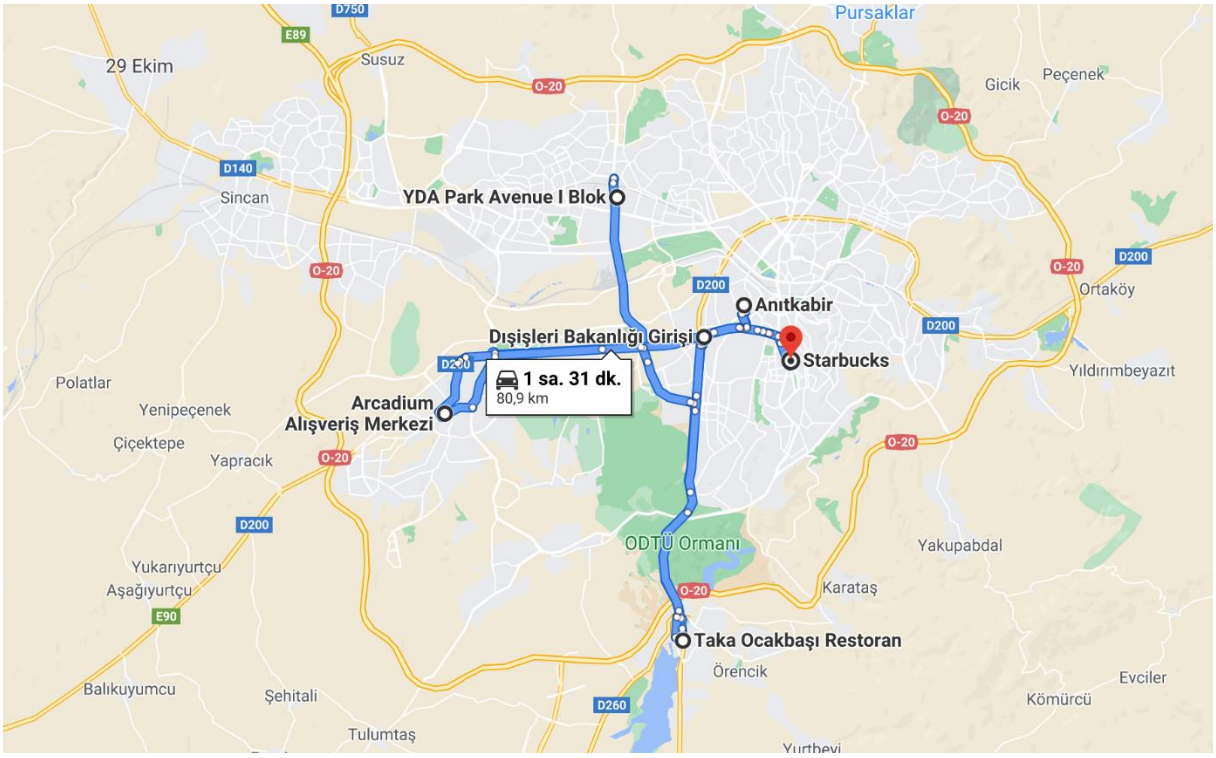


Figure 1. 1: The most visited route of a user in daily life

While you continue carrying out your daily routines, when you're filling up your tank at the gas station for instance, the headquarters will be able to control that station thanks to your rating. Or when you are shopping at a mall, you will be able to make a gain by providing your impressions of the store as an inspector. You can also create missions or carry out existing missions in another location, saving on time and resources.

Our study can carry out countless missions, more than you can imagine, with the opportunities that it provides and with its local, liberal and innovative approach. It provides solutions to hundreds of thousands of problems via business intelligence without any time loss, inconvenience and using suitability control (Watson, et al., 2003).

1.3 Geographic Data

Geographic data can be defined as data that are formed with significant discipline which haves a direct or indirect reference to any position on the earth's surface. Studies in this area gained momentum in academic circles during the mid-20th century. One of the first discussions about the applications of geographic information systems discipline and alternatives was written by the famous geographer Professor Michael Frank Goodchild in November 1992. A working group named ISO/TC 211 aims to establish a resolved standard discipline for geographic data (Goodchild, et al., 1992). Vector folders that include spatial data such as latitude and longitude, raster folders that transfer land cover spheres formed out of pixels and grids, and a geographic database that bear these folders and web folders and many temporal data are designated as geographical data (Figure 1.2).

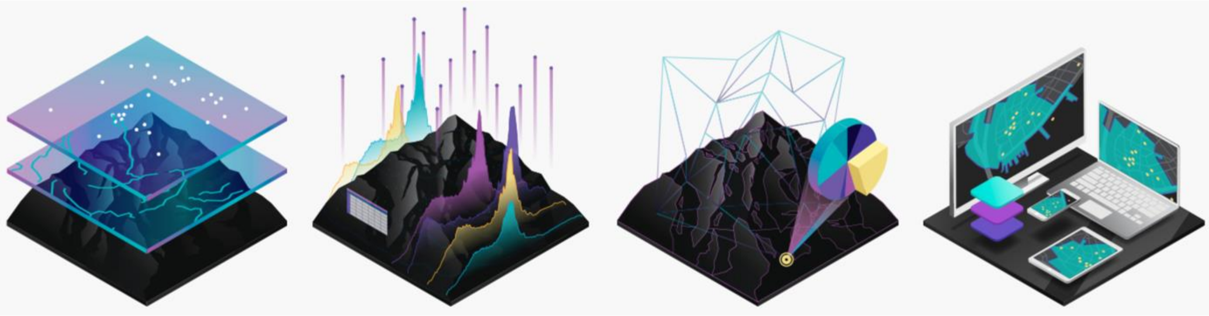


Figure 1. 2: Layers of Geographical Data

Public enterprises and private companies use the applications developed by analyzing the geographic data on the maps in many areas. These applications both make our lives easier and presents stimulating developments in many fields. These include users who get address directions, the logistics firms that research the fastest and shortest routes, insurance companies that evaluate natural disasters and crime using spatial data analysis and their pricing model designations, the aviation sector's flight operations and location analysis for maximum profit regarding new branch samplings.

In the areas of scientific research such as machine learning, internet of things, deep learning, big data, data mining and artificial intelligence, geographic data provides the answer to the question "where?" When there is no answer to this question in the research folders, there are big error forecasts.

The most widely used applications in the world in any field ask for your location as their first action. Even the simplest weather forecast application would not provide results without sharing your location. In the forensic field, courts find out the answer to the question "where" from the defendant's Historical Traffic Search (HTS) records. The increasing feasibility of the location marks in large amounts creates unique opportunities for many sector operations in order to transform the data discovery paradigm (Wolfgang, et al., 2012). Many unnecessary forced realization processes in many business fields create unproductive energy consumption in the sectors. In the application that we employ in this thesis, a user's geographical data will be obtained while they go about their daily routine. This creates a movement map. The users will be appointed some duties through this map using business intelligence and machine learning methods. For example, while the user is browsing a shopping center he regularly visits, the headquarters of any of the stores will be asked to report his observations as a secret auditor of the store to the user. In this way, a brand will have audited all of its branches by paying a fee far below these to the user without impacting workloads such as road, fuel and accommodation without assigning personnel to its branch.

1.4 Business Intelligence

The rapidly developing evolution of technology and industry continues to ignore all authorities. While these developments have been taking place, research has focused on business intelligence. Business intelligence is defined as the processes that enable the processing of big data by analyzing them in raw form and transforming them into meaningful and efficient information (Luhn, et al., 1958). This term was first added to the terminology in 1958 by technology researcher Hans Peter Luhn. Considering the great impact of social media platforms with the introduction of democracy to the Internet with Web 2.0, business intelligence has reached a much more functional place than where it started. With increased analytical capability, the transformation of these big data into a big impact has now reached an automatic solution with machine learning technologies (Roger, et al., 2012).

Standart business intelligence standard solutions are complex event analysis, performance comparisons, reporting solutions, process and data mining, multi-dimensional analysis methods, logical analysis and analytical predictions. All of these solutions analyze information and provide guidance to support decision-making processes. The information required to perform certain routine operations in corporate structures is stored in a database within enterprise resource planning (ERP) applications. Business intelligence tools, on the other hand, unite all the data in data warehouse, which is a different database structured for data mining, instead of processing previously processed a data and revealing meaningful data (Özekes, et al., 2003).

In the business intelligence project that we will implement in thesis, it will be ensured that a map is created by analyzing data on the mobile devices that people use while performing their daily routines. Here, apart from personal information, the user’s age, gender, education level and the map he creates will be added to the data warehouse. Then, while performing these routines, estimates will be made, such as the shortest path for logistics operations, the shortest, the most appropriate cost and the shortest time for inspections, and determining the branch location for the most profitability in human population analysis.

1.5 Research Question

Questions to solve the problems that are discussed in this research thesis;

- What is the state of mobile applications on business intelligence around the world?
- What research is there on geographical data and machine learning methods?
- What are the consequences of desktop and mobile internet use habits of users?
- Why is geographical data vital for mobile applications?
- Can mobile solutions be provided for business processes that are disrupted by restrictions for an extended amount of time due to the pandemic?
- How can processing geographical data using machine learning methods increase business intelligence productivity?
- Would it be possible to save time and resources by defining the business processes to be carried out by the mobile application?

1.6 Methodology

Research has been conducted to find out if the world of technology which evolves upon the newest technologies, new concepts on business intelligence and the habits of internet users has anything to say or has any kind of a solution for today (Asheibi, et al., 2009). An objective analysis of various statistical data such as user experiences, internet traffic tendencies, usage percentages of new trends as well as concepts such as big data, data mining, geographical data has been carried out.

The thesis presents new ideas for the implementation of ever-growing technologies in business intelligence.

1.7 Significance of the Study

The study presents a feasible solution in various sectors for not only Turkey but also for all countries, with a global perspective. It is aimed to create an ecosystem that turns the users which are only consumers, into producers or service providers with a new and feasible point of view by analyzing mobile applications used for various sectors on all platforms.

A system that eliminates inefficiency, saves time and resources, and comes with a large human resource for the employer is being proposed. It is a solution that helps users that complete missions that are already around their location make gains without any additional workload (Sennaroglu, et al., 2018). This study can be defined as the evaluation of the results of the solutions that new technologies can carry out on business intelligence.

1.8 Organization of the Study

This study is expressed under six main headings in detail. Chapter one consists of previous research related to our thesis, definition of the problem, geographical data, business intelligence, research questions that our study is based on, methodology, the significance of the study and the organization. Through these subheadings, the main headings of our study have been explained.

In chapter two; a literature review was conducted on studies about national and regional mobile applications, geographical data analyses, machine learning methods and processes affecting business workstreams. For K-means algorithms, which is the main element of our study, a detailed literature review has been carried out.

The third chapter involves detailed work on the applications and technologies used in our study. The technologies used in the study were explained in detail, which are: Python programming language, Pandas library, Numpy library for large and multidimensional arrays and matrices, Java Spring Boot framework, React.js for desktop solutions, React Native for mobile applications and finally, the database application Postgre SQL (Douglas, et al., 2005).

In chapter four; machine learning technology, which is the cornerstone of our thesis, is explained and the algorithms of these technologies are mentioned under subheadings in further detail. As machine learning technologies were being defined, fields of use and solutions were thoroughly studied and conveyed. The subheadings of this technology are Native Bayes, Support Vector Machine, Logistic Regression, K-Means Algorithm and Long Short algorithms.

Chapter five consists of the findings and our application. Introduction, our research strategy, general structure of the system, software features of the application, features of the database module, features of user interface, design of the mobile interface, background coding features of the application and classification based on geographical data are explained. All of the intricate details we tried to explain with our study are stated within this chapter. All of the technologies that we have explained in detail are tried to be embodied in an application.

Chapter six ends up with conclusions and recommendations.

CHAPTER 2

LITARATURE REVIEW

Data mining, which is used in many different areas especially in the business world, was shown among 10 technologies that will change the world according to the statement published by MIT (Massachusetts Institute of Technology) in 2001. Recently, researchers have been building many studies on the data mining discipline to make data more significant and effective within big data. New studies on data mining (Parcello, et al., 2017), which will hold even more substantiality in the future, are being conducted day by day. New areas of use are being added to data mining every day, which is used in many different sectors from medicine to space sciences; and the rise of more advanced algorithms will have even more importance upon the use of data mining in business intelligence (Erişti, et al., 2010).

Researchers generally classify the studies on data mining as classification, binning, estimation, prediction and clustering. Clustering, which is one of the main aims, is used in many areas such as statistical data analysis and pattern recognition (Cielen, et al., 2016). Clustering algorithms that bring together objects with similar features, placing the data within the database under groups or clusters, are of great importance in the field of data mining.

With the increase in population and technologies in recent years, power quality and energy efficiency have become some of the most important parameters of power systems due to the rapid increase in energy demand and energy costs. To increase this efficiency and balance energy use, the researchers have published many articles that mention data mining k-means clustering algorithm solutions on energy data. Non-linear charges draw current from the electrical grid that contains harmonic components (Gersho, et al., 1992). Harmonics, which are defined as a power quality problem, create serious problems on the network and receivers, and they have become an important issue that has been focused on within this field in recent years. Today, a rapid increase in the placement of harmonic monitoring systems in electrical distribution systems is being seen to detect and reduce harmonic distortion problems. Great sizes of data are being reached due to the increase in harmonic systems and it is becoming harder to monitor harmonic distortions and the state of energy visually through the data. Many researchers felt the need to manage data this large with algorithms under the data mining discipline. Overall, geographical-based real-time business intelligence has been tried to be produced with a clustering approach to data (Grady, et al., 2001).

The UBER mobile application, which is used in many developed countries, also uses data mining techniques in many ways, from fare calculation to the positioning of the tools in the best way possible, in order to maximize user management and profit (Güngör, et al., 2008). UBER travel data sets are being used to get results in order to analyze and monitor the GPS data of the vehicles and the status information of the users. Classification algorithm studies such as k-means under data mining are being conducted using real-time data, sometimes also including geographical data, to make applications that have similar business intelligence more effective and to make logistics management easier.

Researchers have been using data mining algorithms effectively in many fields such as marketing, banking and insurance. It is also being used heavily in medicine. A software was developed using the k-means algorithm with the laryngeal cancer surgery data collected by Kocaeli University Faculty of Medicine, Department of Otorhinolaryngology. The software developed by the researchers makes it easier for medical doctors to make predictions for the future by analyzing retrospective records, providing a business intelligence software that might be helpful in decision making. With this analysis software, many different evaluations can be made while analyzing retrospective data using variable parameters; current and future cases can be predicted for all different situations, the probability of tumor recurrence and the prospective patients' survival probability after surgery can be evaluated for current and future cases, correctly predicted preoperative stages can be visualized and analyzed, and thus, preoperative prediction success can be evaluated. It was demonstrated that by monitoring information of successful surgeries, it is possible to get an idea for future surgery preferences (Piatetsky, et al., 2019).

Many different parameters specific to air transportation create big data; such as technical capabilities of airports, passenger capacities and populations of the cities where the airports are located. It has been an area of interest for the researchers, to significantly process this big data and to reflect this on business intelligence (Schmidt, et al., 2015). In addition, when we consider the costs, workforce and time loss in the aviation industry within the scope of

Digital Transformation; it is seen that digital optimization which is aimed to be provided with big data, machine learning techniques and other elements of current technology not only makes it convenient for the aviation sector but also for almost all different areas of life. This is the reason why it has been the subject of many research studies and has contributed to the development of the business intelligence platform (Raschka, et al., 2020).

CHAPTER 3

TOOLS AND TECHNOLOGIES

3.1 Python

Python is an object based, advanced level programming language that functions on all platforms. The first version was published by Guido van Rossum in 1991. When first published, it was evaluated as sufficient by the technologic and academic circles due to its functional, open source and free structure compared to other programming languages. The most important factors that Python features are firstly its readable code writing technique and secondly increasing improving transaction productivity with fewer code lines(Mark, et al., 2001). Some of the internet giants that use Python are Google, Facebook, Instagram, Spotify, Quora, Netflix and Dropbox.

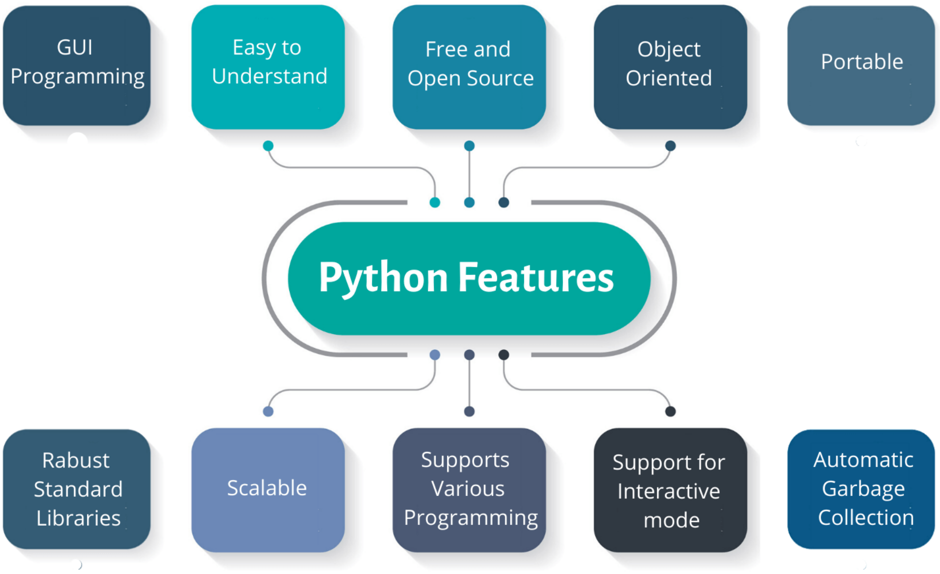

GUI Programming
Easy to Understand
Free and Open Source
Object Oriented
Portable
Python Features
Rabust Standard Libraries
Scalable
Supports Various Programming
Support for Interactive mode
Automatic Garbage Collection

Figure 3. 1: Python features

The characteristics of Python are shown in the table above (Figure 3.1). As far as the versions are concerned: Pyhton version 1.0 was published in January 1994, Python 2.0 in October 2000, Python 3.0 in December 2008 and finally Python 3.9.0rc2 beta in September 2020. Python is known with her free structure that maintains its existence in all platforms. Some of the application fields are: Web applications, Desktop Applications, Console Applications, Software Development applications (In order to operate as a support language, Control Management, Test), Scientific and Equntitative Applications (Artificial Intelligence and Machine Learning), Business Applications (E-commerce and ERP solutions), Audio and Video Based Applications, 3D CAD Applications, Corporate Applications and Sight Processing Applications (Andreas, et al., 2017).

Python stands out in machine learning and explaining operations with deductive code analytics in the scientific research. While increasing the action performance fertility, it gains the strength of owning a library that prioritizes performance from its community. KDnuggest's research conducted on nearly 2000 analytic, data science and machine learning developers proves the thesis that it is the most favored program language (Davy, et al. 2016).

3.2 Pandas

Pandas is a software library that was first coded in Python programming language for high performance flexible tools and financial data analysis by Wes McKinney on January 2008. Eventhough it was written for financial solutions as a starting point by its developer, today it is used for machine learning techniques in scientific circles. Since it is supported by open source code developers, there is no license cost for usage (Chen, et al., 2018).

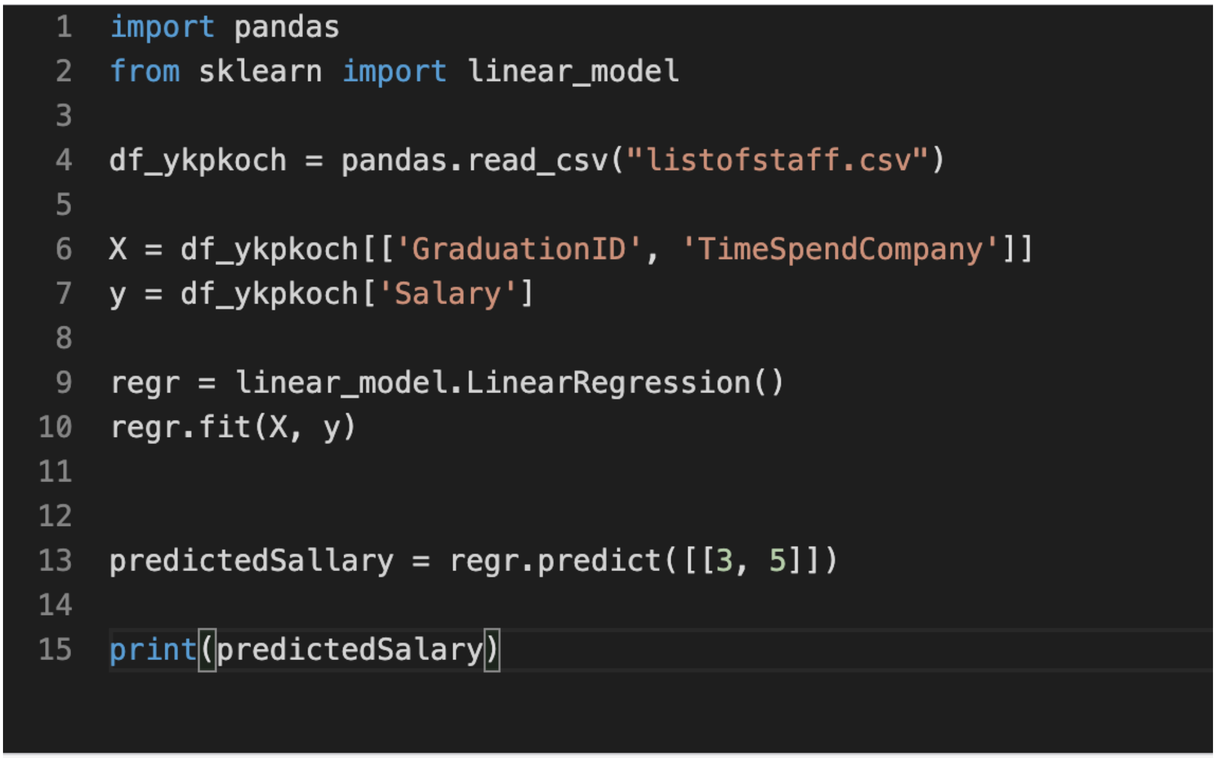
The most important feature is DataFrames. In addition to these, handling of data, alignment and indexing, handling missing data, cleaning data, input and output tools, multiple file formats supported (JSON, CSV, EXCEL, HDF5), merging and joining of datasets, A lot of time series, optimizing performance, Python support, visualizing, grouping, masking data, unique data and performing mathematical operations on the data are the most important factors that make Pandas stand out. To run all these functions requires NumPy, python-dateutil and pytz libraries (McKinney, et al.,2017).

A simple salary forecast study conducted by utilizing pandas libraries and machine learning is illustrated with an example (Jake, et al., 2016). The personnel list can be found below: the salary amount related to years of employment at the company and education level (Primary school 1, Middle school 2, Undergraduate 3, Master's Degree 4, PhD and above 5) are stated in the table.

Table 3. 1: List of staff

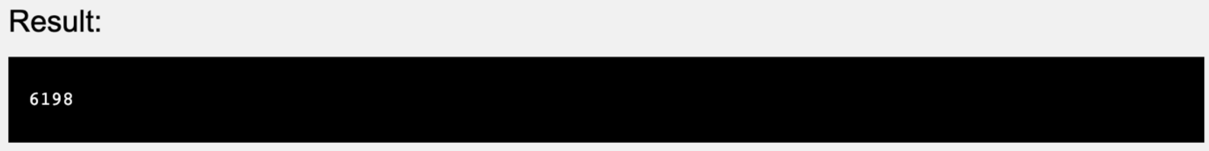
Username	Salary	Years Spent at the Company	Educaiton Level
User1	9854	3	5
User2	8654	5	4
User3	6543	7	3
User4	5489	10	2
User5	3865	15	1
User6	7654	4	4
User7	9432	2	5

Below code was written using predictedSalary.py(Figure 3.2) python and pandas libraries. Here, it became possible to forecast the salary of an employee with a bachelor's degree and five years of experience by importing a listofstaff.csv (Table 3.1) folder in the linear regression method (Figure 3.3).



```
import pandas
from sklearn import linear_model

df_ykpkoch = pandas.read_csv("listofstaff.csv")

X = df_ykpkoch[['GraduationID', 'TimeSpendCompany']]
y = df_ykpkoch['Salary']

regr = linear_model.LinearRegression()
regr.fit(X, y)


predictedSallary = regr.predict([[3, 5]])

print(predictedSalary)
```

Figure 3. 2: PredictedSalary.py folder



Result:

```
6198
```

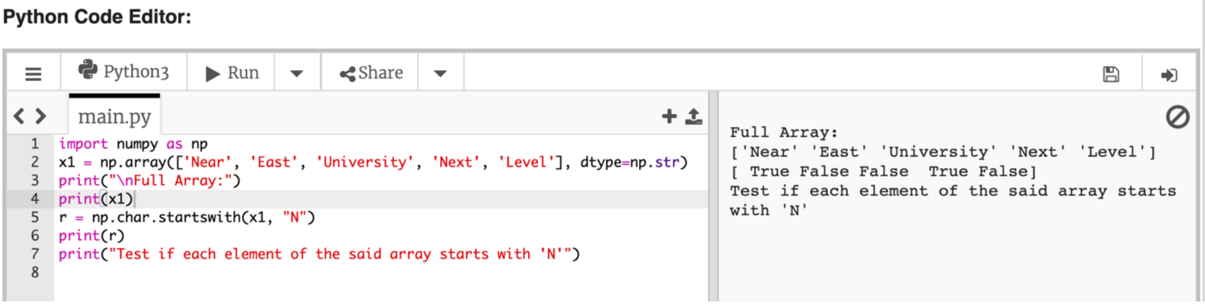
Figure 3. 3: Result of Predicted Salary

3.3 NumPy

When it was first published, Python only supplied solutions for simple mathematical transactions. For the advanced mathematical transactions, the first Numeric library was formed. Later, NumAarray library, which had a flexible and rapid transaction capacity. Was created. NumPy was developed in 2006 by combining the abilities of open source code community Oliphant Numeric (1995) and NumArray libraries. NumPy is a Python programming language library that has several transaction capabilities, such as sequences, matrices, integral and differential equation solutions (Travis, et al., 2015).

Ndarray class lies on the basis of the NumPy library. Intended transactions are carried out on the possible single and multidimensional homogeneous sequence objects. Since it is an open source software, it paves the way to integrate with the codes that were written in different languages (c/c++, Fortran). Eventhough it has a homogeneous sequence structure in identifying the data types and advanced mathematical transactions conducted by the developers, it may perform as a high-dimensional cup function working with generic data. It has such functions that performs several mathematical transactions such as trigonometric, hyperbolic, logarithmic, linear algebras and matrices.

Performance and functionality are enhanced through division of the big data masses into smaller pieces by designated limitations (Ivan, et al., 2015). Matpolib module and graphic drawing have assesment ability. In order to form structures that can integrate with different databases, it puts dtype function to the service of the developers.



```
import numpy as np
x1 = np.array(['Near', 'East', 'University', 'Next', 'Level'], dtype=np.str)
print("\nFull Array:")
print(x1)
r = np.char.startswith(x1, "N")
print(r)
print("Test if each element of the said array starts with 'N'")
```

```
Full Array:
['Near' 'East' 'University' 'Next' 'Level']
[ True False False  True False]
Test if each element of the said array starts
with 'N'
```

Figure 3. 4: NumPy library sample code

On the above example, we carried out an example in Python programming language with Numyp library and related sequences (Figure 3.4). We identified a sequence with five personnel and managed find the sequences starting with starswith function and letter N (Eli, et al.,2012).

3.4 Java Spring Boot

Spring Boot was first developed as an open source framework in 2002 by Rod Johnson. The main goal in this project was to present the collective solutions to the problems that developers faced. This way, rather than writing thousands of lines of codes, software projects could gain productivity, speed, convenience, security and comfort by using domain-based developed libraries (Craig, et al., 2016). Instead of spending time with general routine procedures, the developers could focus on business analytics.

Micro services come with thousands of features and more than 200 JAR libraries that supply reactive, cloud technologies, network applications, independent platforms, incident oriented and automatic tasks. When adding the project to Spring Boot, it would provide automatic configuration according to needs (Josh, et al., 2013). For example, if you would like to connect a database to your project, this connection path is automatically done. Through minimum configuration that prioritizes decreasing the time spent on the project, it is possible to be exempt from the complex XML configurations.

3.5 React.js and Native

React.js is a JavaScript library formed to develop dynamic interfaces by Facebook and the open source developers community in 2013. React Native was published in 2015 in order to develop mobile applications for the Apple and Google Play platforms using this library.

The most important feature of React.js is its use of the virtual DOM (document object model). DOM shelters many labels inside. JavaScript method updates are normally slow. Updating DOM labels one-by-one takes more time than this process. Virtual DOM is a shadow of real DOM and it harbors all of its qualifications. Instead of specifically conducting these update processes one by one in full, Virtual DOM updates the parts that needs updating and adds speed. The control mechanism is ensured by unilateral data validation where project data advance on one layer (Azat, et al., 2017). The project is divided into different components through React.js. All the components are written with XML and JavaScript composition and JSX which includes conditional statements and life cycle methods that give great comfort to the developers.

Mark Zuckerberg wrote the first Facebook code with php in October 2003. In the subsewuent process, hundreds of developers working on the same project faced many problems. Zuckerberg confessed in the following statement: “The biggest mistake we made as a company was betting too much on HTML as opposed to native". PHP brought many problems along with its security and architecture. Facebook's software development team could no longer tolerate these problems and decided to use React technologies. Then, 8868 compenies followed suit in startinged to use React. Uber, Airbnb, Facebook, Pinterest, Netflix, Instagram, Udemy and Twitter (Alex, et al., 2017).

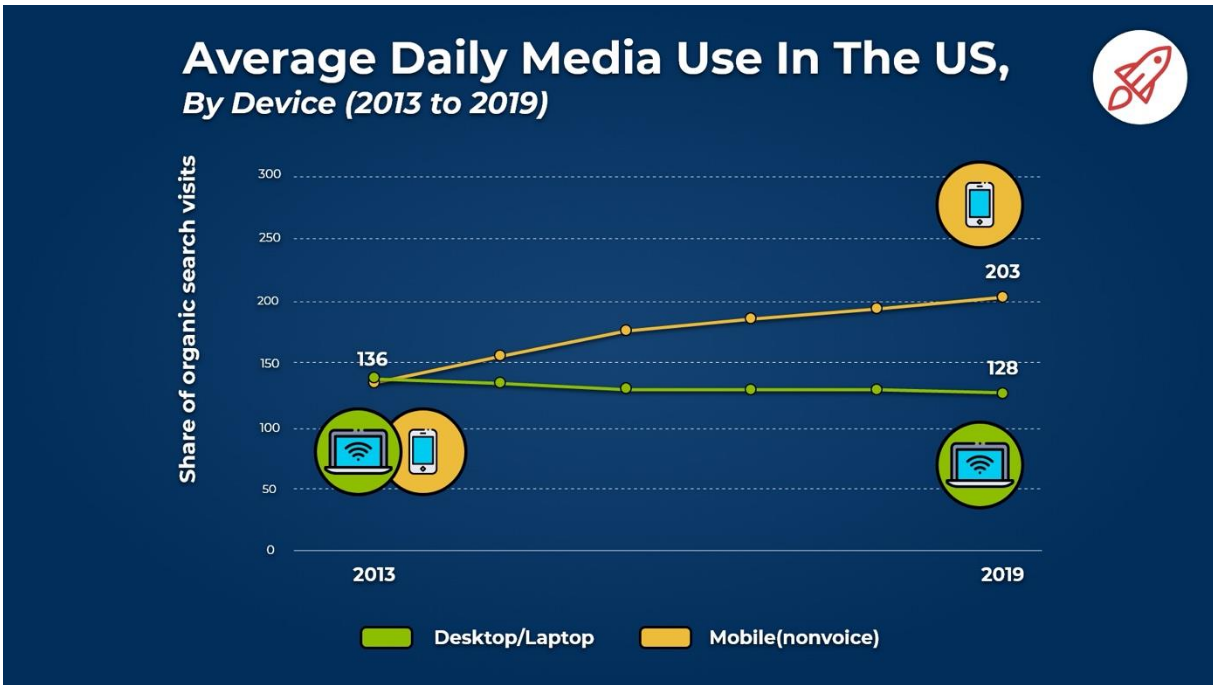


Figure 3. 5: Average daily media use in the US, by device

Source: https://www.broadbandsearch.net/blog/mobile-desktop-internet-usage-statistics

Open source code developer and software educator André Staltz published an article on October 30th, 2017 titled "The Web Began Dying ıi 2014, Here's How". He underlined the role of increasing mobile traffic in the evolving technology world (Figure 3.5). In order to secure customer loyalty, tech giants formed the Apple Store, Google Play and Microsoft Store, which are their background. In this context React technologies gathered speed (Staltz, et al., 2018). Because developing separate software for each ecosystem would affect the cost and time measure and would obligate the different developers who had this capability to work in the same ecosystem. React.js web and React Native brings convenience and security to developers in the same ecosystem by providing appropriate solutions for all mobile platforms.

3.6 PostgreSQL

PostgreSQL is an open-source coded, object-relational database system that protects the complex data masses securely, filters the data, uses Structured Query Language and advances this language. PostgreSQL, which started its path with the POSTGRES project at the University of California in Berkeley in 1986, persists publishing its 13th version on September 24, 2020

The developers who have competence with databases like MySQL and Oracle can easily gain PostgreSQL competence. It is possible to define their data types and form special functions.

Data types stand out with many features such as; standard richness, data collectivity, simultaneity, high performance, security, extraordinary situation saving, JSON and support for international character sets (Korry, et al., 2003).

PostgreSQL, with close to 21% market share is the most written database system in github, where more than 50 million developers produce software. It is used by tech giants such as Apple, BioPharm, Etsy, IMDB, Macworld, Debian, Fujitsu, Red Hat, Sun Microsystem, Cisco, Skype, GitHub, US Navy, NASA, Tesla, Netflix, WeChat, Facebook, Zendesk, Twitter, Zappos, YouTube, Spotify, Microsoft, LinkedIn, PayPal, MasterCard, Autodesk, Sequoia Capital, Atlassian, Optimizely, BNP Paribas, GitLab, Meraki, Delivery Hero, Qubit, LiveRamp and Workable.

It enables the management of millions of preces of data produced in academic and scientific studies (John, et al., 2002). We will use PostgreSQL in our thesis. PostgreSQL comes with PostGIS, which is an add-on that allows the processing of spatial and geographical objects used in geographical information systems that come with a database. For example, it is a technology widely used for vehicle tracking system.

CHAPTER 4

ALGORITHMS

4.1 Machine Learning

Currently through advance technology increase occurs in data storage in data stack and media complexity. In order to make future data meaningful researchers describe these data stacks and Predict data. Machine learning Algorithms through different processes started to describe these analyzes and predictions on large data sets. Machine learning is software of artificial intelligence that enables the system to automatically learn and develop from experiences without any arrangement of programming(Zhongda, et al, 2020). Machine learning focuses on the advancement of computer programs that can reach data and use learning automatically.

The learning method start with research or information such as examples, practice or instruction to observe patterns in the data and make better decisions in the future on the given information sets. The basic goal is to enable computers to gain automatically over stacks of information without human involvement or guidance, and arrange data for future use as required(Alma, et al. 2018).

Although using the standard algorithms of machine learning, text is recognized as a serial of main words; except, a method which is based on semantic analysis mimics the human ability to learn the meaning of a text.

In current thesis, I explained to make it more functional by analyzing the data collected with the software I developed with the machine learning algorithm(Jiaohui, et al, 2020).

4.2 Naive Bayes

Bayes Theorem is a significant subject studied by Thomas Bayes in 1812 in it is included as a conditional probability calculation formula and in probability functions. Bayes' theorem primarily explain the co-relation among conditional probabilities and marginal probabilities in the probability distribution of a random variable. In the diagram below, a formula describing the Bayes theorem can be written as.

$$P(A \mid B) = (\mathrm{P(B \mid A)}\ P(A))/\mathrm{P(B)} \qquad (4.1)$$

P (A | B) : The probability of A being true given that B is true
P (B | A) : The probability of B being true given that A is true
P (A) : The probability of A being true
P (B) : The probability of A being true
P (A | B) = Probability of event A when event B occurs
P (A) = probability of occurrence of A event
P (B | A) = probability of B event occurring when A event occurs
P (B) = probability of event B occurring

Naive Bayes clustering is founded on Bayes' theorem, which has a low learning algorithm and can work on changeable data sets. The method through which the algorithm functions, calculates the probability of each state for a factor or variable and distribute it according to the one with the highest probability result. Very accurate results can be achieved with a little practice data. If a data in the sample set similar to a value not included in the practice set, it returns 0 as a probability value and cannot be estimated. This phenomena is usually express as zero frequency. Researchers use information correction skill to answer this problem. The easiest of these correction skill is known as the Laplace estimation method.

The local Bayesian distribution express the data in different ways that is coded or arranged to the system. The practice data must be classic or categorical. Due to the probability function performed on the given data, the resultant test data entered to the system are evaluated according the previous available and obtained probability values and the according to the category or class in which the presented test data is expressed is explained. If the data is large in number the accuracy of the distribution of the new resultant data will be high.

4.3 Support Vector Machine Algorithm

Support Vector Machine; the foundation of this learning algorithm is based on statistical learning theory. In 1963 this algorithm was discovered by Vladimir Vapnik and Alexey Chervonenkis. First of all it is evaluated in 1963 and then it was established in 1965 by Vladir Vapnik, Berhard Boser and Isbelle Guyon. Thus it is recreated by these researchers and used.

It is one of the very used and simple methods utilized in support vector classification. For classification, a line is drawn among two sets on a plane and used by separating the two sets. This limit should be the furthest from both separated data elements. The rules on how to arrange this limit are selected by the support vector algorithm. For this purpose, two side lines are drawn close to each other and parallel to each other in both sets and a common side line is tried to be inserted by bringing these side lines closer to each other. In figure 4.1 below, we can consider two sets::

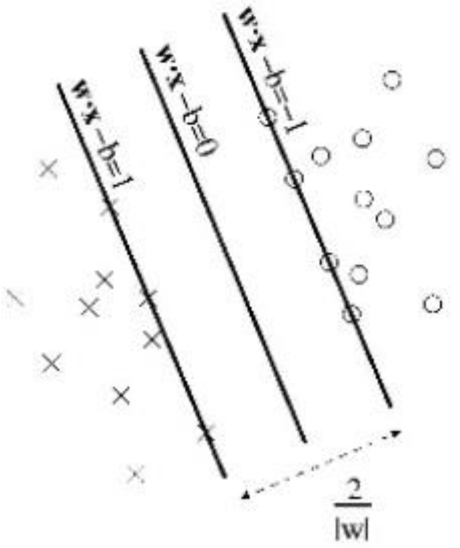


Figure 4. 1: Two groups on a two-dimensional plane

Figure 4.1 display two information sets in a two way data set. Both system and its dimensions are explained as attribute. In simple words, an attribute extraction is made for everyone data set of the system and as a result a totally changed value is obtained in this two dimensional plane, which express each data set. Distribution of these values will ensure that the data sets are accurately distributed according to the predefined attributes. Thus it is expressed as an offset to the interval among both classes and can also ber represented by the formula under for the explanation of each pont in the plane.

In the above given formula: for every x, c pair, Where X is a number in our vector space and c is the number of this point showing that this point is -1 or +1. This group of points starts from i=1 to n. Hence this management refers to the points in the previous figure.

$$D = \left\{(x_i, c_i) \mid x_i \in R^P, C_i \in \{-1, 1\}\right\}_{i=1}^{n} \tag{4.2}$$

Nowadays, For solving many questions researchers are using vector support machine method. The algorithm support vector machine is currently used by many researchers highly in the field of medical and science till now. As an example we can explain that this mechanism is used to explain proteins up to 90% as correctly classified compounds. Permutations tests are based on support vector machine weights have been proposed as a mechanism for interpreting support vector machine models. Hence support vectors weights are used highly, almost in every field to analyze its models.

4.4 Linear Regression

Simple linear regression model is defined as; be a machine learning calculation used to appraise the subordinate variable with the assistance of the autonomous variable when there's a direct relationship among a single free and single subordinate figure. A basic direct relapse is utilized to assess the examination of the straight relationship among two ceaseless factors to calculate the esteem of a subordinate figure based on the number of an free variable.

Actually, linear regression is:

- Express whether the linear regression among two variables is factually noteworthy,

- Express how much of the change in the dependent variable is showed by the independent variable,
- Learn the course and measure of any relationship permits you to appraise the dependent variable values follow by different values of the independent variable,

In order to learn linear regression easily, if we need to show it through a short and concise study, the following study can be found (Figure 4.2).

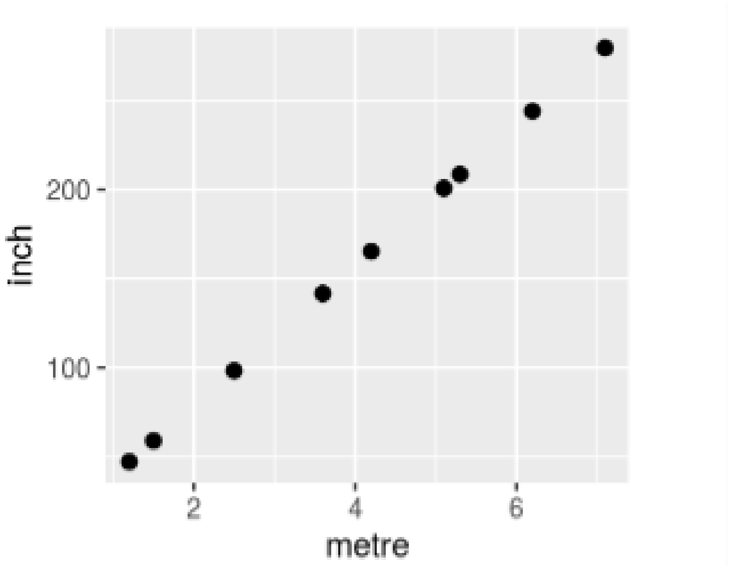


Figure 4. 2: Comparison of Meters and Inches

There's total linearity as over. inches = meters * 39.70 equation applies, Subsequently able to calculate that 10 meters will come to 393.70 inches with the over formula. The precise linear relationships of the units among factors may not be recognized in standard of living. In reality, there's fundamentally randomness. When a linear relationship is observed among variables, the linear regression model is used to predict the future, examine how the variables affect each other and make inferences. This model is comparable to the over model with the basic linear relapse model with the y-axis captured, the incline of the line and the mistake term at the conclusion. The Simple Linear Regression Model having only one independent variable is defined by definition as the Multiple Linear Regression Model having more than one independent parameter.

4.5 Logistic Regression

The method through which categorical or numerical data is classified is known as Logistic Regression. Here in Logistic regression with comparison to linear regression the dependent variable is evaluated through two different values, These value often are yes or no. In this thesis, in the application development phase, in order to get more accurate results Logistic regression classification process is used. Usually researchers utilize this classification algorithm in linear classification studies (Metlek, et al. 2014).

The difference between simple linear regression another method used by researchers and Logistic regression is Logistic. In regression the result will show that the variable is binary or multiple. In parametric model selection and assumption this difference between logistic regression and linear regression is reflected. In logistic regression the variables are described on the basis of two values yes and No while in linear regression analysis all variables are calculated through different values. In addition logistic regression is more simple than linear regression. Although between simple linear regression and Logistic regression there are three specific differences which are given below.

- The dependent variables in logistic regression analysis are discrete in value while the dependent variables in linear regression model are continuous in value.
- As the value of the dependent variable is calculated in liner regression estimation or equation, but in logistic regression model the values that the dependent variable can take is calculated for the probability of realization.
- In linear regression analysis the condition for independent variable is applied that it will express multiple normal distribution, although such type of any condition is not necessary in the case of Logistic regression for Independent variable.

Various studies are available in many categories with respect to the discipline of regression analysis. One of the study among these studies is regarding medicine. This is a study which focuses on examining the factors that affect the birth status of the children. For this research, from hospitals the data of those mothers who gave birth in hospitals were collected and associated factors were created from the data. Regression analysis was performed on the created factors or variables, in addition the relevant parameters were introduced to estimate the factors responsible for the children's birth health and weight.

In the case of logistic regression estimation, the dependent factor is a binary discrete variable such as 0,1; risk is denoted by 1 and the rest by 0. The important value is the result in the regression analysis depending on the values of a provided independent factor is to find the mean value the variable. Moreover this value is known as the conditional mean and denoted by E (Y\ x). Where X is the independent variable and Y is the dependent variable of interest. The conditional mean is assumed to be a linear equation of x in the liner regression model or analysis.

$$\mathrm{E}(Y|x) = \beta_0 + \beta_1 x \tag{4.3}$$

This model express that E (Y \ x) can get any possible value as the interval x changes from 0 and 1. Conditional average in logistic regression analysis must be greater than 0 and less than or equal to 1.

$$0 \leq E(Y|x) \leq 1 \tag{4.4}$$

The left hand side of E(Y\x) = pn + px model gets low probability values among 0-1 in logistic regression analysis although these values are linked with explanatory variables that can get infinite values, to achieve this kind of equality some times not possible. In order to be on the safe side in this type of situation the correct solution is to create the probability values expressed as the result values between –^{o}O and + ^{o}O along various changes. (1,2).

In the estimation of a two level result variable there are various distribution operations available. The very often is used by researchers is logit and profit function. For choosing logistic function there are two main reasons. Firstly in logistic regression analysis the condition of assumption limitation is not present, despite the ease of use, the model is mathematically very flexible obtained from the estimation and the second reason is that it is biologically easy to interpret.

As far as to manage the representation of the formula the values of n (x) - E (Y \ x) can be used to express that the variable Y and X conditional mean is known in the case when logistic distribution is used. The particular form of the logistic regression equation is given below;

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+ e^{\beta_0+\beta_1 x}} \tag{4.5}$$

One function of fl (x) is logit transformation as mentioned above, which will be the main point of our logistic regression analysis. We can draw this transformation in terms of n(x) as follows;

$$g(x) = \left[\frac{\pi(x)}{1- \pi(x)}\right] \tag{4.6}$$

This transformation has all the required characteristics of the linear regression analysis that's why it is very important. Logit g (x) is always linear. It is also continuous in regards of its factors, and can rages among + and – values, depending on the values of x (1).

4.6 K-Means Algorithm

The algorithm through which the researchers perform an unsupervised learning and clustering the information stacks is know is the K-Means algorithm. Where K-Means the total number of clusters along a assumed value of k in the average. Due to this k value an external variable will insert to the algorithm. In reality it is the drawback of this algorithm to take the value of k. Another common algorithm is also available X-Means algorithm that integrates the K value itself. The working procedure of this algorithm is very simple. After selecting the K parameter values determination in the equation the random K main points are considered in the algorithm (Selim, et al. 1984). It estimate or evaluate the interval among each value or information and randomly determine center points and assigns the information to a cluster due to the nearest main point. Again the clustering process is performed for every cluster when the center point is selected, according to the new center point. Until the data sets are completely not stabled the clustering process is performed again and again. Thus the process continue till the stabilization of the all data (Patel, et al, 2020).

This is always problematic for the k means of the algorithm when the central point on the algorithm is selected randomly. For this purpose to avoid this problem, Davir Arthur and Sergei Vassilvitskii created the K-Means algorithm and established the K-MEANS ++ algorithm in order to better select the initial points in 2007. The interval of the selected point from the information are calculated, after selecting a random initial point for the K-Means ++ algorithm. By making certain calculations a new point is selected and the distance is squared then. The condition of this procedure is the radiation analysis O (log k).

The main purpose of the K-means algorithm is to find out to which cluster the new data set belongs, according to more than one property of an extracted data group. The algorithm basically consists of 4 steps. These steps are;

The basic aim of the K-means algorithm is to identify that among the clusters which cluster is belong to the new data set, according to more than one characteristic of an extracted group of data. This algorithm is basically based on four steps. These steps are given below.

1. Creating the center of the cluster.
2. Through distance Reclassification of off-center samples
3. According to new classes defining new centers
4. Until the system is stable, repeating the 2nd and 3rd step for the data.

If we analyze the steps explained above in order of the sample space, the steps as below will take place (Figure 4.3).

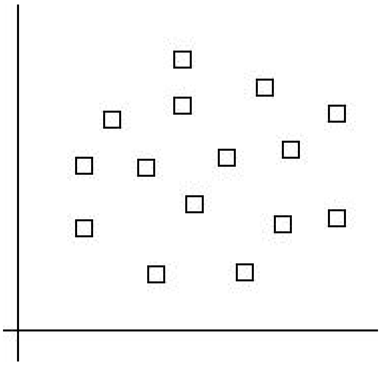


Figure 4. 3: Sample Data Set

At the sample level after determining the sample data set, we define classes for two target sets.

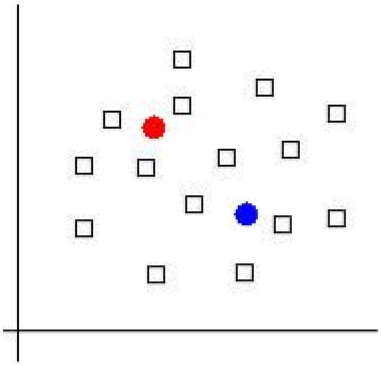


Figure 4. 4: Double Centered Sample Data Sets

We create after this class definition a classification by searching at the intervals of other information according to the main classes defined here (Figure 4.5).

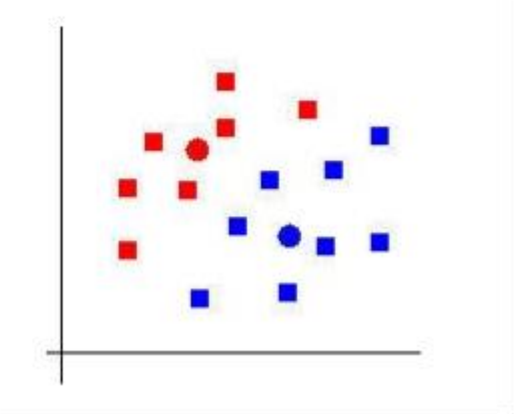


Figure 4. 5: Classified data set

Then we determine a line separating the two classes with the classification formed (Figure 4.6).

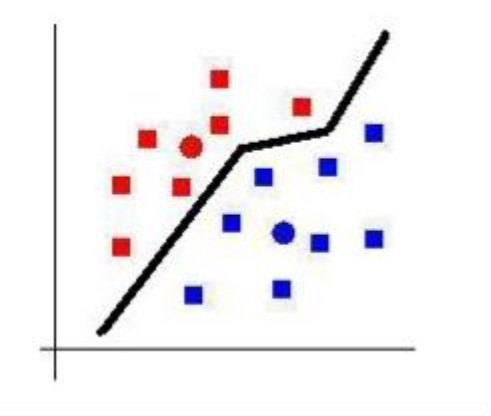


Figure 4. 6: Limitation of two classified data sets I

A few of the samples may be closer to the new focal point, after moving the centers, so we are repeating the old steps to reclassify the sample cluster classifications(Figure 4.7).

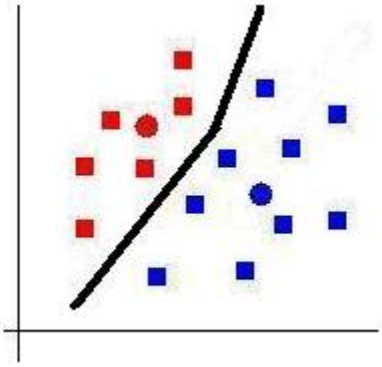


Figure 4. 7: Limitation of two classified data sets II

In many areas the researchers are used the K-Means algorithm. Generally the algorithm is used often in the below given areas and achieves very huge success(Figure 4.8). The the areas in general where the algorithm is frequently used are;

1. Document Classification; According to the labels, subjects and contents of the document clustering is performed. It is very true that this problem is general in the classification of the documents and this algorithm is a valid algorithm for its analysis with the k-means algorithm. Thus it is used very often.

2. Detecting crime scenes; Quality information on crime-prone areas in a city or region can be found from, information on crimes available in particular areas in a city, crime nature, crime location and the association among these two. For this purpose, a classification determined through these relationships can be facilitated through the K-Means algorithm utilization.

3. Customer segmentation; Clustering provides help to marketers by developing their customer base, performance in selected region, and in segmentation of customers through their purchasing behavior history, interests, or practices tracking. Hence according to the classification which is made through using K-Means algorithm, marketers can create their marketing campaign accordance with the organization's policy of sales.

4. Sports analysis; The critical factor of the sport world is to analyze players statistics always and with increasing competition, machine learning has critical role to play here. Now in the analysis performed it is very easy to categorize the struggles of the player and reach the targets required with the K-Mean algorithm.

5. Fraud detection: Machine learning has many uses in automobile, health, insurance and fraud detection and execute a very particular role in fraud detection. Through utilizing old data on fraudulent claims, it may be possible to separate claims based on their proximity t clusters indicating fake patterns.

6. Call record detail analysis; During a customer's call, SMS and internet activity the telecom companies record a call and get the information required known as Call detail record (CDR). The information obtained from call detail record when used with the customer demographic, thus it provide more useful information regarding the needs and wants of the customers. Hence answers to be given through classification of such information may be more effective on the K-Means algorithm.

Moreover using the K-Mean algorithm in many areas such as image processing, apart from the application expressed on the above mentioned items.

CHAPTER 5:

RESULT AND APPLICATIONS

5.1 Introduction

This thesis study develops a business intelligence platform in the form of a mobile application through processing real-time data with machine learning methods.

Today, with the increase of mobile devices and access points of individuals, data are collected on devices. The location of users along with other information is collected in real-time thanks to the mobile applications that have been developed.

5.2 Research Strategy

The mobile application that has been developed in this thesis study is created with the aim of location-based control and mission accomplishment on an individual basis. In this study, two different platforms, namely mobile and web have been planned to be developed. The name of the project has been decided to be YAVER (aide in Turkish) and it aims the missions to be completed by those who want to function as an "aide" and to assign the missions in accordance with location clusters through machine learning. The main reason why this project has been named "Yaver" is that this word means "aide & assistant" in Turkish.

The application called Yaver is planned to have 3 categories. The names of these categories are service, control and survey. All missions to be done are location-based and are presented to users based on a certain algorithm. The mission finishes when the user who assumes the mission informs the user who creates the mission that s/he has completed the mission within the specified deadline. The application that has been developed in this thesis is completely location-based and the aim is to complete the assigned missions.

Such services as setting a software infrastructure that will enable corporate firms to have control mechanisms and that will help the organic user to control the services provided by corporate firms are planned to be amongst future works on this App. Additionally, the control of services provided by public institutions enables organic users to conduct locations based controls and to get necessary feedback. What is more, the algorithm defined within this application will enable to create location-based surveys, to use real users in these surveys and to create requested analysis and graphics. The application that has been developed in this study is a real-time location-based application and presents many services including service, control and survey functions via business intelligence using machine learning categorization algorithms, as well(Vanderplask, et al., 2016).

5.3 General Structure of the System

“The General Structure of the System” shown on the figure displays the running of the software that has been developed. The system is composed of three major sections. The first section is the user interface composed of mobile and PC which presents web and mobile interfaces to the users. Another section is where the server and database are located. This section provides the users with necessary connections to be able to respond to requests coming from interfaces and presents them to users via web services or in form of websites. The last section of the system is the algorithm which is developed by python in which necessary analysis and classifications are made via using machine learning methods to process data collected from users. This python code that has been developed runs via triggering within java codes. The data transferred to the database can be retrieved from the server on the user base and be presented to the users as a web service or html page(Wang, et al., 2020).

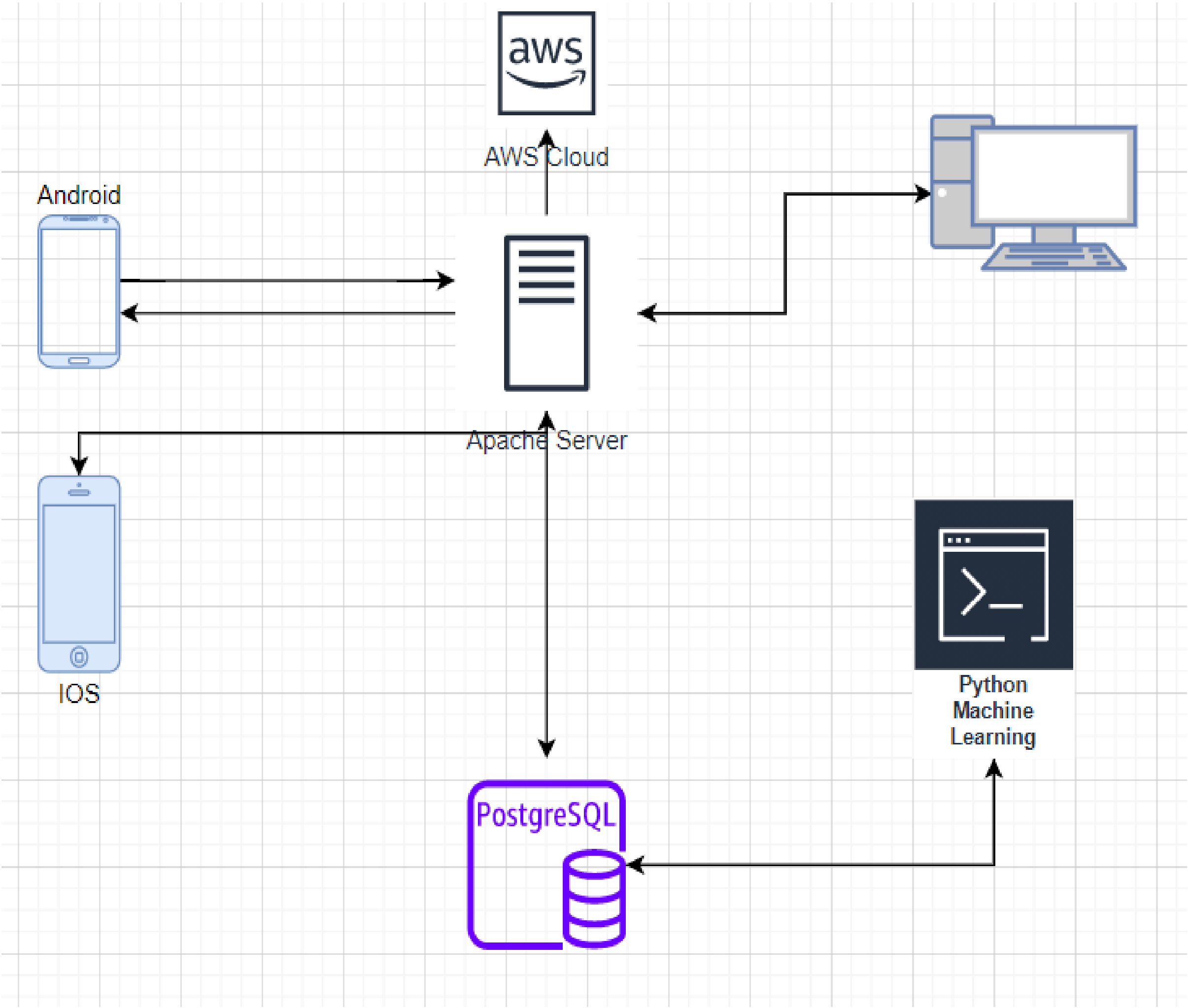


Figure 5. 1: General structure of the system

The following sections of this thesis elaborate on the details of the functions within the general structure of the system (Figure 5.1).

5.4 Software Features of the Application

The interfaces of this application have been designed in two forms, namely as web & Android and IOS. The Web software is mainly designed for system users where condensed data are stored and to be able to make a more complicated analysis. The aim of the mobile software is to offer a more optimized platform in comparison with web software which enables Yaver users to conduct their missions and in which geographical information is collected instantly on this mobile platform (Kresse, et al., 2012). There is a module in the software design which is used for visually mapping the analysis and planning. Apart from this, another database via user-based PostgreSQL database has been created due to the heavy data processing during the designing of the applications. The records of all processes made by the users are stored in the database Distant analysis property has been added via analysis programs set on the server by the system administrator (Xu, et al., 2020).

When developing the interfaces, two different defined application development tools have been used. When developing a Web interface, tools such as Java 8, Springboot, Firebase Notifications, Hibernate, and Amazon Web Service library have been used. The interface designs used by the system administrator have been designed by using React.js library. Additionally, developing web services to respond to the needs of the software that has been developed has been made via this tool. The tool that has been developed by the web software module has enabled the coding and collecting the codes of web interfaces and background code.

Another module, Yaver mobile application has been developed both for IOS and Android devices by using React Native. Additionally, python has been used for geographic mission distribution through machine learning methods (Matwin, et al., 1998). Python has been preferred since it includes lots of third-party libraries for data processing and analysis.

5.5 Features of Database Module

The database module is a module that is jointly used by software modules that have been developed and that stores and administrates all data. Postgresql has been selected for the database. The reason why this database has been selected is that it has open resources and a flexible structure. Additionally, Postgresql displays a better performance in comparison with other databases in storing big data and in processing indexes and searches. This module includes information of users and tables which displays parameters to be used by the application along with the results of the analysis. The tables on the database have been developed via PGadmin tool.

We needed to use a library called Postgresql Connector so that the tables developed through PGadmin can be accessible by applications developed by java. Additionally, the connection is secured through making configurations via Spring Framework. This library can be accessed through the website of Postgresql Oracle. The library that you may access through Oracle hosts JDBC libraries used by java. Since library administration is made through the maven connection administrator on the application, JDBC is automatically added to the library. The below-stated code block shows the connection between java and PostgreSQL.

```java
public static Connection getDBConnection() throws SQLException,
InstantiationException, IllegalAccessException,
ClassNotFoundException {
Class.forName("com.mysql.jdbc.Driver");
Connection connection
DriverManager.getConnection("jdbc:postgresql://"+PATH_TO_DB,
USERNAME,PASS);
connection.setAutoCommit(false);
return connection;
}

```

The connection is not directly written this way within the application. Since the application uses Springboot, necessary parameters have been written on the properties file and it automatically defines the parameters included in the above-stated code. The below code line, on the other hand, shows the definition within Springboot (Figure 5.2).

```
## Spring DATASOURCE (DataSourceAutoConfiguration & DataSourceProperties)
spring.datasource.url=jdbc:postgresql://localhost:5432/yaver
spring.datasource.username= postgres
spring.datasource.password=root
```

Figure 5. 2: Definition of springboot

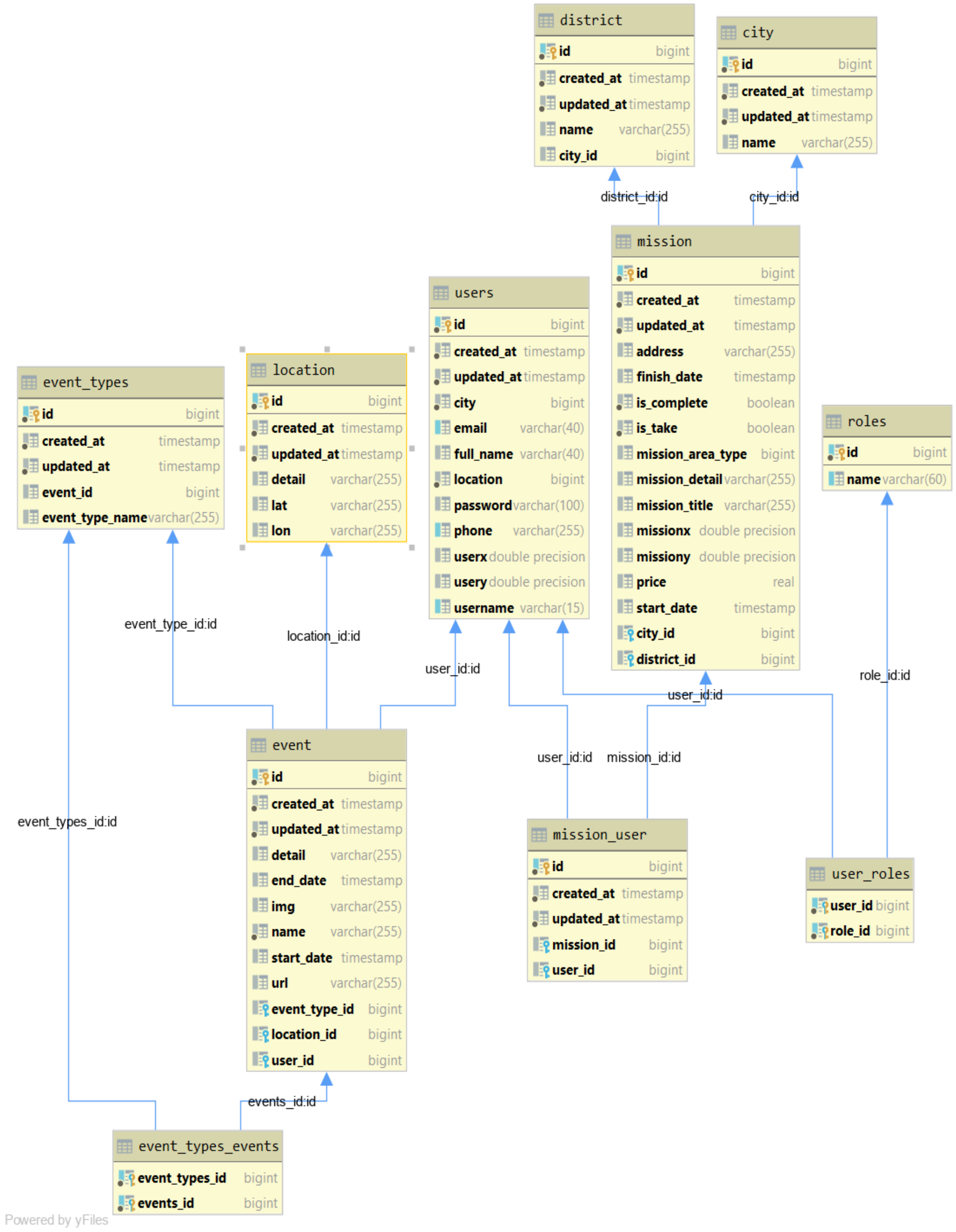


Figure 5. 3: UML classification diagram of the structure of the database

The above figure shows the UML diagram of the application that has been developed. You may kindly find below information on the tables that have been included in the database (Figure 5.3).

1) Mission: The table that shows the information regarding missions that are created on the application.
2) Mission User: The table that shows which user has selected the created mission.
3) Users: The table which shows information about the users.
4) Location: The table which shows information regarding the location to be used on the application on x and y coordinates.

5.6 Features of User Interface

The application to be presented to the users should have an ergonomic interface and should be user friendly. The interfaces of the application created for this thesis study has been developed by using different optimizations. User interfaces have been created by interactive system design architecture. Two different interfaces, namely for web and mobile, have been developed as part of this thesis.

5.6.1 Design of mobile interface

By using react native language, native based mobile applications which are suitable both for IOS and Android operating systems have been developed. Screens where users can create missions and manage the time for the completion of the missions have been created for the mobile application. Upon classification made by the machine learning for the management and distribution of the missions, users receive a notification. Mobile application includes the below stated sections.

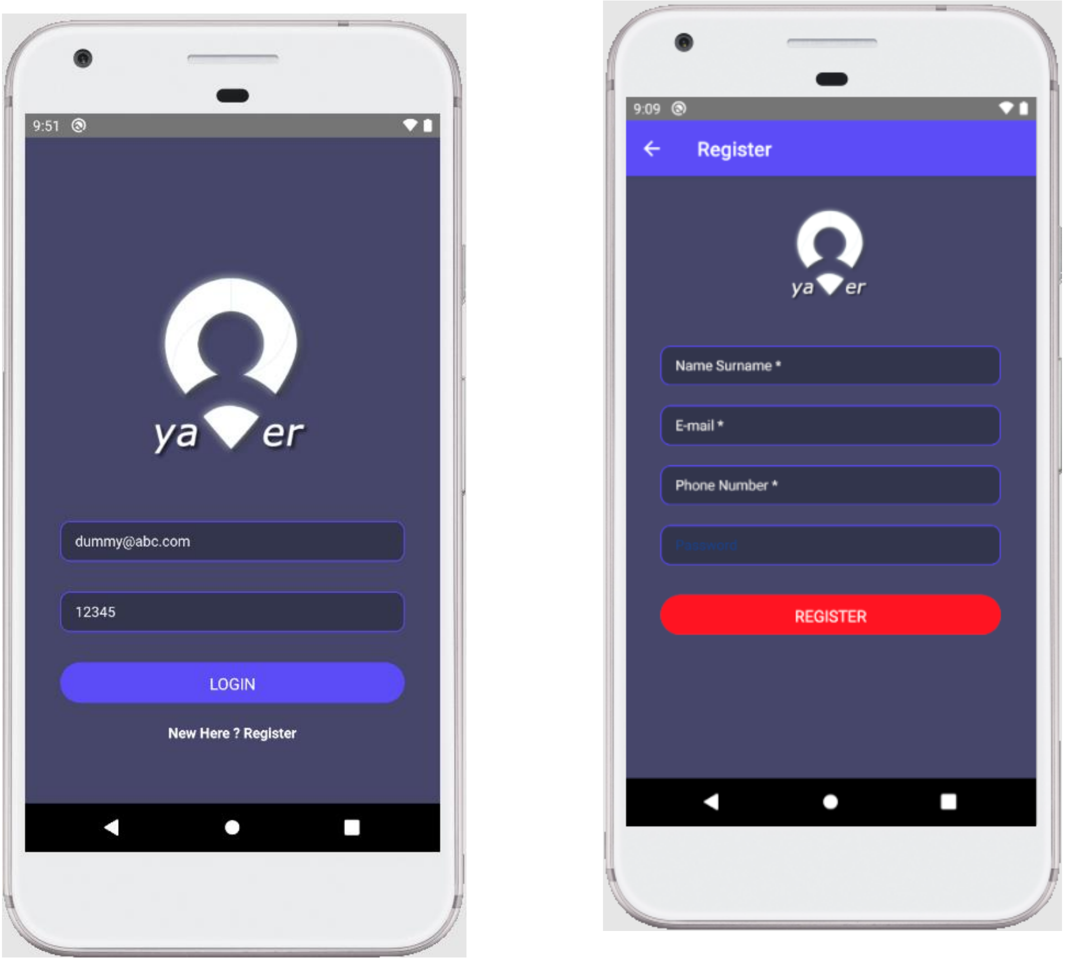
These are as follows:

- Form for Creating a Mission
- Mission List
- Mapping Screen
- Completed Mission List

- Screen for Time Management for Mission Completion
- Management Screen for User Settings

The below stated picture shows the screens for login and sign up sections of the mobile application. Before being able to login, the user must first sign up and authenticate through email. Login is only possible through an authenticated email and entering the password which is secured by token by backend. Java Web Token is used to ensure security on background code. Upon login by the user, it is registered on the database by generating a token. This token provides sending the necessary notification to the user.

The screen of the mobile application has been designed as shown on the below picture and the background code has been written based on the methods shown on the below code block. There are two methods here. One of them is the post method of the web service which is necessary for signing up a user. The other method, on the other hand, ensures the verification of user login and creates token subsequently. These methods are under AuthController class (Figure 5.4).



dummy@abc.com

12345

Figure 5. 4: Sign up and login screens of yaver

```
    @PostMapping("/signin")
    public ResponseEntity<?> authenticateUser(@Valid @RequestBody LoginRequest loginRequest) {

        Authentication authentication = authenticationManager.authenticate(
                new UsernamePasswordAuthenticationToken(
                        loginRequest.getUsernameOrEmail(),
                        loginRequest.getPassword()
                )
        );

        SecurityContextHolder.getContext().setAuthentication(authentication);

        Optional<User> userOpt = userRepository.findByEmail(loginRequest.getUsernameOrEmail());

        String jwt = tokenProvider.generateToken(authentication);
        return ResponseEntity.ok(new JwtAuthenticationResponse(jwt,userOpt.get().getId()));
    }

    @PostMapping("/signup")
    public ResponseEntity<?> registerUser(@Valid @RequestBody UserDto signUpRequest) {
        if(userRepository.existsByUsername(signUpRequest.getEmail())) {
            return new ResponseEntity(new ApiResponse(false, "Username is already taken!"),
                    HttpStatus.BAD_REQUEST);
        }

        if(userRepository.existsByEmail(signUpRequest.getEmail())) {
            return new ResponseEntity(new ApiResponse(false, "Email Address already in use!"),
                    HttpStatus.BAD_REQUEST);
        }

        if(userRepository.existsByPhone(signUpRequest.getPhone())){
            return new ResponseEntity(new ApiResponse(false,"Phone already in use!"),HttpStatus.BAD_REQUEST);
        }

        // Creating user's account
        User user = new User(signUpRequest.getFullName(), signUpRequest.getEmail(),
                signUpRequest.getEmail(),signUpRequest.getPhone(),0L,
                0L, signUpRequest.getPassword(),signUpRequest.getUserX(),signUpRequest.getUserY());

        user.setPassword(passwordEncoder.encode(user.getPassword()));

        Role userRole = roleRepository.findByName(RoleName.ROLE_USER)
                .orElseThrow(() -> new AppException("User Role not set."));

        user.setRoles(Collections.singleton(userRole));

        User result = userRepository.save(user);

        URI location = ServletUriComponentsBuilder
                .fromCurrentContextPath().path("/api/users/{username}")
                .buildAndExpand(result.getUsername()).toUri();

        return ResponseEntity.created(location).body(new ApiResponse(true, "User registered successfully"));
    }

```

The below stated code line which is developed by react native shows the method which verifies the user via server application upon the entry of user information on the screen and which shows the token processes received by the application.

```
const LoginScreen = props => {
  let [userEmail, setUserEmail] = useState('');
  let [userPassword, setUserPassword] = useState('');
  let [loading, setLoading] = useState(false);
  let [errortext, setErrortext] = useState('');

  const handleSubmitPress = () => {
    setErrortext('');
    if (!userEmail) {
      alert('Please fill Email');
      return;
    }
    if (!userPassword) {
      alert('Please fill Password');
      return;
    }
    setLoading(true);
    var dataToSend = { usernameOrEmail: userEmail, password: userPassword };
    var formBody = [];
    for (var key in dataToSend) {
      var encodedKey = encodeURIComponent(key);
      var encodedValue = encodeURIComponent(dataToSend[key]);
      formBody.push(encodedKey + '=' + encodedValue);
    }
    formBody = formBody.join('&');

    fetch(ApiUrl.baseUrl + '/auth/signin', {
      method: 'POST',
      headers: {
          'Content-Type': 'application/json'
      },
      body: JSON.stringify(dataToSend)
    }).then(response => response.json())
      .then(responseJson => {
        //Hide Loader
        setLoading(false);
        console.log(responseJson);
        // If server response message same as Data Matched
        if (responseJson.accessToken) {
          AsyncStorage.setItem('accessToken', responseJson.accessToken)
          console.log(responseJson.accessToken);
          props.navigation.navigate('DrawerNavigationRoutes');
        } else {
          setErrortext('Please check your email id or password');
          console.log('Please check your email id or password');
        }
      })
      .catch(error => {
        //Hide Loader
```

```
            setLoading(false);
            console.error(error);
        });
    };

```

In Figure 5.5, we may see the main screen which appears after the user logs in on the mobile application and the menu of the app. User can see the open missions on the main screen. He or she may assume these tasks if s/he deems appropriate for himself/herself. Additionally, once the user logs in, then the application can send him/her notifications. The menu shows the tasks that the user can do depending on his/her authorization.

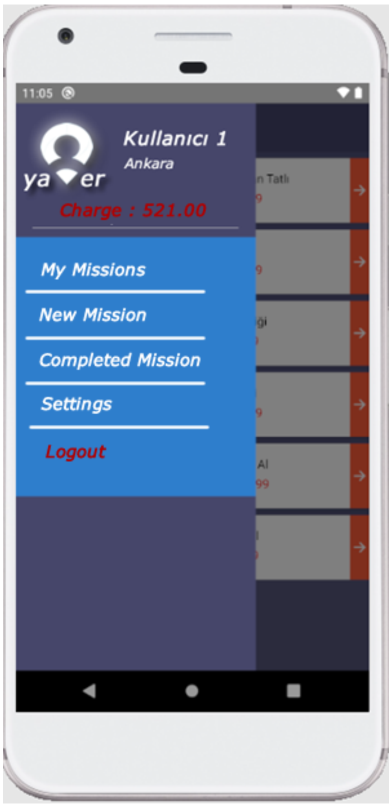


Figure 5. 5: User menu and main screen

Figure 5.6 shows the screen where the information regarding the mission that the user wants to assume will be filled. The user enters the user information on this form. The most important information that needs to be filled in at this stage is the information regarding the location of the mission to be completed. Users can determine the location from GPS instantly through the mobile device or s/he can select the location from the map shown in figure 5.7. Entering the location is necessary for location-based classification which is made by a machine learning classification algorithm within the background code. Therefore, the location should be exact and true. Additionally, the form requires the entry of the open address of the task to be completed. This open address is used for the required data analysis. Other Yaver users receive a notification based on the location which has been determined either through GPS or selecting on the map. In so doing, the completion of the mission in a timely and appropriate manner is secured.

```
@PostMapping("/add/{userId}")
  public ResponseEntity<?> addMission(@RequestBody MissionDto missionDto,@PathVariable Long userId) {

      System.out.println(missionDto);

      missionService.addMission(missionDto);

      HttpHeaders responseHeaders = new HttpHeaders();
      //responseHeaders.setLocation(location);
        responseHeaders.set("MyResponseHeader", "MyValue");
        return new ResponseEntity<String>("Hello World", responseHeaders, HttpStatus.OK);
    }

```

On the above stated code line, we may see the web service endpoint which is called when the user fills the form to complete a mission. Then, it is recorded on the database table after the authorization and information of the user is checked. Additionally, the validation of the areas is controlled at this stage.

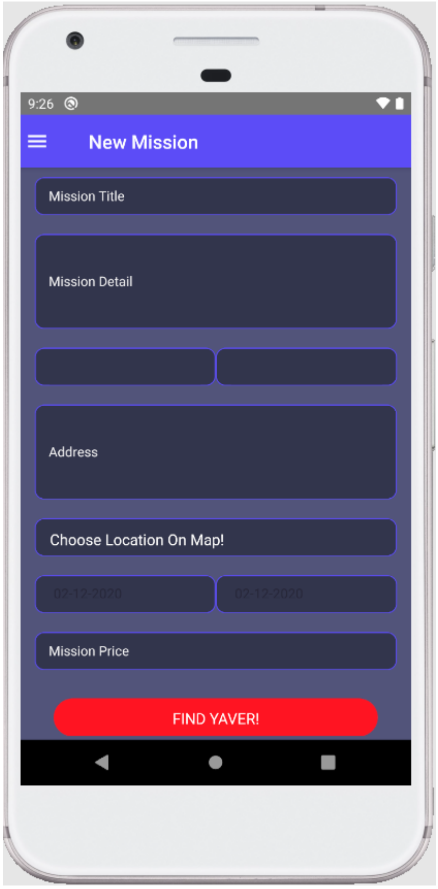


Figure 5. 6: Mission opening form

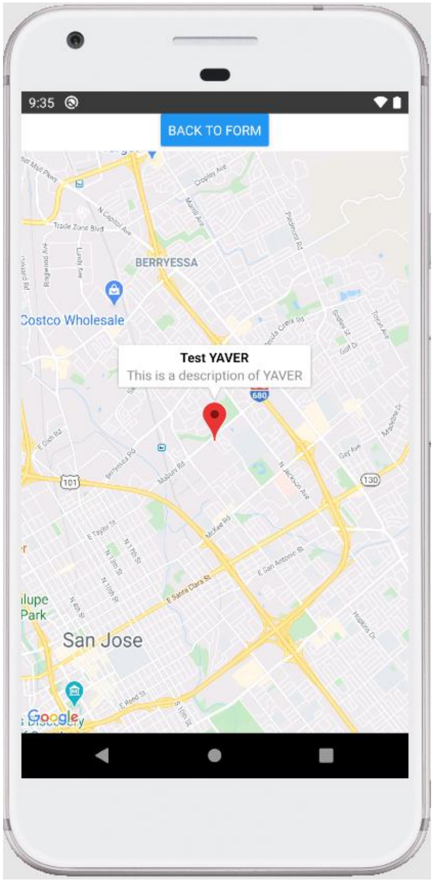
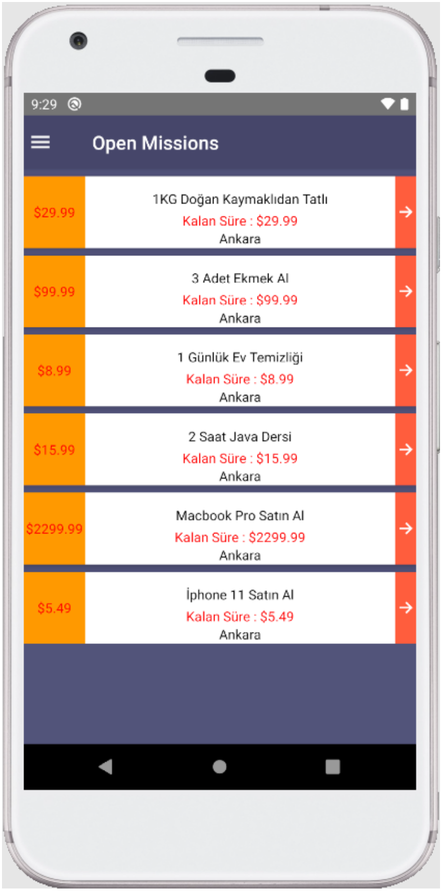


Figure 5. 7: Choosing a location on the Map

In Figure 5.6, we may see the open missions. On this mission list, the missions that are opened by the users who have not yet been assumed or completed by other users called Yaver are displayed. The missions on the list are updated throughout the mission deadline. Once the deadline of the mission expires, the mission automatically disappears from the mission list. The list appears on the screen of the user through an algorithm that is run within the background code in accordance with the instant GPS information of the user.

In Figure 5.7, we can see the details of the mission once the user clicks son a mission on the list or open the notification s/he receives. On this screen, the user is expected to assume a mission in accordance with his/her preference. The user who assumes a mission is expected to finish a task within the time specified. Once the mission is completed, the payment is made automatically. However, the Yaver user should follow the necessary process to complete a task because s/he needs to strictly adhere to the process which will secure the duly completion of the mission which is assigned(Figure 5.8, 5.9).

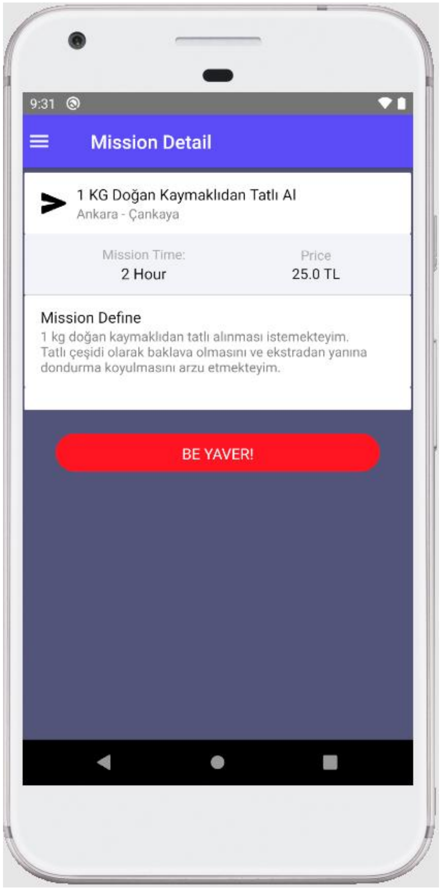


9:29

Open Missions

$29.99	1KG Doğan Kaymaklıdan Tatlı Kalan Süre : $29.99 Ankara	→
$99.99	3 Adet Ekmek Al Kalan Süre : $99.99 Ankara	→
$8.99	1 Günlük Ev Temizliği Kalan Süre : $8.99 Ankara	→
$15.99	2 Saat Java Dersi Kalan Süre : $15.99 Ankara	→
$2299.99	Macbook Pro Satın Al Kalan Süre : $2299.99 Ankara	→
$5.49	İphone 11 Satın Al Kalan Süre : $5.49 Ankara	→

Figure 5. 8: Mission list



9:31

Mission Detail

1 KG Doğan Kaymaklıdan Tatlı Al
Ankara - Çankaya

Mission Time:	Price
2 Hour	25.0 TL

Mission Define
1 kg doğan kaymaklıdan tatlı alınması istemekteyim. Tatlı çeşidi olarak baklava olmasını ve ekstradan yanına dondurma koyulmasını arzu etmekteyim.

Figure 5. 9: Mission detail

```
@GetMapping("/getMissionDetail")
    public MissionDto getMissionDetail(@RequestParam Long missionId) {

        try{
            MissionDto missionDto = missionService.getMissionDetailById(missionId);
            return missionDto;

        }
        catch (Exception ex)
          {
              System.out.println("Error getMissionDetail"+ex.getMessage());
          }
          return null;
    }

    /*
     */
    @PostMapping("/takeMission/{userId}")
    public ResponseEntity<?> takeMission(@RequestParam Long missionId,@PathVariable Long userId) {

          try {
              MissionUser missionUser = missionService.takeMissionByUser(missionId,userId);

              HttpHeaders responseHeaders = new HttpHeaders();
              //responseHeaders.setLocation(location);
              responseHeaders.set("MyResponseHeader", "MyValue");

              return new ResponseEntity<String>("Success", responseHeaders, HttpStatus.OK);
          }catch (Exception ex)
          {
              System.out.println(ex.getMessage());
              HttpHeaders responseHeaders = new HttpHeaders();
              //responseHeaders.setLocation(location);
              responseHeaders.set("MyResponseHeader", "MyValue");
              return new ResponseEntity<String>(ex.getMessage(), responseHeaders, HttpStatus.OK);
          }

    }

```

On the above stated code line, we may see the endpoint on the details of a mission based on an ID and the methods of the processes to record the information of the user to assume a task on the database.

5.7 Background Coding Features of the Application

The coding of the application that has been developed has been made by using Java Enterprise Edition (J2EE), Spring Framework, React Native Mobile, Python Data Mining Modules and web service technologies in accordance with MVC architecture. The system is composed of five main structures that are interconnected. The mentioned five main structures are the mobile application, web server application, the database module, web result application and the application on which classification algorithm is made(Işık, etal., 2006).

In addition to the main structures, there is also a security wall that runs integrated with the system in order to provide security for the application. The security wall prevents the access of the user information and mission details by the third parties and the possible trade loss which may stem from any leak(Kaufman, et al., 1990).

The thesis is composed of two blocks in accordance with the system functioning scheme displayed in Figure 5.10. The first block is called the business layer which administrates all missions and related tasks. The business layer runs and business logic algorithm and records on the database via server application located on the middle layer in accordance with the task and the geographical locations provided. The block which is highlighted with red is the application that has been developed by using python language for k-mean classification algorithm for data regarding the coordinates and machine learning. Postgresql database which is used in the application serves both blocks. Additionally, the application handles the request and response processes completely through web service.

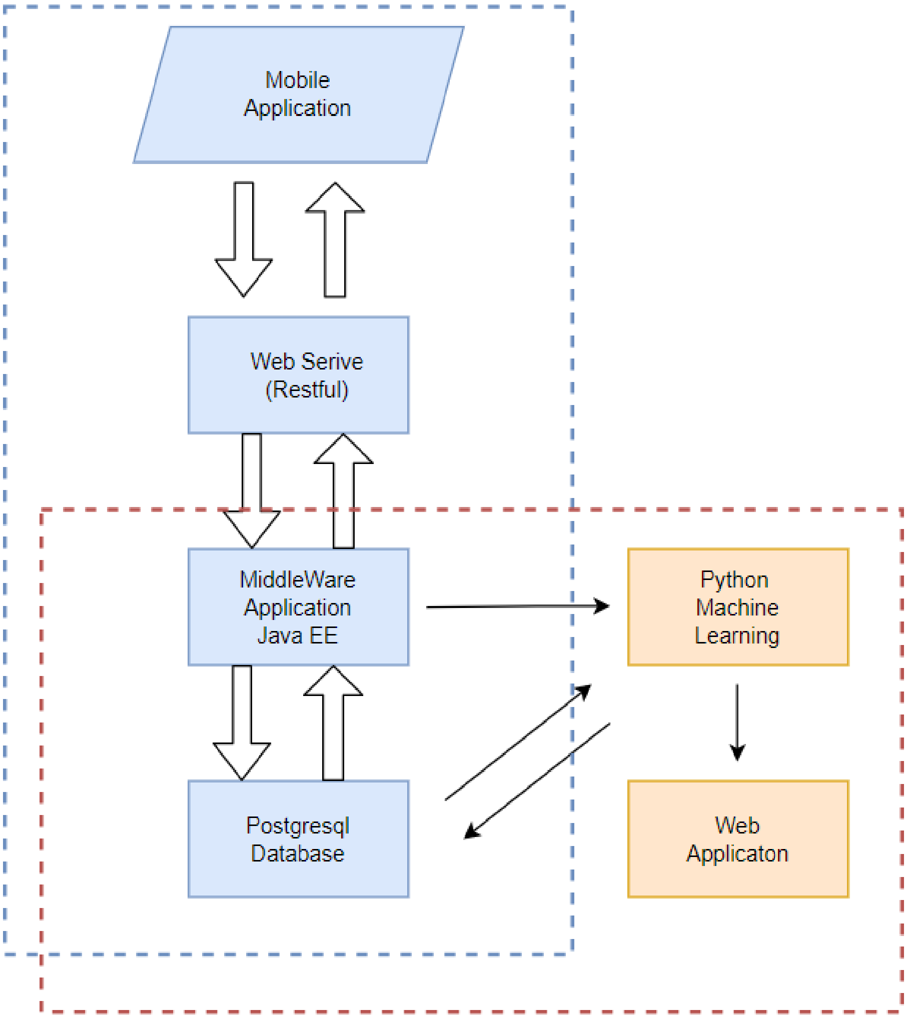


Figure 5. 10: Classification Based On Geographical Data

Inferring new meaning through the data collected on the application through machine learning and classifying them in accordance with certain parameters are amongst the topics which are found interesting by the researchers (Figure 5.10). We may ensure the effective and efficient running of the application through the classification made based on the data collected. The data that has been collected via the application called Yaver has been classified by using python k-means algorithm. In the running process of K-means algorithm, K number object is randomly selected to determine the center point of each cluster and their mean. The resting objects, on the other hand, are included in similar clusters in accordance with their distance to the centroid point. Subsequently, the mean value of each and every cluster is calculated, and the centroid points of the clusters are determined. The algorithm continues until there is no change. The main data in this classification is the data on geographical location. Clustering is made on the missions assumed and completed by Yaver users. A study is made on the missions and the missions completed by the user through a classification algorithm developed by Python. This study is on finding who will be a suitable candidate for the next mission and who is competent on that mission (Jain, et al., 1988).

A sample data set has been used on the Yaver application while developing the classification algorithm. The sample data set is collected on data from Ankara province.

```
%matplotlib inline
import pandas as pd, numpy as np, matplotlib.pyplot as plt
from scipy.cluster.vq import kmeans, kmeans2, whiten
import psycopg2

```

On the above code line, the identification of the modules to be used in python is made. On the below code line, on the other hand, shows the connection of the Yaver application on the database that is created by Postgresql.

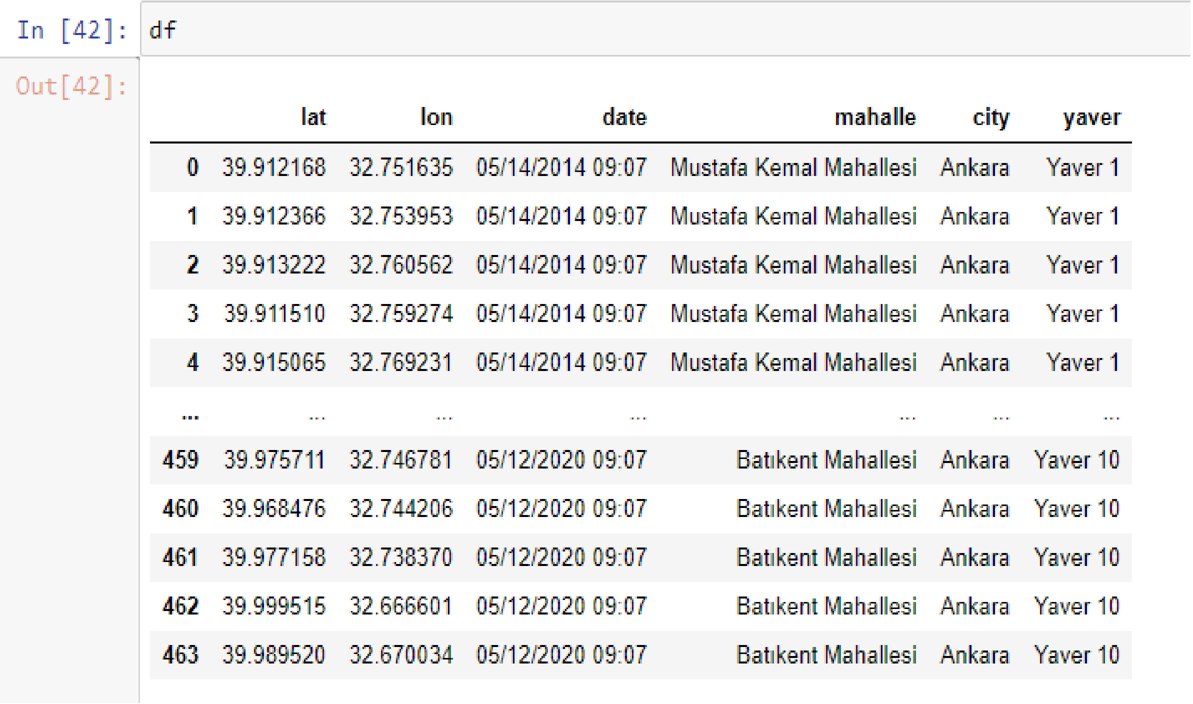
```
connection = psycopg2.connect(user = "postgres",
                                  password = "root",
                                  host = "127.0.0.1",
                                  port = "5432",
                                  database = "yaver")

```

The code that is written in order to retrieve data from the table which records the geographical data is shown below(Table 5.1):

```
qSelect = "SELECT m.missionX as lat, m.missionY as lon, m.start_date as date, a.mahalle,a.city,  FROM mission m inner join mission_user on mu m.id=mu.mission_id inner join address a on m.address_id=a.id"
cursor.execute(qSelect)
df = cursor.fetchall()
df.head()
```

```

```

Table 5. 1: Geographical data space



```
In [42]: df
```

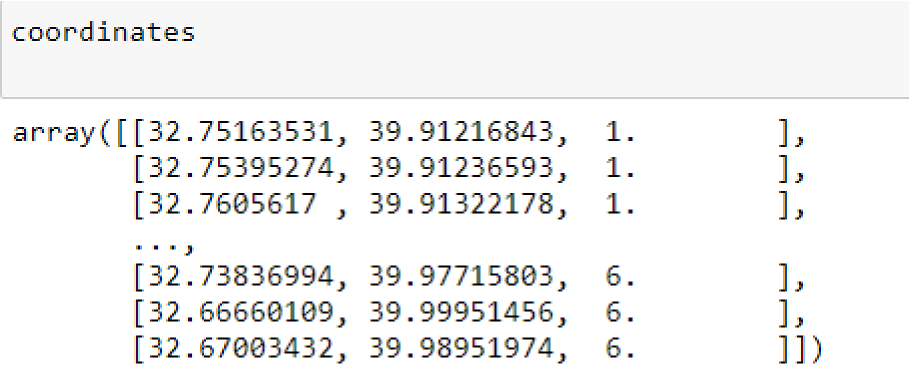
Out[42]:

	lat	lon	date	mahalle	city	yaver
0	39.912168	32.751635	05/14/2014 09:07	Mustafa Kemal Mahallesi	Ankara	Yaver 1
1	39.912366	32.753953	05/14/2014 09:07	Mustafa Kemal Mahallesi	Ankara	Yaver 1
2	39.913222	32.760562	05/14/2014 09:07	Mustafa Kemal Mahallesi	Ankara	Yaver 1
3	39.911510	32.759274	05/14/2014 09:07	Mustafa Kemal Mahallesi	Ankara	Yaver 1
4	39.915065	32.769231	05/14/2014 09:07	Mustafa Kemal Mahallesi	Ankara	Yaver 1
...	...	...	...	...	...	...
459	39.975711	32.746781	05/12/2020 09:07	Batıkent Mahallesi	Ankara	Yaver 10
460	39.968476	32.744206	05/12/2020 09:07	Batıkent Mahallesi	Ankara	Yaver 10
461	39.977158	32.738370	05/12/2020 09:07	Batıkent Mahallesi	Ankara	Yaver 10
462	39.999515	32.666601	05/12/2020 09:07	Batıkent Mahallesi	Ankara	Yaver 10
463	39.989520	32.670034	05/12/2020 09:07	Batıkent Mahallesi	Ankara	Yaver 10

The connection is made through psycopyg2 module in order to retrieve data from the PostgreSQL database via Python. When the geographical data is called on the python code, the sample space table 5.2 of the data on the parameter list is as follows:

```
coordinates = df[['lon', 'lat',’id’]].to_numpy()

```

Table 5. 2: Set of coordinates



```
coordinates
```

```
array([[32.75163531, 39.91216843,  1.        ],
       [32.75395274, 39.91236593,  1.        ],
       [32.7605617 , 39.91322178,  1.        ],
       ...,
       [32.73836994, 39.97715803,  6.        ],
       [32.66660109, 39.99951456,  6.        ],
       [32.67003432, 39.98951974,  6.        ]])
```

A set of lines has been made which includes x, y and yaver id data by identifying the coordinates parameter to be used to classify lat,lon,date,mahalle,city and yaver info retrieved from the database.

```
most_index = df['mahalle'].value_counts().head(6).index
most = pd.DataFrame(df[df['mahalle'].isin(most_index)])
most.drop_duplicates(subset=['mahalle'], keep='first',
inplace=True)

plt.figure(figsize=(10, 6), dpi=100)
co_scatter = plt.scatter(coordinates[:,0], coordinates[:,1], c='b',
edgecolor='', s=15, alpha=0.3)

plt.title('Scatter plot of the full set of GPS points')
plt.xlabel('Longitude')
plt.ylabel('Latitude')

for i, row in most.iterrows():
    plt.annotate(row['mahalle'],
                 xy=(row['lon'], row['lat']),
                 xytext=(row['lon'] + 1.5, row['lat'] + 0.6),
                 bbox=dict(boxstyle='round', color='k', fc='w',
alpha=0.6),
                 xycoords='data',
                 arrowprops=dict(arrowstyle='->',
connectionstyle='arc3,rad=0.5', color='k', alpha=0.8))

plt.show()

```

The graphic that shows the distribution of the two-dimensional coordinate data created by the code line on the XY coordinates system is made by using matplotlib.pyplot module. The x, y geographic data graph is shown in Figure 5.11.

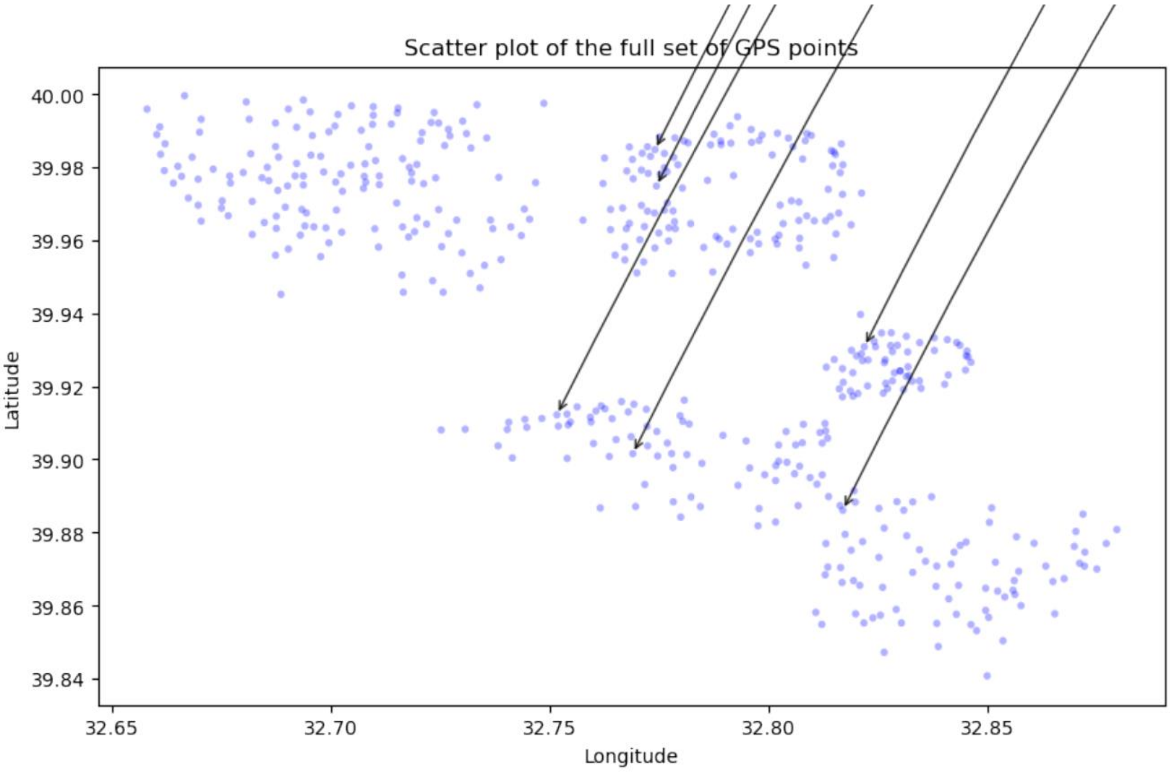


Figure 5. 11: The distribution of missions on geographical coordinates

```
N = len(coordinates)
w = whiten(coordinates)
k = 50
i = 1000
cluster_centroids1, distortion = kmeans(w, k, iter=i)
fig=plt.figure(figsize=(10, 10), dpi=100)
ax = fig.add_subplot(111, projection='3d')
ax.scatter(cluster_centroids1[:,0],
cluster_centroids1[:,1],cluster_centroids1[:,2], c='m', marker='o')
ax.set_xlabel('Lat Centroid')
 ax.set_ylabel('Lon Centroid')
 ax.set_zlabel('YAVER ID')
 plt.show()
```

After the coordinates list is created, the length of the list is taken in order to conduct a classification on kmeans algorithm. Then, the set of coordinates has been normalized via the whiten function. The number of necessary classifications for Kmeans classification algorithm is defined as (k) and the number of iteration on the data is defined as (i). 1000 iteration has been conducted throughout this thesis study. The result of the classification can be more meaningful and proximate when the number of iteration is increased. The centroid point of the classified geographical data is determined via K-means algorithm. The graphic design of the centroid points is displayed below (Figure 5.12):

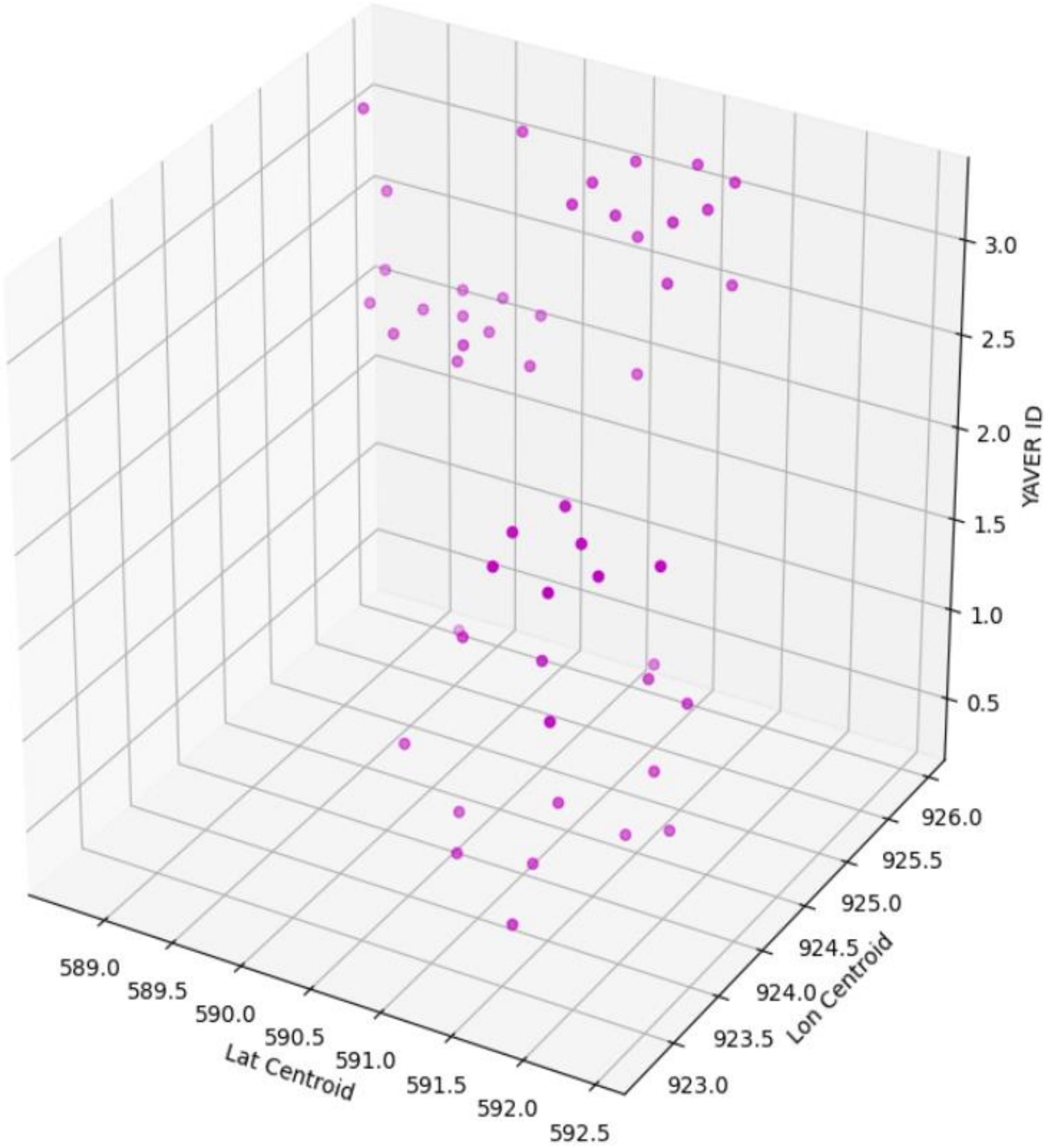


Figure 5. 12: Centroid points with K-means first classification

```
cluster_centroids2, closest_centroids = kmeans2(w, k, iter=i,
minit='points')

fig=plt.figure(figsize=(10, 10), dpi=100)
ax = fig.add_subplot(111, projection='3d')
```

```
ax.scatter(cluster_centroids2[:,0],
cluster_centroids2[:,1],cluster_centroids2[:,2], c='r', marker='o')
ax.scatter(w[:,0], w[:,1],w[:,2], c='k')
ax.set_xlabel('Lat Centroid')
ax.set_ylabel('Lon Centroid')
ax.set_zlabel('YAVER ID')
 plt.show()

```

We want the Kmeans2 function to make researches in accordance with the centroid point determined as 50 by using k-means algorithm additionally, by making 1000 iterations on the algorithm as discussed above, the algorithm will be able to research the closest points and detect nearby centroid points. The comparison of the results it finds and the data that we have are presented in the below Figure 5.13:

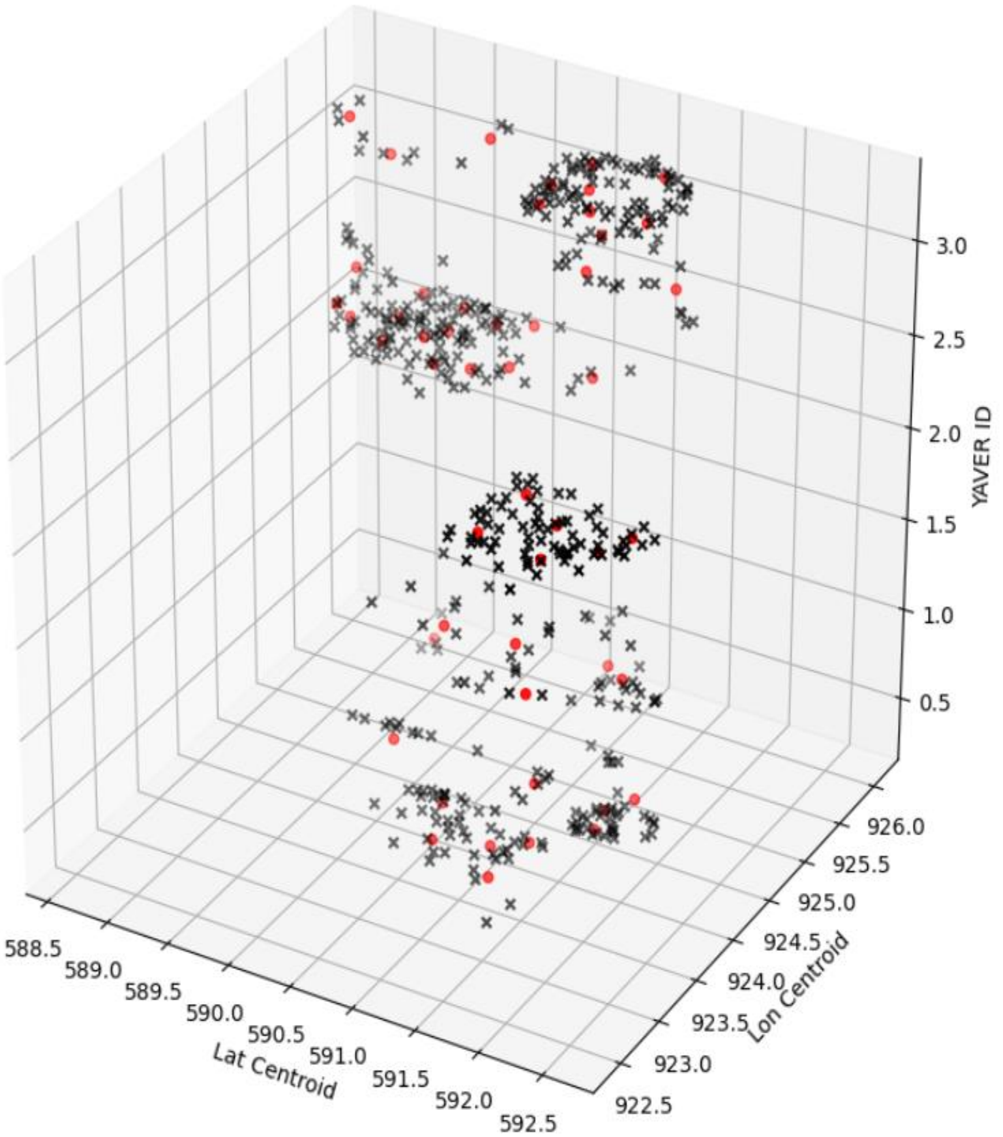


Figure 5. 13: Comparison of the data with centroid

Figure 5.14 shows the colored sample data set retrieved from the database in accordance with the emerged close centroids via the running of K-means algorithm. The colors have been randomly chosen in accordance with their distance to the centroid.

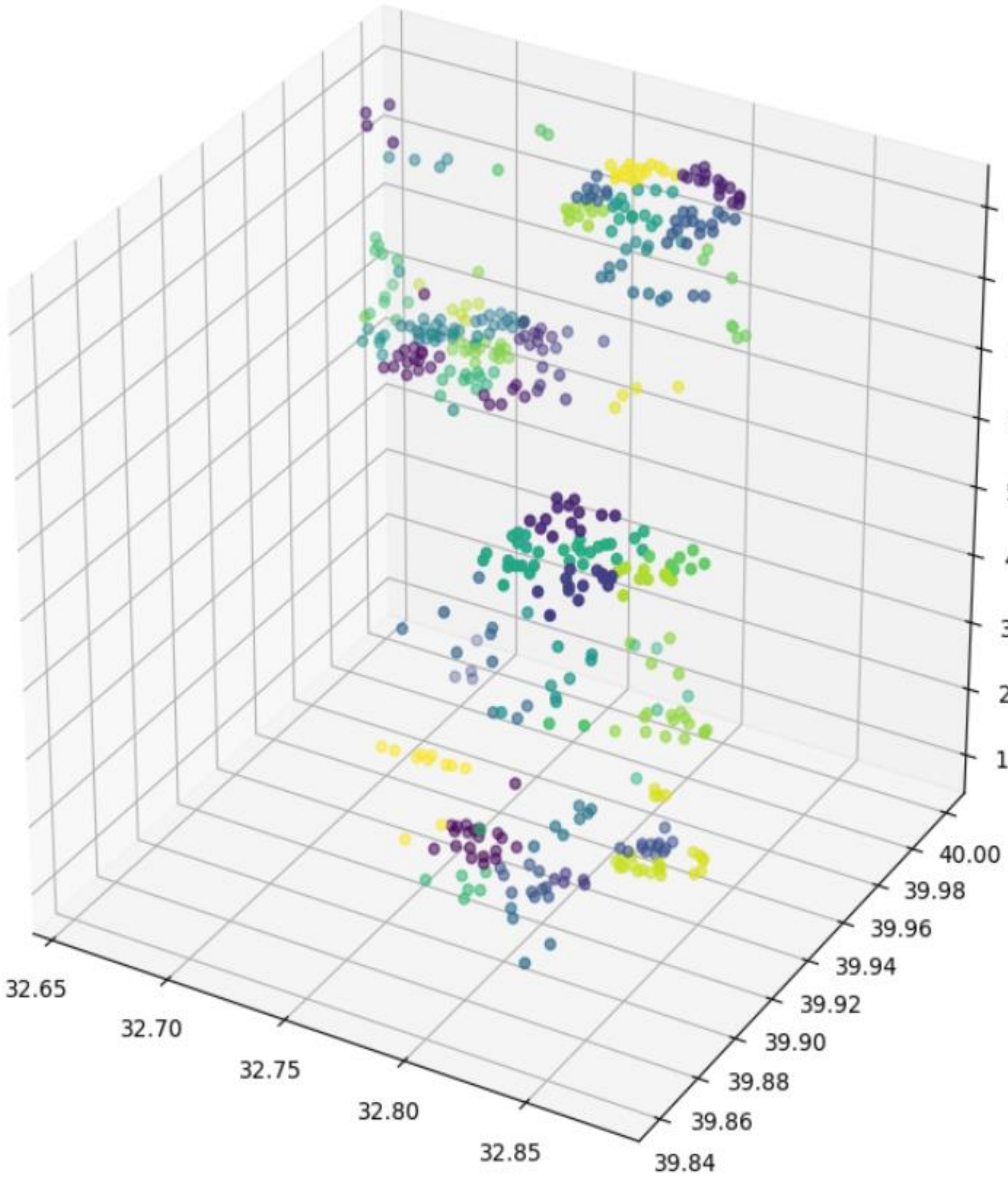
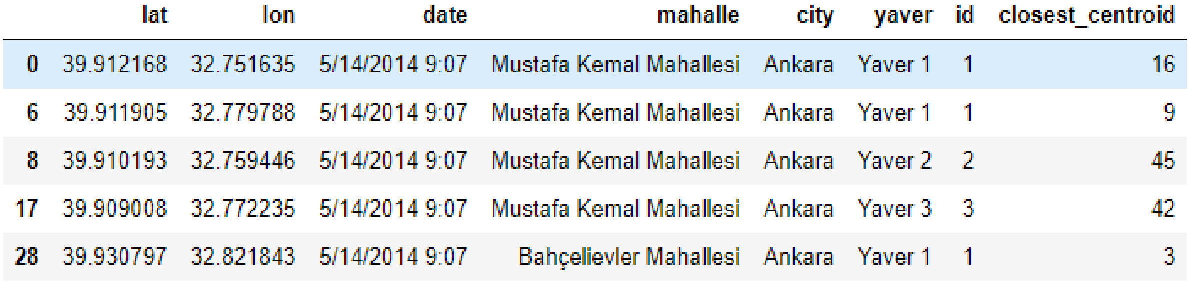


Figure 5. 14: Coloring of sample data in accordance with classification

```
rs = pd.DataFrame(df)
rs['closest_centroid'] = closest_centroids
rs.drop_duplicates(subset=['closest_centroid'], keep='first', inplace=True)
rs.head()

```

With the above code line, a new column called closest_centroid has been added to our data set. With the addition of this column, the same number of closest_centroid is generated as the number of data set with the running of k-means algorithm. However, there are as many unique data as the total number of clusters within the generated data. The sample list of the most recent generated data is shown in Table 5.3:

Table 5. 3: The list on which the closest_centroid is added



	lat	lon	date	mahalle	city	yaver	id	closest_centroid
0	39.912168	32.751635	5/14/2014 9:07	Mustafa Kemal Mahallesi	Ankara	Yaver 1	1	16
6	39.911905	32.779788	5/14/2014 9:07	Mustafa Kemal Mahallesi	Ankara	Yaver 1	1	9
8	39.910193	32.759446	5/14/2014 9:07	Mustafa Kemal Mahallesi	Ankara	Yaver 2	2	45
17	39.909008	32.772235	5/14/2014 9:07	Mustafa Kemal Mahallesi	Ankara	Yaver 3	3	42
28	39.930797	32.821843	5/14/2014 9:07	Bahçelievler Mahallesi	Ankara	Yaver 1	1	3

```
plt.figure(figsize=(10, 6), dpi=100)
plt.scatter(rs['lon'], rs['lat'], c='m', s=100)
plt.show()

```

The final version of the code with the running of the algorithm on the classification of the geographical data is shown below Figure 5.15:

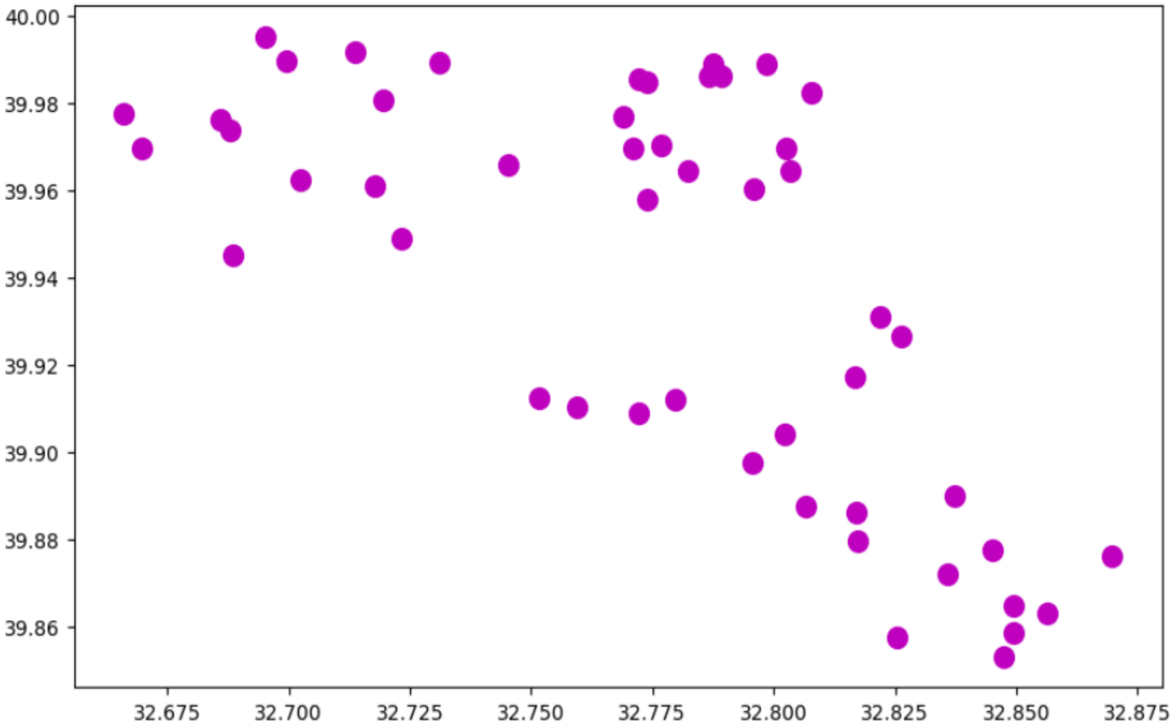


Figure 5. 15: Classification of geographical data

```
plt.figure(figsize=(10, 6), dpi=100)
plt.scatter(cluster_centroids2[:,0], cluster_centroids2[:,1], c='r', alpha=.7, s=150)
plt.scatter(w[:,0], w[:,1], c='k', alpha=.3, s=10)
plt.show()
plt.figure(figsize=(10, 6), dpi=100)
rs_scatter = plt.scatter(rs['lon'], rs['lat'], c='m', alpha=.7, s=150)
df_scatter = plt.scatter(df['lon'], df['lat'], c='b', alpha=.3, s=5)

plt.title('Data set vs k-means reduced set')
plt.legend((df_scatter, rs_scatter), ('Data Set', 'Reduced Set'), loc='upper left')
plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.show()

```

The above script shows in Figure 5.16 the cluster centers of our sample data set as a result of our study formed via k-means algorithm and their location. Additionally, another graphic created by the script shows the centroid of the data we have and the classified data through the running of the k-mean algorithm in Figure 5.17.

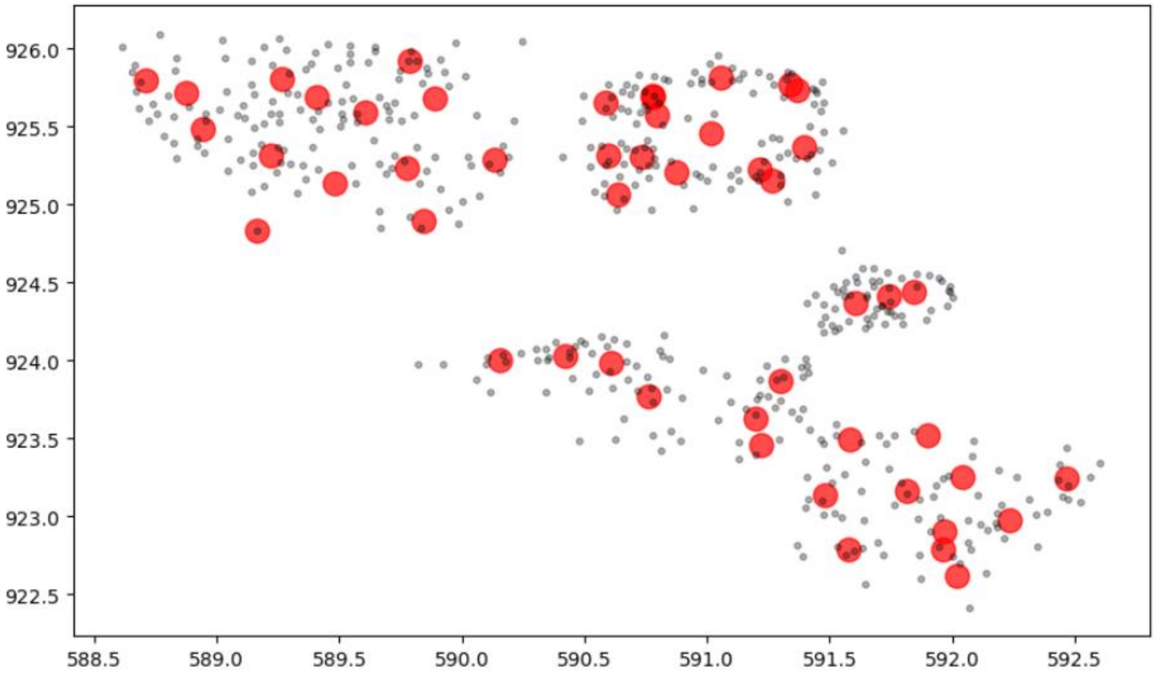


Figure 5. 16: Centroid and data sets

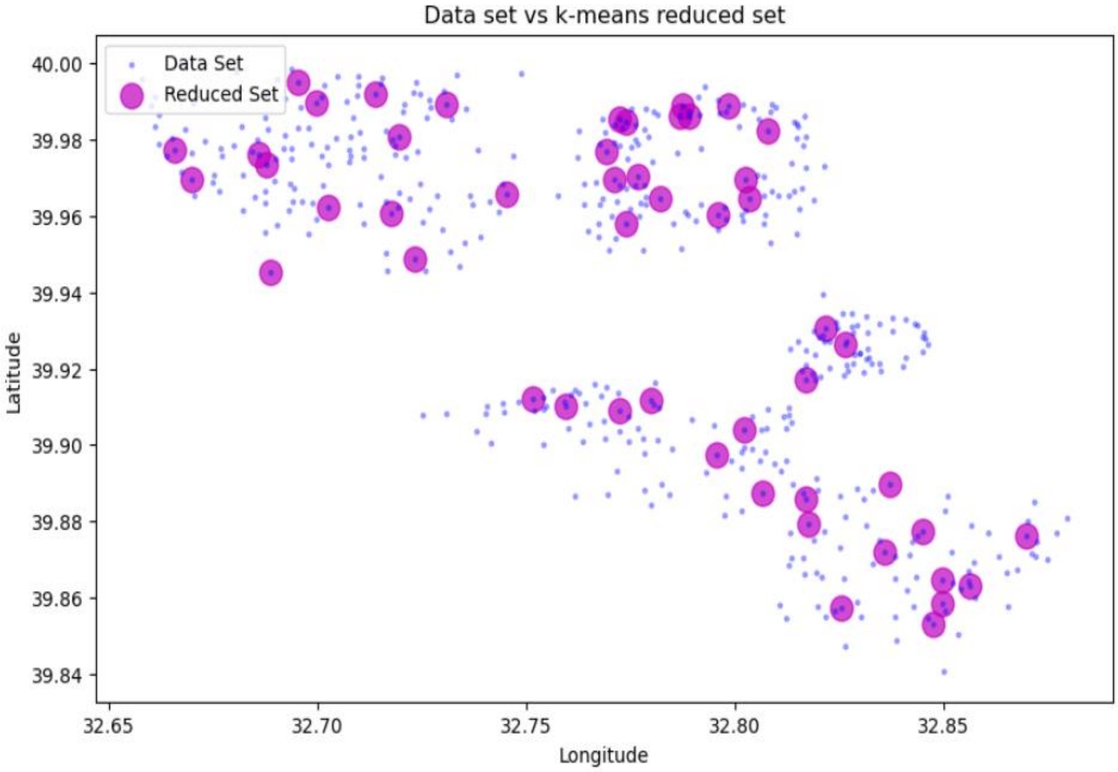


Figure 5. 17: Data set and classified cluster centers

In this thesis study, a business intelligence platform has been developed by creating an end to end application, using real-time geographical information and running a k-mean algorithm within machine learning. First of all, a mobile application is developed and in so doing, the data has been collected real-timely manner. The collected data has been run in sync with k-means algorithm, a classification within python, and the above stated results have been obtained (Lutz, et al., 2001). The main goal here is to designate a new real-time mission to the right Yaver user and to manage the business intelligence platform in a proper manner.

CHAPTER 6

CONCLUSION AND RECOMMENDATION

A study on providing new solutions by implementing real-time geographical data and new technologies such as machine learning on business intelligence is presented as a thesis. The application that is developed is based on the completion of a mission for the most part. One of the biggest problems during the completion process of the mission is that the user that will fulfill the logistics mission must complete the task completely on time. Nowadays, many researchers have been trying to develop applications using machine learning and artificial intelligence for logistics and distribution channels (Staub, et al., 2015). The accessibility of data collection devices has increased with the development of technology. With the increase in devices, lots of data has emerged. Now that mobile devices are easy to access by people, business intelligence platforms beneficial for companies or institutions also became widespread.

With the spread of business intelligence, different data from users have begun to be collected, and many tools and algorithms have been developed by researchers to record and analyze this data. Data mining or big data disciplines have recently emerged in the processing of this data. To give meaning to the data and to use it effectively; algorithms on processes such as clustering, cleansing and future analysis have been worked on under these disciplines.

This thesis study develops a business intelligence platform in the form of a mobile application through k-means algorithm within machine learning with the help of Python modules. The aim of the study is to classify user data in logistics distribution processes, in order to save resources and to complete the process without losing time. In the thesis, latitude and longitude information of mobile application users are used based on the clustering algorithm on k-means algorithm. Geographical data, latitude and longitude, is of great importance for applications in business intelligence platforms with processes such as logistics. It is seen that disciplines such as data mining and artificial intelligence are used to make the distribution processes quicker and to lower costs within the platforms. In addition to this data, the completion of the missions by suitable Yaver users is ensured through data retrieved with the help of the clustering process made through k-means algorithm using various parameters. The business intelligence platform developed for this thesis based on real-time geographical data enabled the efficient use of the application through the clustering process and assigning the missions to a competent Yaver user. With the integration of the business intelligence platform on many applications performing logistics missions, desired classifications can be carried out with relevant parameters and business processes can be reduced to a minimum.

Alongside this thesis, sham transactions can be revealed by using artificial intelligence algorithms to increase reliability between users and machine learning algorithms with further research. Besides, online payments through applications are used extensively today, however, payment processes should be carried out safely through the system. For this purpose, studies are carried out to use all of the individuals' data properly to keep the security in payment systems at a high level in platforms such as business intelligence. For this thesis, payment systems can be defined as further research. In the future, real-time management of the application can be ensured through the addition of auto-control.

REFERENCES

Aarts, E., & Lenstra, J. K. (2003). *Local Search in Combinatorial Optimization.* Atlanta, USA: Princeton University Press.

Alanis, A., Arana-Daniel, N., & López-Franco, C. (2018). *Bio-inspired Algorithms for Engineering.* Butterworth-Heinemann.

Arrillaga, J., & Watson, N. (2003). *Power System Harmonics, 2nd Edition.* John Wiley & Sons, Ltd.

Asheibi, A., Stirling, D., & Sutanto, D. (2009). *Analyzing Harmonic Monitoring Data Using Supervised and Unsupervised Learning.* IEEE Transactions On Power Delivery.

Banks, A., & Porcello, E. (2017). *Learning React: Functional Web Development with React and Redux 1st Edition.* California, USA: O'Reilly Media.

Bressert, E. (2012). *SciPy and NumPy: An Overview for Developers.* O'Reilly Media.

Chen, D. (2017). *Pandas for Everyone : Python Data Analysis.* Boston, USA: Addison-Wesley Professional.

Chen, H., Chiang, R., & Storey , C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly,* s. 1165-1188.

Cielen, D., Meysman, A., & Ali, M. (2016). *Introducing Data Science: Big Data, Machine Learning, and more, using Python tools 1st Edition.* Manning Publications.

Douglas, K. (2005). *PostgreSQL.* Sams Publishing.

Drake, J., & Worsley, J. (2002). *Practical PostgreSQL O'Reilly Media.* O'Reilly Media.

Erişti, H., & Demir, Y. (2010). A new algorithm for automatic classification of power quality events based on wavelet transform and SVM. *Expert Systems with Applications, 37*(6), 4094-4102.

Gersho, A., & Gray, R. (1992). *Vector Quantization and Signal Compression.* New York: Springer US.

Goodchild, M. (1992). A hierarchical spatial data structure for global geographic information systems. *CVGIP: Graphical Models and Image Processing*, 31-45.

Grady, W., & Santoso, S. (2001). Understanding Power System Harmonics. *IEEE Power Engineering Review*.

Güngör, Z., & Ünler, A. (2008). K-Harmonic means data clustering with tabu-search method. *Applied Mathematical Modelling*, 1115-1125.

Idris, I. (2018). *NumPy: Beginner's Guide - Third Edition 3rd Edition.* Packt Publishing.

Işık, M. (2006). Data Mining Applications with Division Clustering Methods. *Institute of Science*, 15-41.

Jain , A., & Dubes, R. (1988). *Algorithms for Clustering Data (Prentice Hall Advanced Reference Series : Computer Science).* New Jersey: Pearson College Div; First Edition .

Jain, A., Murty, M., & Flynn, P. (1999). Data Clustering: A Review, ACM Computing Surveys. *The dblp computer science bibliography* , 3-31.

Kaufman, L., & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley and Sons.

Kresse, W., & Danko, D. (2012). *Springer Handbook of Geographic Information.* Berlin: Springer-Verlag.

Kubat, M., Holte, R., & Matwin , S. (1998). Machine Learning for the Detection of Oil Spills in Satellite Radar Images. *Machine Learning*, 195–215.

Kuncheva, L. (tarih yok). *Combining Pattern Classifiers: Methods and Algorithms, 2nd Edition.* 2004: John Wiley & Sons.

Li, L.-L., Chang, Y.-B., Tseng, M.-L., & Lim, M. (2020). Wind power prediction using a novel model on wavelet decomposition-support vector machines-improved atomic search algorithm. *Journal of Cleaner Production*.

Long, J. (tarih yok). *Spring Framework LiveLessons.* Addison-Wesley Professional.

Lu, D. (2007). A survey of image classification methods and techniques for improving classification performance. Int. J. Remote Sens. 28, 823–870. *International Journal of Remote Sensing*, 823–870.

Luhn , H. (1958). *A Business Intelligence System.* New York: IBM Journal of Research and Development.

Lutz, M. (2001). *Programming Python, Second Edition.* O'Reilly Media.

Mardan, A. (2017). *React Quickly: Painless web apps with React, JSX, Redux, and GraphQL 1st Edition.* Manning Publications.

McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython 2nd Edition.* O'Reilly Media.

Müller, A., & Guido, S. (n.d.). *Introduction to Machine Learning with Python: A Guide for Data Scientists.* New York, USA: O'Reilly Media.

Oliphant, T. (2015). *Guide to NumPy: 2nd Edition.* CreateSpace Independent Publishing Platform.

Osei-Bryson, K.-M., & Inniss, T. (2007). *A hybrid clustering algorithm.* Department of Mathematics. Atlanta: Computers & Operations Research.

Özekes, S. (2003). Data mining models and application areas. *Istanbul Commerce University Journal of Science*, 65-82.

Patel, E., & Kushwaha, D. (2020). Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. *Procedia Computer Science*(171), 158-167.

Piatetsky . (2019). *Python Leads the 11 Top Data Science, Machine Learning Platforms: Trends and Analysis.* Kdnuggets: https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html adresinden alındı

Raschka, S., Patterson, J., & Nolet, C. (2020). Review Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. *Information 2020*, 4-15. doi:https://doi.org/10.3390/info11040193

Schmidt, R., Möhring, M., Härting, R.-C., & Reichstein, C. (2015). Industry 4.0 -Potentials for Creating Smart Products: Empirical Research Results. *18th International Conference on Business Information Systems.* Poznań, Poland: LNBIP.

Selim, S., & Ismail, M. (1984). K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality. *Institute of Electrical and Electronics Engineers*, 81-87.

Sennaroglu, B., & Celebi, G. (2018). A military airport location selection by AHP integrated PROMETHEE and VIKOR methods. *Transportation Research Part D: Transport and Environment*, 160–173.

Sobie, C., Freitas, C., & Nicolai, M. (2018). Simulation-driven machine learning: Bearing fault classification. *Mechanical Systems and Signal Processing* , 403–419.

Staltz, A. (2018, October). The Web Began Dying In 2014, Here's How. *Arka Kapi Magazine*, 1-10.

Staub, S., Karaman, E., Karapınar, H., & Kaya, S. (2015). Artificial Neural Network and Agility. *Social and Behavioral Sciences*, 1477–1485.

Stichhauerova, E., & Pelloneova, N. (2019). An Efficiency Assessment of Selected German Airports Using the DEA Model. *Journal of Competitiveness*, 135–151.

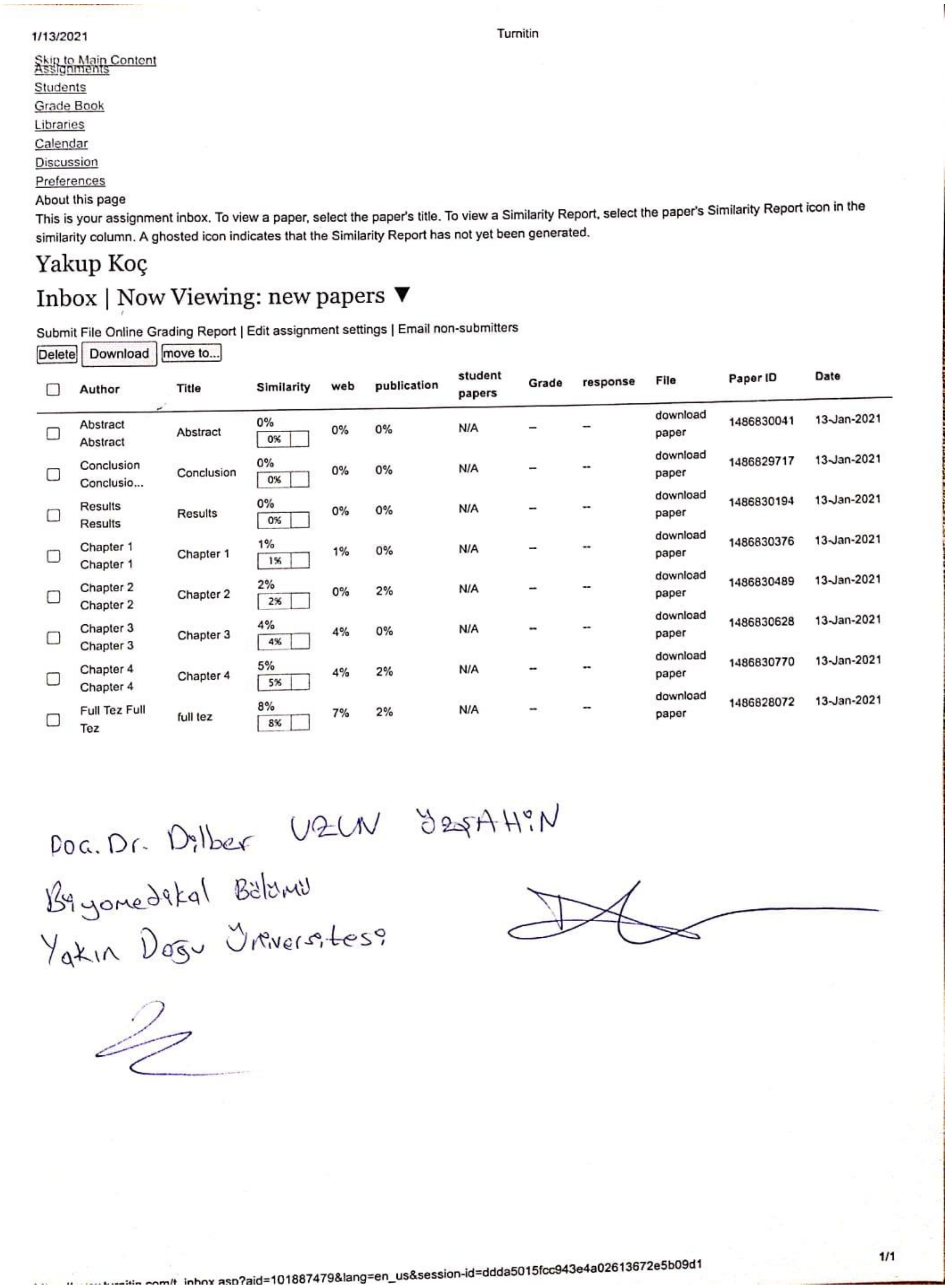
Şişeci, M., Metlek, S., & Cetisli, B. (2014). Accelerating the image segmentation using sub-block technique and clustering methods. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 655-664.

VanderPlas, J. (2016). *Python Data Science Handbook : Essential Tools for Working with Data.* O'Reilly Media.

Walls, C. (2016). *Spring Boot in Action 1st Edition.* Colorado, USA: Manning Publications.

Wang, F., Wang, Q., Nie, F., Li, Z., & Yu, W. (2020). A linear multivariate binary decision tree classifier based on K-means splitting. *Pattern Recognition*(107), 43-76.

Xu, J., Tan, W., & Li, T. (2020). Predicting fan blade icing by using particle swarm optimization and support vector machine algorithm. *Computers & Electrical Engineering*.



1/13/2021 Turnitin

Skip to Main Content
Assignments
Students
Grade Book
Libraries
Calendar
Discussion
Preferences

About this page

This is your assignment inbox. To view a paper, select the paper's title. To view a Similarity Report, select the paper's Similarity Report icon in the similarity column. A ghosted icon indicates that the Similarity Report has not yet been generated.

Yakup Koç

Inbox | Now Viewing: new papers ▼

Submit File Online Grading Report | Edit assignment settings | Email non-submitters

□	Author	Title	Similarity	web	publication	student papers	Grade	response	File	Paper ID	Date
□	Abstract Abstract	Abstract	0%	0%	0%	N/A	--	--	download paper	1486830041	13-Jan-2021
□	Conclusion Conclusio...	Conclusion	0%	0%	0%	N/A	--	--	download paper	1486829717	13-Jan-2021
□	Results Results	Results	0%	0%	0%	N/A	--	--	download paper	1486830194	13-Jan-2021
□	Chapter 1 Chapter 1	Chapter 1	1%	1%	0%	N/A	--	--	download paper	1486830376	13-Jan-2021
□	Chapter 2 Chapter 2	Chapter 2	2%	0%	2%	N/A	--	--	download paper	1486830489	13-Jan-2021
□	Chapter 3 Chapter 3	Chapter 3	4%	4%	0%	N/A	--	--	download paper	1486830628	13-Jan-2021
□	Chapter 4 Chapter 4	Chapter 4	5%	4%	2%	N/A	--	--	download paper	1486830770	13-Jan-2021
□	Full Tez Full Tez	full tez	8%	7%	2%	N/A	--	--	download paper	1486828072	13-Jan-2021

Doç. Dr. Dilber UZUN ÖZŞAHİN
Biyomedikal Bölümü
Yakın Doğu Üniversitesi

...turnitin.com/t_inbox.asp?aid=101887479&lang=en_us&session-id=ddda5015fcc943e4a02613672e5b09d1 1/1