Instituto Tecnológico
y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Electrónica, Sistemas e Informática

Maestría en Sistemas Computacionales

Land Use Identification of the Metropolitan Area of Guadalajara using Bicycle Data: An Unsupervised Classification Approach

---

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
MAESTRO EN SISTEMAS COMPUTACIONALES

Presenta: **DULCE MARÍA GRACIA RIVERA**

Director **DR. IVÁN ESTEBAN VILLALÓN TURRUBIATES**

Tlaquepaque, Jalisco. Noviembre de 2021.

ACKNOWLEDGEMENTS

I would like to acknowledge:

My thesis assessor Dr. Iván Villalón, for his disposition, understanding, cheerful attitude, suggested the topic and always support me.

My thesis reviews and professors Dr. Francisco Cervantes and Ing. Rodolfo Luthe for their knowledge shared during classes.

To my mom and Dad who have always been there to listen and support me.

To Eduardo who always was there with a cheerful attitude and with his support.

To Oracle Corporation, the company that I work for, provides resources and flexibility that allowed me to continue my professional development.

To Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO) for the resources provided for the development of the investigation.

To Consejo Nacional de Ciencia y Tecnología (CONACYT) for the financial support received with the scholarship by the number 501617.

AGRADECIMIENTOS

El autor desea dar las gracias a:

Mi asesor de tesis Dr. Iván Villalón, por su disposición, comprensión, actitud alegre, sugerir el tema y apoyo siempre.

Mis revisores de tesis y profesores Dr. Francisco Cervantes e Ing. Rodolfo Luthe por su conocimiento compartido en clase.

A mi Mamá y Papá quienes siempre estuvieron ahí para escucharme y apoyarme.

A Eduardo quien siempre estuvo ahí con su actitud alegre y apoyo.

A la empresa Oracle en la cual trabajo que brindo los recursos y flexibilidad que me permitió continuar con mi desarrollo profesional

Al Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO) por sus recursos provistos para el desarrollo de esta investigación.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el soporte financiero recibido con el número de beca 501617.

DEDICATION

To my family who have always given me their support, love, understanding and taught me that we can always achieve whatever we proposed.

To my friends and fellows from the master that have shared knowledge, support and laughs.

DEDICATORIA

A mi familia que siempre que siempre me ha brindado su apoyo, amor, comprensión y enseñarme que siempre podemos lograr los que nos proponemos.

A mis amigos y compañeros de la maestría que compartieron su conocimiento, apoyo y risas.

ABSTRACT

The following work proposes different ways to solve a problem that is currently encountered, which is to do research in the area of land-use, land-mapping and human behavior evaluating its movement through information sources that contain geo-referenced information.

MiBici was used as a source of information, which is a bicycle sharing platform that exists in the city of Guadalajara, Jalisco, which shares month after month a consolidated file of the trips that are made each month, it is worth mentioning that the access of this information is totally free. The methodologies used were agile for project planning, KNN, Decision Trees and KMeans for the cauterization of the zones, the programming language used was Python, in addition an implementation proposal was attached using the Amazon Web Service platform with the aim of proposing a "simpler" solution to implement, but with the same value as doing it with pure open sources.

The process was divided primarily into 3 parts where the first was to clean data and understand it, machine learning algorithms were applied, which were Decision tree and KNN. The second stage evaluating the results of the previous stage, where added new fields to improve the results and KMeans was applied to create groups and as a last step a flow was created that began with cleaning the raw data using AWS tools and ended with the interpretation of the final results.

The results obtained were too encouraging since the groups that were obtained were too marked and reviewing it with the areas related to the nodes a great relationship was found. Without a doubt, there is still too much work to be done in this area of research.

RESUMEN

El siguiente trabajo propone diferentes maneras de resolver una problemática que se encuentra en la actualidad, que es el hacer la investigación en el área de land-use, mapeo y comportamiento humano evaluando su movimiento por medio de fuentes de información que contienen información geo-referenciada, también se comparte la meta de clasificar diferentes secciones y su relación entre ellas.

Se utilizó como fuente de información MiBici que es una plataforma de compartimiento de bicicleta que existe en la ciudad de Guadalajara, Jalisco, la cual comparte mes tras mes un archivo consolidado de los viajes que se realizan en cada mes, cabe mencionar que el acceso de esta información es totalmente libre. Las metodologías utilizadas fueron agile para planeación del proyecto, KNN, Decision Trees y KMeans para la cauterización de las zonas, el lenguaje de programación utilizado fue Python, además se anexo una propuesta de implementación utilizando la plataforma de Amazon Web Service con el objetivo de proponer una solución más “sencilla” de implementar, pero con el mismo valor que hacerlo con puros recursos libres.

El proceso se dividió primordialmente en 3 partes en donde la primera fue limpiar datos y entenderlos, se aplicaron algoritmos machine learning que fueron Decision tree y KNN, para la segunda etapa evaluando los resultados de la etapa anterior se hicieron modificaciones a los datos en donde se agregaron nuevos campos para mejor los resultados y se aplicó KMeans para la creación de grupos y como último paso se creó un flujo que inicio con la limpieza de los datos en crudo utilizando herramientas de AWS y se terminó con la interpretación de los resultados finales.

Los resultados obtenidos fueron demasiados alentadores ya que los grupos que se obtuvieron fueron demasiados marcados y revisándolo con las zonas relacionadas a los nodos se encontró una gran relación. Sin duda alguna queda aún demasiado trabajo a desarrollar en esta rama de investigación.

TABLA DE CONTENIDO

1. INTRODUCTION 14

1.1. Background 15
1.2. Justification 15
1.3. Problem 16
1.4. Objective 16
1.4.1. General Objective 16
1.4.2. Specific Objective 16
1.5. Scientific or technologic share / Innovation 16

2. STATE OF THE ART OR THE TECHNIQUE 17

2.1. Social Media related work 18
2.1.1. Not Image related research 18
2.1.1.1. Reveling the relationship between spatio-temporal distribution of population and urban function with social media data 18
2.1.1.2. Fine-Grained subjective partitioning of urban space using human interactions from social media 19
2.1.2. Image related research 20
2.1.2.1. Putting people in the picture: combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in Shanghai 20
2.1.2.2. Improved classification of satellite imagery using spatial feature maps extracted from social media 21
2.1.2.3. Towards an automatic urban settlement mapping from multi-temporal InSAR trained by social media 22
2.1.2.4. Potential analysis of feature extraction based quick response for environmental change with social media photos 23
2.1.2.5. Using social media data to map urban areas: ideas and limits 23
2.2. Bicycle training data 24
2.2.1. Identification of land-use characteristics using bicycle sharing data: A deep learning approach 24

3. THEORIC/CONCEPTUAL FRAMEWORK 26

3.1. Introductory Concepts Related to the investigation 27
3.2. Statistics Terminology 28
3.3. Methodologies 29
3.4. Libraries 33
3.5. Cloud Computing 36
3.6. Social Media 38
3.7. Information Technologies 38

4. DEVELOPMENT METHODOLOGY 40

4.1. How the idea was born – Background 41
4.2. Requirements 41
4.3. Analyze Source Data, Dataset Prepossessing and Cleaning 43

4.4. STAGE 1 – (KNN AND DECISION TREES) 45
4.4.1. Analyze Data *45*
4.4.2. Clean Data *47*
4.4.3. Implementation of Algorithms *48*
4.5. STAGE 2 (KMEANS) 48
4.5.1. Prepare Data Set *48*
4.5.2. Implement Algorithm *50*
4.6. STAGE 3 50
4.6.1. Prepare Data *50*
4.6.2. APPLIED CLASSIFICATION ALGORITHM 52
4.6.3. Analyze Results *53*
5. RESULTS AND DISCUSION 55
5.1. RESULT STAGE 1 56
5.1.1. Decision Tree *57*
5.1.2. K-NN *59*
5.2. RESULT STAGE 2 59
5.2.1. Only Origin *60*
5.2.2. Only Destiny *61*
5.2.3. Origin and Destiny *65*
5.2.4. Origin and Destiny Difference *67*
6. CONCLUSION 69
6.1. CONCLUSIONS 70
6.2. FUTURE WORK 70

LIST OF FIGURES

Figure 3.1 Image showing how similar data points typically exist close to each other 30
Figure 3.2 Example of Elbow Method 31
Figure 3.3 Example of Decision Tree 32
Figure 4.1 Year of Birth Histogram 45
Figure 4.2 Details from the station 51st 46
Figure 4.4 Records with null data in the DataSet 49
Figure 4.5 Final DataSet 49
Figure 4.6 Diagram of the implementation at AWS 50
Figure 4.7 Data part of the Bucked Raw Data 51
Figure 4.8 DataSet to be clean using AWS Glue Data Brew 51
Figure 4.9 Job created for export the clean data 52
Figure 4.9 Recipe Jobs 52
Figure 4.11 Notebook to be run at AWS Sage Maker 53
Figure 4.12 Results from the classification exported to Quick Sight 53
Figure 4.13 Results interpretation 54
Figure 5.1 Characteristics Correlation 1st DataSet 56
Figure 5.2 Characteristics Correlation 2nd DataSet 57
Figure 5.3 Decision Tree Results 57
Figure 5.4 Decision Tree max deep 2 58
Figure 5.5 K value and precision rate accordingly 59
Figure 5.6 Elbow Method for Origin 60
Figure 5.7 K-Means Origen Cluster 61
Figure 5.8 Elbow Method for Destiny 62
Figure 5.9 K-Means Destino Cluster 63
Figure 5.10 K-Means Destino 2 clusters 64
Figure 5.11 Elbow Method for Origen and Destino 65
Figure 5.12 K-Means Origen and Destino 66
Figure 5.13 Elbow Method for Origen and Destino Diferentes 67
Figure 5.14 K-Means Origen and Destino Different 68
A.1 Plot Cluster Comparation 73
A.2 Gradient Boosted Decision Trees 74

LIST OF TABLES

Table 4.1 Features to Develop in the project 41
Table 4.2 Stations that are part of MiBici 43
Table 4.3 Interesting Data of the total Trips 45
Table 4.4 Distribution of the trips per Month 47
Table 4.5 Characteristics of the algorithms used 48
Table 5.1 K-NN Accuracy 59

LIST OF ACRONYMS AND ABBREVIATIONS

AWS	Amazon Web Service
API	Application Programming Interface
CIE	Comission Internationale de l´Eclairage
CNN	Convolutional Neural Network
DBSCAN	Density Based Spatial Clustering of Application with Noise
EO	Earth Observation
GIS	Geographic Information System
InSAR	Interferometric SAR
KNN	K Nearest Neighbor
LCZ	Local Climate Zones
LBS	Location Based Services
NaN	Not A Number
OCSVM	One Class Support Vector Machine
POIs	Point Of Interest
SAR	Synthetic Aperture Radar
VAC	Voting Active Cluster

1. INTRODUCTION

Several investigations have been shown that the use of information generated by users either by social media, internet companies or other services that at the end of the day provide urban - land use characteristics data which are used to analyze distinct aspects such as human behavior, human activities, mapping of cities and regions (map urban areas), among others. The previous ones have been showing that work is even better than traditional methods such as Earth Observation by remote sensing.

In the modern world that we live in it is important for us as humans adapt to modern technologies and try to innovate and improve with the resources that we have in our environment. Normally the previous data sources of data that were listed before such as Twitter, Telcent, Weibo that have in common sharing geographical, economics and political factors that help not only to map better areas but also adding geolocation to analyzes allow to understand the spatio-temporal distribution, human behavior, urban dynamics.

The methodologies used before incorporating analysis of SAR images, telephony data by MatLab, ArcGI, statistics and ML have shown that add social media data improves mapping urban zones. Although most of them use very rich data sources with load of data or time cost computer. Based on those implementations, this research was focus on using source data that is mean full and with low computer cost.

The main source of data used was MiBici that is a bicycle rent service located at Guadalajara, Jalisco, using ML models. It was divided by 3 distinct stages where first was implemented using supervise methods and only reviewed the period of January 2019 to December 2019, the remaining stages developed using unsupervised method and the period of time would extends to December 2014 to December 2020. Nevertheless, dataset preparation is required in every stage, and it is important to mention that preparing and understanding the data would be an extremely critical step to follow in every part of the process.

1.1. Background

There is an increasing trend at using social media data to map human activity. Social media, such as Twitter and Facebook, are commonly considered as valuable information streams, well beyond their initial role as communication tools. The proliferation of the use of social media provides every user with rich digital contents, such as geographic coordinates, photographs, videos, resulting into significant sources of" big data." Although massive data from social media are often published without scientific intent, often carry little scientific merit, and the use of social media tools is not uniformly distributed across the world, they provide a source of useful information that may help to avoid, for instance, long and expensive in situ data collection for remote sensing image classification [1].

Urban areas, which are characterized by high population densities and extensive human features, are an essential component of human society. Urbanization cuts both ways: it brings modernization, industrialization as well as ecological and environmental problems, such as the greenhouse effect, urban heat island effect, and air-pollution consequences. Thus, accurate and timely urban settlement information is critical for monitoring urbanization process and providing answers to the balance between the urbanization and environmental protection [3].

The emergence of social media brings new opportunities to understand urban settlement from a distinct perspective. Social media collects massive spatiotemporal data that can help people understand the physical as well as social environment of urban areas. Whilst some research has been carried out on urban zoning identification from social media data, far too little attention has been paid to investigate sampling scheme of social media data for urban settlement extraction [3].

The rise of big geo-data and network analysis science brings new opportunities for the subjective partitioning of urban space. Accompanied by the subjectivity of human beings, various unprecedented types of activity data have been produced. These activity data, such as taxi GPS data, mobile phone data, and social media data, are regarded as inherent driving forces for the subjective partitioning of urban space. Combined with human activity data, network analysis methods have been expanded as a support for the subjective partitioning of urban space [4].

1.2. Justification

Every day as humans, we create an exceptionally large amount of information that can be used to create solutions to complex problems, from being able to predict a meteorological event, to being able to monitor climate change through images uploaded to social media. Even to be able to map a city based on the use that its habitants give it.

Although there are many fields of study in this area, there are few studies realized, most of them present use of resources that can be expensive.

The proposal of this research is to carry it out using open source resources and contribute to the mapping area of a city in order to understand the behavior of humans who in this case use the MiBici service in the city of Guadalajara with the goal to propose better city planning mapping.

1.3. Problem

The problem that I am addressing is to make this type of research more accessible. Since most of the documentation that exists encourages the use of images from different sensors and satellite and / or text and full metadata from different social media applications or electronic companies to complement the data normally used that is georeferenced coming from the different sources listed before. As well as the use of different paid software, which this information is not as accessible to the public in general or it may be expensive to carry out this type of research, so all this can become a little detrimental to the investigation depending on the environment in which you want to carry out the investigation. The goal of this research is to classify different sections and their relation at Guadalajara, Jalisco.

1.4. Objective

1.4.1. General Objective

To investigate and create new knowledge using accessible data and open source resources.

1.4.2. Specific Objective

To accomplish the general objective, the following specific objectives are required

- To analyze source data
- To clean source data and create one complete DataSet
- To apply supervise classification methods to data
- To discuss the results of supervising methods
- To refine data for used with unsupervised classification methods
- To discuss the results from unsupervised methods
- To validate results with AWS

1.5. Scientific or technologic share / Innovation

This research aims to be able to contribute knowledge in the area of Urban Settlements and human behavior. By proposing and implementing open-source resources for future work. Also, set a different and friendly implementation using cloud services.

2. STATE OF THE ART OR THE TECHNIQUE

In the following chapter the abstract of the research that has been reviewed and taken as a base for the development of the investigation could be found.

All the researchers agree that the use of the data that now days as humans create (social media, shared bicycle data) is valuable to plan better, understand urban settlements in a unique perspective and even with environmental purposes and natural disasters.

To go a little deeper into some of the important concepts mentioned in this chapter, they are reviewed in the appendix section.

2.1. Social Media related work

2.1.1. Not Image related research

2.1.1.1. Reveling the relationship between spatio-temporal distribution of population and urban function with social media data

In 2016 Li, Shen, and Hao [3], followed an investigation had as an objective analyst the dynamic of urban space. They used POIs and LBS produced by Tencent't social netwoking application, recollected from July 29 to August 2 in 2015.

The objective of the investigation was to attempt to uncover the relationship between spatio-temporal dissemination and urban capacities utilizing an uncommon high-resolution and broad-coverage-crowd LBS information from Tencent, one of the greatest web companies in China, recording populace with hour as a time scale. They then examine how their work can help get it the dynamic nature of our city and how our work can help urban arranging decision-making [3].

The have create a spatial entropy model based on POIs and extend it to examine the temporal repartition of population by hours on weekdays and weekend [3].

During the investigation they noticed that higher transient entropy indicates that individuals show up more regularly in certain areas amid distinctive time periods; lower temporal entropy means that individuals show up more habitually in a certain range in as it were certain time periods; a larger population and higher worldly entropy demonstrate an area with 24-h movement [3].

The understanding of how people move and take portion in different exercises within a city and how they connected with urban space is crucial to empower a people-centric urban arranging hone that accounts for open requests. For instance, individuals control off their versatile phones at night, and this causes private regions to seem to have relatively higher entropy. This would cruel that a number of individuals are showing up at these places at any time, which isn't genuine in reality. In addition, the actual temporal populace dissemination is closely connected to the estimate of a zone. Bigger frameworks are steadier, and the same as the transient dissemination of populace on a land parcel.

Working zones with less facilities and moo building thickness tend to appear high population and unevenly transiently conveyed population. Large commercial centers appear tall population with unevenly transiently disseminated population, while district-grade commercial centers appear high population with equitably transiently distributed population. Distinctive urban capacities have different temporal populace dissemination bends [3].

2.1.1.2. Fine-Grained subjective partitioning of urban space using human interactions from social media

In 2019 Qiao, Wang, Wu, Luo, Ruan and Gu [4] developed an investigation that present new method for fine-grained subjective partitioning of urban space based on the combination of network analysis and human interactions from social media by taking into account the importance of the attraction of nodes in shaping an urban space. They have found that social media can refine comes about subjective partitioning.

The displayed approach incorporates three primary parts to achieve fine-grained subjective apportioning of urban spaces by using human intelligent gotten from social media data. First, the substantial and invalid movement flows are decided, and a cut-off point identified as the basis for progressive apportioning. Moment progressed hierarchical spatial systems are developed by consolidating a gravity model. Third, the progressive spatial community recognized to get progressive and fine-grained partitioning results [4].

They have built an information crawler outline to gather human activities from the Sina Weibo platform (a famous social media platform in China) inside the ponder zone. Information with GPS data implanted was collected through the application programming interfacing (APIs) given by Sina Weibo. With an extended period of collection (from 2014.02.01 to 2014.09.31), commotion expulsion (expelling notices, marketing accounts and virtual individual accounts) and preprocessing (removing clients with less than 3 messages posted amid our collection period), they finally obtained 4,111,364 substantial streams and 2,854,369 invalid streams for the fine-grained subjective dividing of urban space [4].

As result can be found the following three characteristics:

1) The regulatory boundaries and expressways decided in a top-down design have preconceived influences on genuine human movement spaces.
2) The real human action spaces are not only constrained by authoritative boundaries and expressways but too affected by inclination, religion, and financial level and so on.
3) The top-level apportioning comes about are more refined than those of the official district-level boundaries, especially in urban areas

Also, that the subjective dividing of urban spaces by utilizing human interactions can viably uncover the common ways that people interact and the genuine human action spaces, which has great suggestions for the official administration and configuration of cities [4].

2.1.2. Image related research

2.1.2.1. Putting people in the picture: combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in Shanghai

At 2016 Jendryke, Balz, McClure and Liao [2], lead an investigation about the contextual urban information in Shanghai, motivated by the need for an enhanced and timely analysis of urban processes in China cities.

For the investigation, they have used SAR data and Location-Base Social Media Messages. The use of both data enhances land classification and interpretation of urban development and urban activity pattern.

They have shown that it is possible to detect ‘high’ and ‘low’ built-up areas as well as human activity patterns from each data set [2].

The process that they have follow it was split up in two process one to analyst the SAR images stack and the other to analyst the data collection, the result of the processes are combined into one layer to finally get the representation of four classes that are:

1. Build-up with human activity
2. Non-build-up and have no human activity
3. Build-up without human activity
4. Non-build-up with human activity

They have seen that the extracted build-up information from remote sensing also the human activity data from social media are not randomly spreader, but sort of clustered. Entirely dispersed build-up areas and human activity data would result in checkerboard-like appearance [2].

As result of the investigation, they proposed an approach that provides an improvement interpretation for urban dynamics and found that the existence of human activity expressed by social media messages correlate with urban build-up areas. As well they have addressed some limitations that are important to keep an eye on whenever working with social media data, for example, that they don’t know who is using the network and what percentage of the actually human activity in the area of interest. So, investigations normally restricted to people that owns a device [2].

Most of the time is need it to create assumptions as they did that younger people and middle-aged people most likely are active on the platform of choice (in this case Weibo social media) [2].

2.1.2.2. Improved classification of satellite imagery using spatial feature maps extracted from social media

In 2018 Leichter, Wittich, Rottensteiner, Werner and Sester [1] follow and investigation with the goal of determine a classification for LCZ by exploring social media data in the context of Remote Sensing and Spatial Information sciences.

All the investigation used local area around Washington DC, because it was seemed to be the most suitable data for the objective due to the high amount of twitter users as well as the availability of open satellite imaginary. The dataset for the investigation generated by filtering 680,982,894 tweets collected between 09 February 2018 and 19 June 2018, filtering by the target location. The result were 3925559 tweets, most of the tweets were present in the urban agglomerations. In addiction Sentinel-2 earth, observation images were also use [1].

They have generated six feature maps, which were used in two experiments and were selected due to their potential relevance for LCZ classes. The features were:

- Tweet count
- Mean text length
- Mean friends count
- Mean time
- Mean tag count
- Mean user mentions count

The investigation was divided into two experiments, for the first one only the six features listed before to predict LCZ classes using boost trees. Important to mention that they have prepared the data feature maps for their training by means of linear normalization to the interval from 0 to 1 [1].

For the second experiment, they use a convolutional neural network for a pixel wise prediction of LCZ classes based on different bans of satellite imaginary and the six twitter feature maps. Therefore, we can summarize that the first experiment referred to a baseline model, infers the LCZ classed based on the satellite images only and the second experiment additionally uses the twitter data and was refer to as augmented model [1].

They conclude for the first experiment that the twitter information contains data suitable for LCZ classification and that the quality is low, because of the missing explicit content information and noisy nature of the twitter data. For the second experiment, they gathered additional proves for the positive effect of twitter data and more realistic classification scenario [1].

In addition, they were able to prove their initial hypothesis about the beneficial contribution of twitter information on land use classification and our classification approach in suitable for data fusion.

2.1.2.3. Towards an automatic urban settlement mapping from multi-temporal InSAR trained by social media

At 2018 Miao, Wu, Shi, Gamba and Jang [5] publish and investigation with propose to address 3 issues that were:

- The effect of de-speckling that is usually underestimated to the extent that the improvement of the SAR image quality is ignored in some investigations.
- Is missing a method that combines full InSAR information.
- Normally training samples are time-consuming and labor-intensive or even impractical when classifying satellite data at the regional/global scale.

The paper they present an automatic method for urban settlements mapping trained by multi-temporal InSAR using social media [5].

As data sources they have used Twitter as social media, the tweets were collected between February 17, 2016, to December 17, 2016. In addition, they have used ERS-1 SAR data from Rotterdam, Netherlands. The study area located in the city center were four main land covers founded:

- Urban
- Forest
- Grass
- Water

The steps they follow in order to accomplish their goal were first to remove noises if SAR imagery and a filtering method to remove redundancy and error of social media data. Important to mention that only geo-tagged data from social media were explored to generate the training samples, which will be subsequently applied to train one class classifier to classify multi-temporal InSAR imagery to produce urban settlements maps [5].

Once the data was, clean they generated training samples and since only training samples of urban class were, generate from social media data and training samples of other land-cover classes are vacant, the researchers used OCSVM to extract urban settlements. The main idea of using OCSVM is because they would be able to infer a decision boundary from training samples of the simple class. So, if the test sample locates within the decision boundary it will be classify as urban class, otherwise, will be classify as non-urban class [5].

They conclude by presenting a new method to delineate urban settlements based on the integration of social media data and multi-temporal InSAR imagery. Social media data provide a new strategy to generate training samples in near real time and free cost that significantly improves the efficiency of practical urban settlements extraction, with that, could be confirm the hypothesis achieves a comparable performance with common methods [5].

2.1.2.4. Potential analysis of feature extraction based quick response for environmental change with social media photos

In 2018 Wu, Gao, Liao, Gamba and Zhang [8] proposed a framework base on color feature extraction of social media photos and correlation analysis with air quality parameters to monitor environmental health.

The investigation used Panoramio website, which photos include attributes as location, description, geo-position (latitude and longitude), imaging, time, and others, were collect 846 photos from April 2008 to October 2013, important to mention that were depicting portions of Beijing Olympic Park area [8].

They first processed the remote sensing images performing geometric correlation, radiometric correlation, and atmospheric correlation. However, social media photos contain many forms of information, so it is necessary to filter the available data sets before analyzing them [8].

Color models can be used for feature extraction from social media photos. They have use RGB where value could be map into the CIE, once the color data is obtained from social media photos they are stored in an organized database with the number of attributes, like the geographical position, date and so on [8].

Remote sensing images can address data to support wide area decision-making several challenges remain in emergency management and quick response, especially in case of air pollution monitoring [8].

They have performed the goodness-of-fit check to diagnose the adequacy of the fitted model. Social media photos do not have enough good fit to the sun-photometer observation ground, definitely new color model needs to be defined to represent sky colors feature. Even thought, the results obtained are a proof of concept that social media photos have an interesting potential for air parameter estimate and remote sensing parameter validation with low cost [8].

So as conclusion, social media photos have interesting validation with low cost but inconsistent data inexact temporal occurrence of the social media photos; It’s important to keep consistent data collection, social media photos have an interesting potential for air parameter estimate and remote sensing parameter validation with minimal impact.

2.1.2.5. Using social media data to map urban areas: ideas and limits

At 2019 Miao, Iannelli and Gamba [6] publish research with the approach to map urban areas using SAR images from sentinel-1 data and exploiting either Weibo or Twitter data for two cities, Beijing, and Taipei. The study used three measures: namely-precision, recall and F-score to validate the maps obtained by the weighed OCSVM.

They have use one class classifier. The method they propose consists mainly of two parts

- The automatic generation of training samples from geo-referenced social media data: they have considered two different source data, but only one consider by the time. Social media data are point –wise measurements, while remotely senses data rest on a grid, with a given spatial resolution.

  One of the challenges they have face is having duplicated data and that mostly data is coming from urban zones, but not always, so they have study two features, frequency, and similarity of social media data.
- The detection of urban settlement extends using a one class classifier: They have use OCSVM to classify a satellite image and extract urban settlements extends, fits a hyper-sphere with minimum volume that contains the maximum number of training samples of a single class.

The results showed that the approach they suggested is both more accurate and stable with respect to the location. The results depend on the social medium that is used, the classification accuracy obtained by using earth observation is stable and the coverage is uniform independently from geographical location [6].

2.2. Bicycle training data

2.2.1. Identification of land-use characteristics using bicycle sharing data: A deep learning approach

At 2019 Zhao, Fan and Zhai [7] had presented research that used bicycle ridership data and land-use planning data provided by Ford Gobike system and San Francisco planning department in the San Francisco Bay Area, California USA. They proposed a model that combines time series feature extraction and deep neural network proposed in order to identify regional land use characteristics and quantify land use, enable to calibrate the land use planning, and provide dynamics reference for urban spatial structure optimization in the feature.

This research combines time series features extraction and deep neural network, in order to identify regional land use characteristics and quantify land-use intensity based on bicycle sharing data. This type of data normally includes starting or ending point of the original trip chains, which is more useful for land analysis, authors said, that the distribution of shared bicycles is highly coincident with the distribution of hotspot facilities such as urban residents, companies, restaurants, and rail transit stations [7].

They also wanted to target three problems that remain:

- It is not quite common for researchers to use both geographic and traffic ridership data to conduct analysis of urban land use.
- Few studies have quantitatively analyzed the relationship between traffic demand and land use characteristics.
- Neural Networks models are good choices when relationship is existing although difficult to describe and there are considerable number of data samples. Few research

has studied the relationship between transit ridership and urban land use through neural networks.

To fill the gap, they develop a method to identify the urban land use characteristics using GIS and machine learning methods.

The investigation consists of four main parts:

- A set of land-use characteristics labels are evaluated based on planning and geographic information systems (GIS) data
- An ensemble clustering method is used to determine the segmentation points of ridership time series
- The statistical characteristics of the segmentation time series are extracted and used as input to the neural network
- A deep neural network established and trained based on processed ridership features and land-use labels

They have use ArcGIS and MATLAB for the development and KMeans, DNN, VACs, mean square error and R-square as methodologies for different proposes during the development [7].

The approach they develop is able to analyze whether current land use features meet the goal of land use management and planning, the study also provide a valuable new land use analysis method for big data driven land use planning, transportation planning and/or trip distribution research. Experimental results show that the method they proposed can identify land use characteristics effectively and analyze the changing process of metro external characteristics based on the historical ridership data. In addition, were found that the model has better performance for extensions on characteristics of residence, while the effect of work, consumption and transit are sometimes oscillatory and unstable [7].

As disclaimer, they found out that season has a significant impact on the use of sharing bicycles, so, training models based on this type of data of different seasons may lead to more stable conclusion in the extended analysis [7].

3. THEORIC/CONCEPTUAL FRAMEWORK

The following chapter describes the concepts related to the investigation. This section is divided into 7 different areas that are:

1. Introductory Concepts Related to the investigation: Concepts that will help for better understanding, will work as an introduction for the research.
2. Statics Terminology: The elements in this section are related to statics concepts in the research.
3. Methodologies: Concepts related to the methodologies related and used in the investigation.
4. Libraries: Libraries and programming languages used in the development of research.
5. Cloud Computing: Concepts related to cloud computing that will help to explain and understand better the development in the research.
6. Social Media: Elements in this field are the ones related to social media in used in the research
7. Information Technologies: Concepts relate to the area of information technologies that will help to understand better the development.

3.1. Introductory Concepts Related to the investigation

MiBici

MiBici is a System of Public Bicycles and individual transport that provides service every day of the year. [11] This service offers rental of bicycles available in stations located in the most important points of the city. It can be obtained by annual subscription or by a temporary subscription of 1, 3 or 7 days.

MiBici began in 2014 with the sum of efforts of cycling groups and government agencies that have promoted the use of bicycles in the metropolis. Currently the program has 2925 bicycles and 274 stations in the municipalities of Guadalajara, Zapopan and Tlaquepaque [11].

Guadalajara

Guadalajara is the capital and largest city of the Mexican state of Jalisco [12]. It's located in the central region in the Western-Pacific area of Mexico. Guadalajara is the 10th largest city in Latin America and the second most populous metropolitan area in Mexico. The city is named after the Spanish city of Guadalajara, meaning "river/valley of stones."

Guadalajara is the cultural center of Mexico, considered by most to be the home of mariachi music and host to a number of large-scale cultural events such as the Guadalajara International Film Festival and globally renowned cultural events which draw international crowds [12].

Georeferenced

The term "georeferenced" refers to the ability of a digital map or aerial photo's internal coordinate system to be linked to a terrestrial system of geographic coordinates [13]. Users can discover where every point on a georeferenced digital map or aerial shot is located on the Earth's surface by tying it to a known Earth coordinate system.

Though there are numerous alternative ways for performing georeferencing, the appropriate coordinate transforms are often contained inside the image file (GeoPDF and GeoTIFF are examples of georeferenced file formats). Basic map analysis, such as pointing and clicking on the map to find the coordinates of a location, calculating distances and areas, and determining other information, are all possible with georeferencing in the digital file [13].

Trip

A trip is the process of travelling from one place to another, staying there, usually for a brief time, and coming back again [14].

Human Behavior

The way people act and interact is referred to as human behavior. It is impacted and shaped by a variety of elements, including genetic makeup, culture, and personal values and views [15].

Location-Based Services (LBS)

Services based on a mobile user's location as determined by network and/or device-based technology [15]. Advertisements, billing, information, tracking, and safety are just a few of the services that may be provided to mobile device users using location data.

Point of Interest (POI)

A designated geographic entity, such as a milestone, an institute, a heritage site, or a business headquarters, is referred to as a point of interest (POI) [17]. The majority of the data enabling location-based applications is based on points of interest. Housing places, cafes, petrol outlets, parking spaces, tourism attractions, and other POI categories are among the most common. POIs can be both permanent and temporary, such as heritage sites and monuments, or moveable in nature, such as shops and restaurants. They are shown on web maps and made available to users of location-based smartphone apps.

Urban Dynamics

Urban dynamics are the changing elements that make up an urban environment: the opportunities and the threats [18].

3.2. Statistics Terminology

Population

A population is a collection of individuals from the same species who interact in the same area [19].The way people connect with each other and with their surroundings influences a population's health and behavior [19].

Distribution (Population Distribution)

The phrase "population distribution" is used to describe how people are distributed across a certain area. The global population distribution can be measured, or a smaller region within a country or continent can be measured [20].

Spatial-Temporal

Space is referred to as spatial. The term temporal relates to the passage of time. When data is collected over both place and time, the term spatial-temporal or spatial temporal, is used in data analysis [21]. It refers to a phenomenon that occurs at a certain location and time, such as shipping movements over a geographic area over time. By visualizing how objects move in space and time, a person can solve multi-step issues using spatial-temporal thinking.

3.3. Methodologies

Machine Learning

Machine learning is a discipline of artificial intelligence (AI) and computer science that focuses on using data and algorithms to mimic the way humans learn, with the goal of steadily improving accuracy [22].

Algorithms that utilize statistics to discover patterns in massive* volumes of data are known as machine-learning [22] and data here refers to a wide range of items, including numbers, texts, photos, clicks, and so on.

If anything can be saved digitally, it can be fed into a machine-learning system. Machine learning is the technique that underpins many of the services we use today, including Netflix, YouTube, and Spotify's recommendation systems; Google and Baidu's search engines; Facebook and Twitter's social media feeds; and Siri and Alexa's voice assistants [23] and the list goes on and on.

Classification Algorithms

The popular algorithms in supervised learning are called classification algorithms [24]. Classification is a process of setting up the boundary conditions to predict the target class and provides a classifier so as to determine the possible outcome based on the independent variables. For example, in the allocation of the flats to the customers based on the salary range of the customer, type of the locality (urban, semi-urban), previous customer of the company to predict the outcome the best method is applicability of classification techniques.

The most frequently used classification techniques in the literature of ML are Logistic Regression, Decision Trees, Random Forest, Naïve Bayes classifies, K-nearest neighbor, Support Vector Machine and Neural networks [24].

K – Nearest Neighbor (KNN)

The k-nearest neighbors (KNN) technique is a supervised machine learning algorithm that can be used to tackle classification and regression problems [25].

A flexible approach may also be used to fill in missing values and resample datasets. As the name implies, K Nearest Neighbors (Data points) are used to predict the class or continuous value for a new Data point [25].

The KNN algorithm [26] assumes that items that are similar are close together. To put it another way, related items are close together.

The figure 3.1 is an example that shows the similarity data points typically existed close to each other [26].

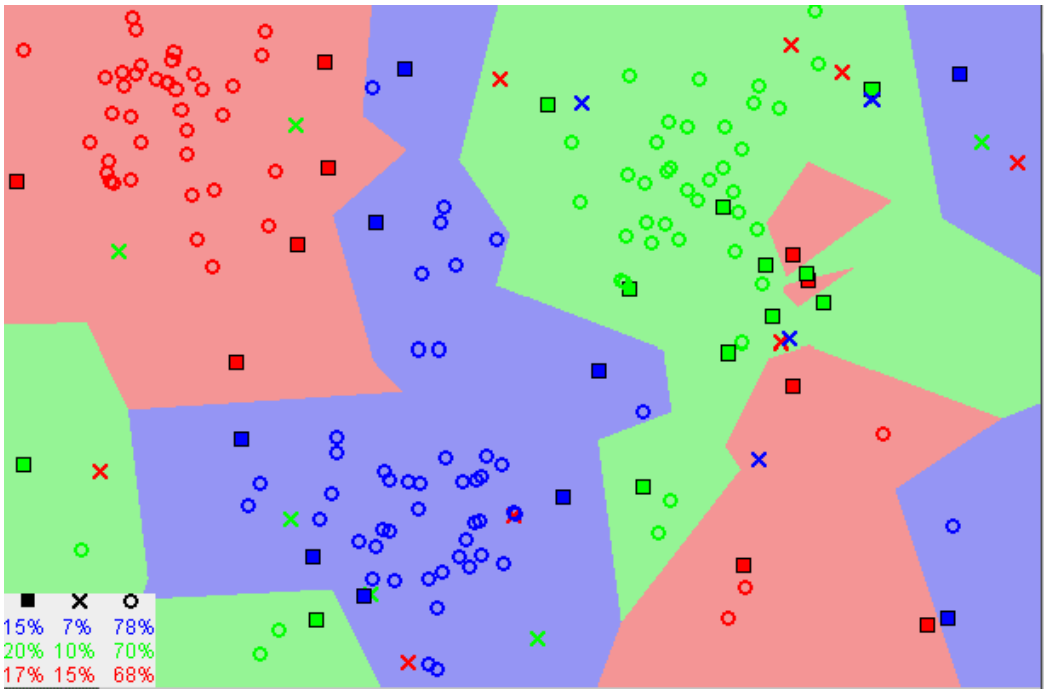


Figure 3.1 Image showing how similar data points typically exist close to each other

Data Analysis

Once the data gathering process from a source has been completed successfully, the next stages are to retrieve important information hidden within it for further interpretation [27]. Data analysis is the process of doing computations and procedures in order to obtain this important information.

Simple data is easily organized, however complicated data necessitates proper processing, which entails recasting and dealing with data until it is ready for analysis. The goal of stated analysis is to summarize the acquired data and organize it in such a way that it answers the initial queries.

The data analysis process must be set up in such a way that conclusions can be drawn for subsequent modification. This also entails properly structuring the data. From research to study, the problem and complexity of data processing differs [27].

Zone Mapping

Zone maps are used to identify locations within a field that have comparable soil qualities (structure, organic content, depth, and drainage), nutrient levels, topography (flatland, rolling hills), and historical crop growth and yield [28].

Elbow Method

In k-means clustering, the elbow approach is used to estimate the ideal number of groups. The elbow technique plots the cost function value produced by various k values [29]. When you might expect, as k rises, average distortion falls, each cluster has fewer constituent examples, and the instances are closer to their respective centroids. As k grows larger, however, the average distortion improves less. The elbow is the value of k at which the improvement in distortion diminishes the most, and at which we should cease dividing the data into more clusters.

The figure 3.2 is an example of elbow method [29]. Disclaimer the figure is not related to any work or investigations is just illustrative.

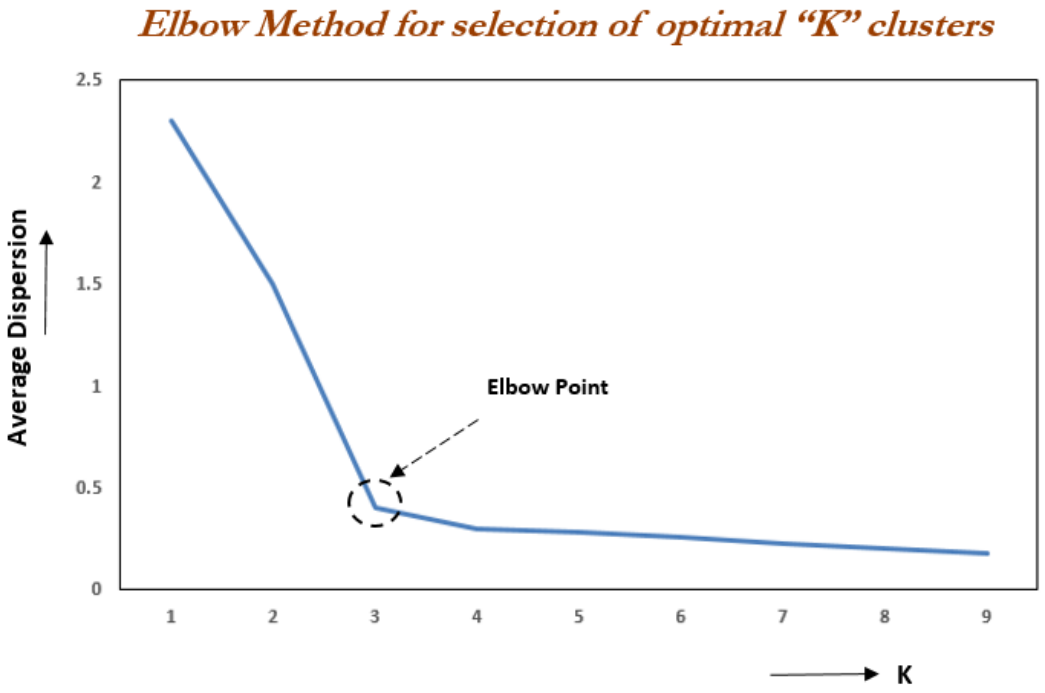


Figure 3.2 Example of Elbow Method

Decision Tree

For classification and regression, Decision Trees (DTs) are a non-parametric supervised learning method [30]. The goal is to learn simple decision rules from data attributes to develop a model that predicts the value of a target variable. A tree is an approximation to a piecewise constant.

In the example below, decision trees use a series of if-then-else decision rules to estimate a sine curve using data. The decision criteria become more complex as the tree grows deeper, and the model becomes more accurate.

The problem is solved using the tree representation, in which each leaf node corresponds to a class label and characteristics are represented on the tree's internal node [31]. The decision tree can represent any Boolean function on discrete attributes.

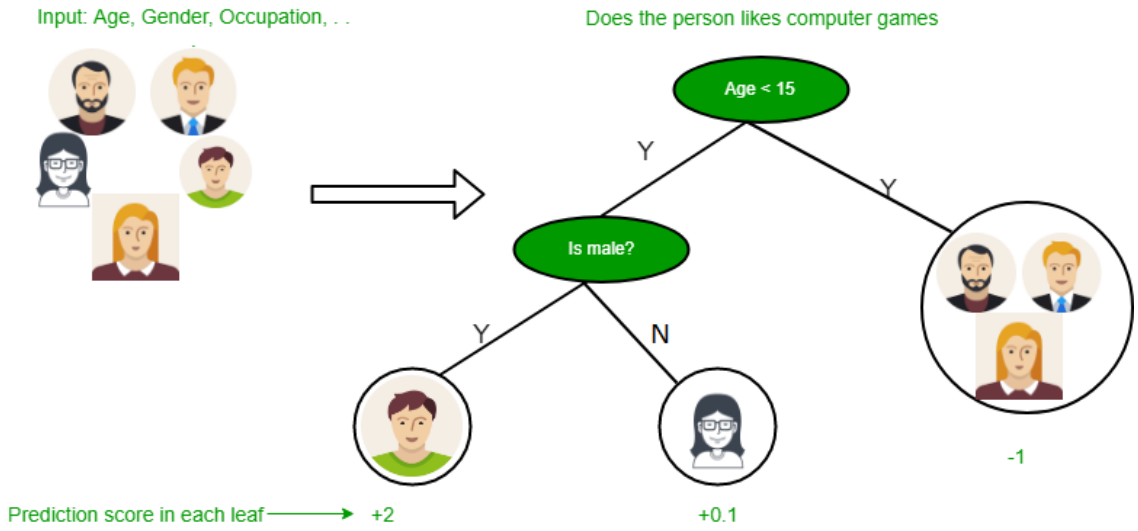


Figure 3.3 Example of Decision Tree

KMeans

K-means is an approach for training a model that groups things that are similar. The k-means algorithm accomplishes this by associating each observation in the input dataset with a point in dimensional space n. (where n is the number of attributes of the observation). Temperature and humidity readings at a specific place, for example, may be stored in your datasets as points (t, h) in a two-dimensional space [32].

Consider the case where you wish to train a model to recognize handwritten digits using the MNIST dataset. Thousands of handwritten digit images are available in the dataset (from 0 to 9). You can opt to make ten clusters in this example, one for each digit (0, 1, ..., 9). The k-means technique divides input photos into ten clusters during model training. The MNIST dataset contains 784 images, each of which is a 28x28 pixel image. Each image represents a point in a 784-dimensional space, which is analogous to a two-dimensional point (x,y). The k-means method searches for the distance between a point and all of the cluster's centers to determine which cluster it belongs to, next, select the cluster with the closest center as the image's home cluster [32].

Data Cleaning

Most individuals think that your insights and analytics are only as good as the data you're employing when it comes to data [33]. The garbage data that comes in is essentially garbage analysis. If you want to build a culture around making excellent data decisions, data cleaning/data scrubbing is one of the most significant steps you can do.

The practice of cleaning or eliminating erroneous, corrupted, poorly packaged, duplicated, or incomplete data from a dataset is known as data cleansing. There are several ways for data to be duplicated

or wrongly categorized when different data sources are combined. The results and algorithms are unreliable if the data is inaccurate, even if they appear to be correct [33].

Because the methods will differ from one dataset to the next, there is no way to specify the exact phases in the data cleansing process. However, it's critical to create a template for your data cleansing process so you can be sure you're doing it correctly every time.

Unsupervised Methods

You don't have to monitor or exchange the labeled data with the model if you use this method or strategy [34]. Instead, the model's algorithm will recognize the data and begin learning from it without any assistance. Due to the algorithm's design, the model will employ unlabeled data to detect new patterns and information.

Using this method, we can identify new and previously unrecognized data. This form of learning is comparable to how people learn. Consider how we collect data, learn, and recognize objects by analyzing and observing the surroundings. Machines that use unsupervised learning algorithms discover patterns in order to produce valuable results. For example, by knowing both the qualities and characteristics of the animals, the system can distinguish between cats and dogs [34].

Unsupervised algorithms operate without the need for any prior training. It starts working as soon as the data is received. The algorithm takes its own conclusions and determines how to categorise the variables and determine whether they are compatible. Another advantage of this strategy is that no labeled data is required. The system will look at the data and create rules based on what it finds [34]. In an unsupervised learning method, the output has a defined working process.

3.4. Libraries

GIT

Git is a distributed version management system that is free and open-source with the cooperation of several contributors; it keeps track of projects and files as they change over time [35].

Git is a program that keeps track of code changes if you get into a fatal error when coding and don't know what's causing it, Git allows you to revert to a stable state. It also allows you to observe how the code has evolved over time [35].

Python

Python is a dynamically semantic, interpreted, object-oriented high-level programming language. Its high-level integrated data structures, together with dynamic typing and dynamic linking, make it ideal for rapid application creation, as well as a scripting language or glue for connecting existing components [36].

The syntax is straightforward and simple to grasp. It promotes readability, which lowers the program's maintenance costs. Modules and packages are supported by Python, which fosters program

modularity and code reuse [36]. The Python interpreter and substantial standard library are free to download in source or binary format for all major systems and can be used without restriction.

NumPy

One of the most popular Python packages for scientific computing is NumPy. It's a Python library that includes a multidimensional array object, derived objects (masked arrays and matrices), and a set of routines for performing fast array operations, such as mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation, and other useful libraries [37].

Keras

Keras is a Python-based deep learning API. It is built on top of TensorFlow, a machine learning platform. It was created with the goal of allowing for quick experimentation. The API "follows best practices for decreasing cognitive burden" and was "built for human beings, not machines." Individual modules like as neural layers, cost functions, optimizers, initialization techniques, activation functions, and regularization schemes can be combined to form new models. New modules, like new classes and methods, are straightforward to add. Models are defined in Python code rather than separately configured model files [38].

Tensor Flow

Tensorflow is an open-source library for numerical computation and large-scale machine learning developed by the Google Brain team. TensorFlow combines a variety of machine learning and deep learning models and algorithms into a single metaphor that makes them useable [39]. It uses Python to give a handy front-end API for constructing framework programs, but it executes them in C++.

Deep neural networks for handwritten digit classification, image recognition, word embeddings, recurrent neural networks, sequence-to-sequence models for machine translation, natural language processing, and PDE (partial differential equation) based simulations can all be trained and run with TensorFlow [39].

Dataflow graphs—structures that explain how data passes through a graph, or a sequence of processing nodes—can be created with TensorFlow. Each node in the graph represents a mathematical process, and each link or edge between nodes is a tensor [39], which is a multidimensional data array.

Pandas

Pandas is a widely used open-source Python library for data science, data analysis, and machine learning activities [40]. It is based on NumPy, which allows you to work with multi-dimensional arrays within the Python ecosystem, it integrates effectively with a variety of different data science components

Pandas has libraries and functions to assist with the following activities, which are popular in data analysis and data science projects:

- Data cleansing
- Data fill
- Data normalization
- Mergers and joins
- Data visualization
- Statistical analysis
- Data inspection
- Loading and saving data

Sklearn

Scikit-learn is a Python module that integrates a wide range of cutting-edge machine learning methods for supervised and unsupervised issues on a medium-scale. This package focuses on using a general-purpose high-level language to bring machine learning to non-specialists. Ease of use, performance, documentation, and API consistency are all prioritized. It has few dependencies and is released under a simplified BSD license, making it suitable for usage in both academic and commercial environments [41].

Jupyter Notebook

Jupyter Notebook [42] is an interactive computing environment that allows users to create notebook documents that can include:

- Live code
- Interactive widgets
- Plots
- Narrative text
- Equations
- Images

These documents provide a complete, self-contained record of a calculation that can be converted to various formats and shared with others via email, Dropbox, version control systems (such as git/GitHub), or nbviewer.jupyter.org [42].

The Jupyter notebook [42] combines three components:

- The notebook web application: an interactive web application for writing and executing code interactively and creating notebook documents.
- Cores: Separate processes initiated by the laptop's web application that execute users' code in a certain language and return the output to the laptop's web application. The kernel also handles things like calculations for interactive widgets, tab completion, and introspection.
- Notebook documents: Stand-alone documents that contain a representation of all content visible in the notebook web application, including calculation inputs and outputs, narrative text, equations, images, and rich multimedia representations of objects. Each notebook document has its own core.

3.5. Cloud Computing

Amazon Web Services (AWS)

Amazon Web Services (AWS) is the most widely used and comprehensive cloud platform in the world, with over 200 data center services available worldwide [43]. AWS is used by millions of clients, including the fastest-growing startups, the largest corporations, and the most powerful government agencies, to cut costs, improve agility, and innovate quicker.

AWS offers more services and features than any other cloud provider, including infrastructure technologies like compute, storage, and databases, as well as new technologies like machine learning and artificial intelligence, data lakes and analytics, and the Internet of Things [43]. This allows you to construct practically anything you can think by making moving current apps to the cloud faster, easier, and more cost-effective.

Within those services, AWS also provides the most extensive functionality. AWS, for example, has the most databases that are purpose-built for various sorts of applications, allowing you to pick the ideal tool for your project in terms of cost and performance [43].

Bucket

A bucket is a container for objects. An object is a file and any metadata that describes that file [44].

To store an object in Amazon S3, you create a bucket and then upload the object to the bucket. When the object is in the bucket, you can open it, download it, and move it. When you no longer need an object or a bucket, you can clean up your resources.

Amazon Quick Sight

Amazon Quick Sight is a scalable, serverless, integrable, machine learning-powered business intelligence (BI) service built for the cloud. Quick Sight allows you to easily create and publish interactive business intelligence dashboards that include insights powered by machine learning [45]. You can access Quick Sight dashboards from any device and easily integrate them into apps, portals, and websites.

Quick Sight is a serverless service and can be automatically scaled to tens of thousands of users without having to manage infrastructure or plan capacity. In addition, it is the first business intelligence service to offer pay-per-session pricing, so it only pays when users access its dashboards or reports and is therefore cost-effective for large-scale deployments [45].

With Quick Sight, you can ask business questions related to your data in simple language and receive answers in a matter of seconds.

Advantages:

- Easy scalability from tens to tens of thousands of users
- Access to more detailed information with Machine Learning

- Integration of business intelligence dashboards into your applications
- Ability to analyze based on questions related to business data

Amazon S3

Amazon Simple Storage Service (Amazon S3) is an object storage service that delivers industry-leading scalability, data availability, security, and performance [46]. With Amazon S3, customers of all types and industries can store and protect any volume of data for a variety of purposes, such as using it in data lakes, websites, mobile apps, backup and restore processes, archiving operations, enterprise applications, IoT devices, and big data analytics. Amazon S3 provides easy-to-use management features that allow you to organize your data and configure sophisticated access controls to meet your business, organizational, and compliance requirements.

AWS Glue DataBrew

AWS Glue DataBrew is a visual data preparation tool that lets you clean and normalize data without having to write any code. When compared to custom-developed data preparation, DataBrew can cut the time it takes to prepare data for analytics and machine learning (ML) by up to 80% [47]. To automate data preparation chores like filtering anomalies, transforming data to standard formats, and correcting erroneous values, you can choose from more than 250 out-of-the-box transformations.

Business analysts, data scientists, and data engineers can work together more readily with DataBrew to gain insights from raw data. Because DataBrew has no server, you may explore and transform gigabytes of raw data without having to cluster or manage any infrastructure, regardless of your technical expertise.

You can interactively explore, visualize, cleanse, and alter raw data with DataBrew's easy interface. DataBrew delivers intelligent ideas to assist you in identifying data quality issues that can be difficult to locate and resolve. You can use your time to act on results and iterate faster with DataBrew prepping your data. You can preserve the transformation as steps in a recipe, which you can alter or reuse with new datasets in the future and distribute indefinitely [47].

Amazon SageMaker

Amazon SageMaker is a managed service that allows data scientists and developers to easily construct, train, and deploy machine learning models [48]. SageMaker makes it easier to generate high-quality models by automating the tedious processes of every phase of the machine learning process.

Amazon SageMaker is built to be highly available. There are no scheduled maintenance periods or downtime. The SageMaker API is hosted in Amazon's proven, highly available data centers, with the service stack replicated over three facilities in each AWS Region to ensure fault tolerance in the event of a server failure or an Availability Zone outage [48].

AWS ML Insights

Amazon Quick Sight uses machine learning to help you uncover hidden insights and trends in your data, identify key factors, and predict business metrics [49]. You can also consume this information

in natural language narratives built into the dashboards. With ML Insights, Amazon Quick Sight provides three main features:

- Anomaly detection based on machine learning
- Machine learning-based forecasting
- Automatic narrations

Spice

Because of how data is kept in Quick Sight, when you import data into a dataset instead of doing a direct SQL query, it is transformed to SPICE data. Amazon Quick Sight's SPICE computation engine is a fast-in-memory, parallel, and in-memory compute engine [50]. It's made to do sophisticated calculations and produce results rapidly. Data stored in SPICE is encrypted at rest in the Enterprise edition. Unless the dataset contains uploaded files, you can use SPICE or a direct query while creating or editing it. Data import (also known as ingest) into SPICE can save you time and money because:

- Your analytical queries are processed faster.
- No need to wait for a direct query to be processed.

SPICE data can be reused indefinitely without acquiring extra costs.

3.6. Social Media

Social Media

Through the creation of virtual networks and communities, social media is a computer-based technology that allows the exchange of ideas, opinions, and information. Social media is Internet-based by design, allowing people to share content quickly via electronic means. Personal information, documents, movies, and images are all included in the content. Users interact with social media using web-based software or applications on a computer, tablet, or smartphone [51].

Twitter

Twitter is a free social networking microblogging service that allows registered members to broadcast short posts called tweets. Twitter members can broadcast tweets and follow other users' tweets by using multiple platforms and devices [52].

3.7. Information Technologies

Wizard

A wizard is a piece of software that simplifies difficult activities or guides a user through them. The software wizard, as a type of user interface, could be described as a "digital tutorial" or "online (or desktop) guide" that assists users in achieving their objectives [53].

Wizards, in general, employ a variety of technologies to "guide" human users through a procedure. Many of these wizards are made up of specialized forms in object-oriented programming languages that serve as the framework for the instruction [53]. The wizard may employ menus, drop-down lists, checkboxes, command buttons, and other tools to automate or guide a process through this graphical user interface.

Job

A Job in SQL Server is a container that allows packaging one or more steps in process that need to be scheduled and executed. We can say, it a series of query actions that SQL Server performs [54].

DataSet

A file containing one or more records can be referred to as a data set. A data set is a named collection of records. Data sets can be used to store information such as medical records or insurance records for usage by a system software [55]. Data sets can also be used to store information that is required by applications or the operating system, such as source programs, macro libraries, or system variables or parameters.

You can print or show data sets that contain readable text on a console if they contain readable text (many data sets contain load modules or other binary data that is not printable). Data sets can be cataloged, allowing them to be referred to by name without having to identify where they are kept [55].

4. DEVELOPMENT METHODOLOGY

The next chapter will talk about the process and the steps that were followed to carry out the investigation. Which consisted of seven features that were distributed in three distinct stages.

This part of the research focused a lot on understanding the data in order to manipulate them and once a clear idea of what could be done with them was had, an attempt was made to apply different classification algorithms to be able to reach more solid conclusions based on our environment. An implementation was also carried out using cloud services to evaluate this solution, without neglecting that it was also used to help better implement the data obtained from the investigation.

4.1. How the idea was born – Background

Now day's data is particularly important and the way you use it more. In a world where we are constantly searching for the newest force us to be data driven in many areas of our daily basic activities. It is important for people to understand better how the world is moving, how we as humans drive the world. Ideally, the project was born from the idea to use social media to map an urban area of a city. According to Z. Miao [6]. There is an increasing trend at using social media data to map human activity, to the point that some authors have suggested that these data may be even better than Earth Observation (EO) data to map urban areas [1].

Where he also mentions that the adoption of specific social platform depends on a number of geographical, economic, and political factors, while remote sensing is based on physics [1].

The proliferation of the use of social media provides every user with rich digital contents, such as geographic coordinates, photographs, videos, resulting into significant sources of" big data." Although massive data from social media are often published without scientific intent, often carry little scientific merit, and the use of social media tools is not uniformly distributed across the world, they provide a source of useful information that may help to avoid, for instance, long and expensive in situ data collection for remote sensing image classification [1].

However, it was shown according to previous works, that first needs to be defined which source of information is going to be the best for your investigation. For example, if in the area where you are planning to investigate Twitter is not that much used it does not suit you, but on the other hand, you found that the data is rich and will contribute to your work you should use it.

In our case in Guadalajara City, social media data coming from Twitter does not contain a good amount of georeferenced for the study, that is the reason MiBici was a reliable source of data.

According to Jiahui Zhao, Ample evidence have shown that there is a strong relationship between travel demand and land use characteristics, especially in the urban area with significant functional division [5].

4.2. Requirements

The Requirements for this project was following agile methodology, so Features will be used to plasm the needs that we want the investigation to have, Benefit Hypothesis as the gain the functionality will bring to the user and acceptance criteria as the conditions to mark the feature as complete

Table 4.1 Features to Develop in the project

ID	Feature	Benefit Hypothesis	Acceptance Criteria
01	Analyze source data	Understand data will confirm if the source data is good to use for the investigation	• Able to understand data. • Able to define a cleaning process so

			data could be useful.
02	Clean Source Data and Create One complete DataSet	Having a clean and complete dataset will speed up the development process	• One file • No Null data • No records/fields that will not add to the investigation
03	Apply supervise classification methods to data	Apply classification methods that give results to have a reference point from where to start	• Output from the algorithms applied.
04	Discuss the results from supervised methods	Discuss the results obtained on the supervised methods will allow evaluating their performance and creating new hypotheses about the research.	• New Hypothesis (could be only the idea) for the innovative approach of the investigation • Plan of the steps to take in order to start the new line of investigation
05	Refine data used with unsupervised classification methods	Having a clean, complete, and fit dataset will speed up the development process	• Have complete dataset that applies according to the new method to use
06	Discuss the results from unsupervised methods	Understanding the results will allow new hypotheses to be made in the research based on the approach of the environment and the scope of the same can be decided	• Defined scope of the investigation • Results report from the investigation

07	Validate results with AWS	Having a cloud implementation will allow evaluating the tool for possible research, as well as helping to implement graphics that help with the understanding of the results obtained in the previous stage.	• Concluding thoughts of the investigation, results and conclusion explained

4.3. Analyze Source Data, Dataset Prepossessing and Cleaning

The data comes from MiBici, they provide a compile every month with the trips that occur during this time. This data could be found in [10].

The Investigation was divided into 3 stages:

1. Analyzed data only for the period of time January 2019 – December 2019 where the algorithms that were applied were decision tree and KNN.
2. Analyzed data since MiBici has information on the site at the time this investigation took place that was from December 2014 – December 2020 where K-Means classification algorithm was applied.
3. Analyzed data from December 2014 – December 2020 using tools from AWS as assistant to interpret the previous source data from the analysis.

The data was provided in csv format and contains the following fields:

- Viaje_id: Id of the trip. Every Trip have their own identification.
- Usuario_id: Id of the user creating the trip. For being able to use the application the user need to have an ID.
- Genero: Gender of the user that is taking the trip.
- Año de Nacimiento: Year of born of the user taking the trip.
- Inicio_del_viaje: day and time when the trip started.
- Fin_del_viaje: day and time when the trip finished.
- Origen_id: Id of the station where the trip started. Every station has its own id.
- Destino_id: Id of the station where the trip finished. Every station has its own id.

In addition, the site provides the way they classify the stations with the ID. That looks like the table 4.2

Table 4.2 Stations that are part of MiBici

ID	Nombre de la Estación	ID Estación	Localización	Latitud	Longitud	Estado
2	(GDL-001) C. Epigmenio Glez./ Av. 16 de Sept.	GDL-001	POLÕGONO CENTRAL	20.666378	-103.34882	IN_SERVICE
3	(GDL-002) C. Colonias / Av. NiÒos hÈroes	GDL-002	POLÕGONO CENTRAL	20.667228	-103.366	IN_SERVICE
4	(GDL-003) C. Vidrio / Av. Chapultepec	GDL-003	POLÕGONO CENTRAL	20.66769	-103.368252	IN_SERVICE
5	(GDL-004) C. Ghilardi /C. Miraflores	GDL-004	POLÕGONO CENTRAL	20.69175	-103.36255	IN_SERVICE
6	(GDL-005) C. San Diego /Calzada Independencia	GDL-005	POLÕGONO CENTRAL	20.681151	-103.338863	IN_SERVICE
8	(GDL-006) C. Venustiano Carranza /C. Reforma	GDL-006	POLÕGONO CENTRAL	20.6807519	-103.3443801	IN_SERVICE
9	(GDL-007)C. Epigmenio Glez./Av. Cristobal C.	GDL-007	POLÕGONO CENTRAL	20.666771	-103.350562	IN_SERVICE
10	(GDL-008) C. J. Angulo / C. Gonz·lez Ortega	GDL-008	POLÕGONO CENTRAL	20.681871	-103.350396	IN_SERVICE
11	(GDL-009) Calz. Federalismo/ C. J. Angulo	GDL-009	POLÕGONO CENTRAL	20.681984	-103.353835	IN_SERVICE
12	(GDL-010) C. Cruz verde / C. Joaquin Angulo	GDL-010	POLÕGONO CENTRAL	20.681786	-103.357267	IN_SERVICE
13	(GDL-011) C. Garibaldi / C. FrÌas	GDL-011	POLÕGONO CENTRAL	20.681081	-103.360099	IN_SERVICE
14	(GDL-012) C. JoaquÌn Angulo / C. Ghilardi	GDL-012	POLÕGONO CENTRAL	20.681989	-103.36292	IN_SERVICE
15	(GDL-013) C. Vidrio / C. Marsella	GDL-013	POLÕGONO CENTRAL	20.667623	-103.370499	IN_SERVICE
16	(GDL-014)C. J. Angulo/C. JosÈ Clemente Orozco	GDL-014	POLÕGONO CENTRAL	20.682059	-103.365969	IN_SERVICE
17	(GDL-015) C. Herrera y Cairo /C. Pedro Buzeta	GDL-015	POLÕGONO CENTRAL	20.683191	-103.369349	IN_SERVICE
18	(GDL-016)C. Bernardo de Balbuena/C. J. Angulo	GDL-016	POLÕGONO CENTRAL	20.682007	-103.372642	IN_SERVICE
19	(GDL-017) Av. MÈxico /C. Bernardo de Balbuena	GDL-017	POLÕGONO CENTRAL	20.679086	-103.372841	IN_SERVICE
20	(GDL-018) Av. MÈxico / C. Manuel M. DiÈguez	GDL-018	POLÕGONO CENTRAL	20.679296	-103.370658	IN_SERVICE
21	(GDL-019) C. Juan Manuel / C. AndrÈs Ter·n	GDL-019	POLÕGONO CENTRAL	20.678835	-103.368045	IN_SERVICE
22	Ex-GDL-020 Juan Manuel / C.General Coronado	Ex-GDL-020	POLÕGONO CENTRAL	20.678714	-103.36573	NOT_IN_SERVICE
23	(GDL-021) C. Nicol·s Romero / C. Reforma	GDL-021	POLÕGONO CENTRAL	20.67966	-103.362224	IN_SERVICE
24	(GDL-022)C. San Felipe /C. Enrique D. de LeÛn	GDL-022	POLÕGONO CENTRAL	20.679389	-103.35913	IN_SERVICE
25	(GDL-023) C. Jes¨s / C. San Felipe	GDL-023	POLÕGONO CENTRAL	20.679309	-103.356281	IN_SERVICE
26	(GDL-024) Av. Federalismo / C. Juan Manuel	GDL-024	POLÕGONO CENTRAL	20.678777	-103.354035	IN_SERVICE
27	(GDL-025) C. Reforma / C. Gonz·lez Ortega	GDL-025	POLÕGONO CENTRAL	20.680227	-103.350146	IN_SERVICE

Where:

- ID: is the identifier that the station has. Every station has its own Identifier.
- Nombre de la Estacion: name of the station is the name that the station has. Normally the address where the stations are located with concatenation of the ID station.
- ID Estacion: is the identifier of the station, normally has an abbreviation of in which municipality is located and number to identify.
- Localizacion: localization of the station in the city according to the classification of the station in the city could be POLIGONO-CENTRAL, TLQ-CORREDORATLAS, ZAPOPAN CENTRO.
- Latitud: latitude where the station is located.
- Longitud: longitude where the station is located.
- Estado: status of the station, the values could be IN_SERVICE or NOT_IN_SERVICE

As a note all, the code used for the investigations can be found in the following git repository [9]

4.4. Stage 1 – (KNN and Decision Trees)

Once the origin of the data and the fields are defined, the next step was to try to analyze it in order to define what could be interpreted from the data. As mentioned before, for this stage of the investigation the data collected for the period January 2019 to December 2019.

For this stage divided into 3 main steps, that were the following listed below:

4.4.1. Analyze Data

The total number of trips during this period of time was 4631777. From this number of trips were deleted the ones with null data, so the new total trips used for the analyze were 4620166.

Table 4.3 Interesting Data of the total Trips

Interesting Data of the total Trips		
Measure	Type	Data
Mean	Birth Year	1987
Mean	Age of the Users	32 Years
Mode	Origin Station Most Used	51
Mode	Destination Station Most Used	51

Figure 4.0 explains the distribution between the birth years of the users that used the service.

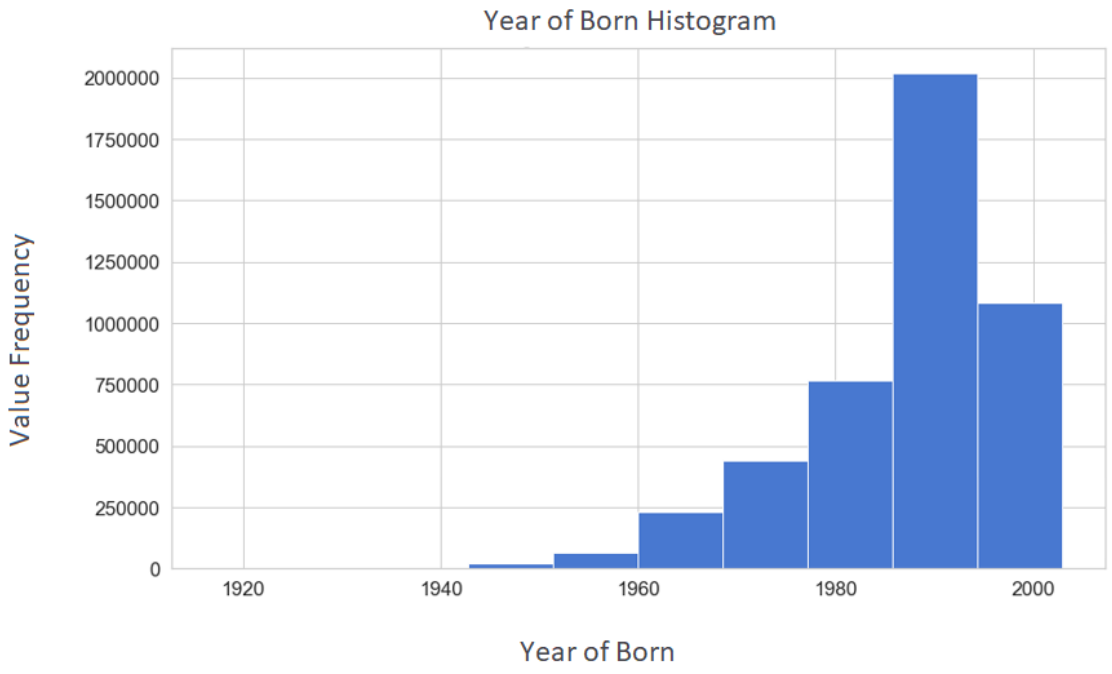


Figure 4.1 Year of Birth Histogram

The data from 51st the station is:

Nombre de la Estación	(GDL-049) Lopez Cotilla/ Marcos Castellanos
ID Estación	GDL-049
Localización	POLÕGONO CENTRAL
Latitud	20.6741
Longitud	-103.356
Estado	IN_SERVICE

Figure 4.2 Details from the station 51st

The following graphic is related to the trips created per month

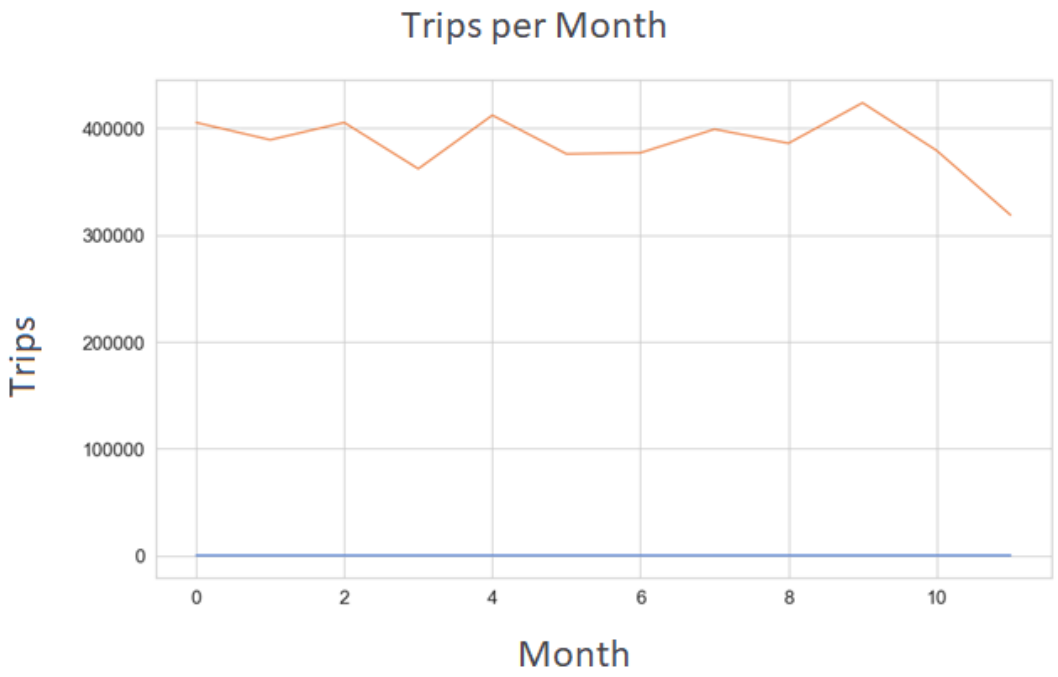


Figure 4.3 Trips per Month Graphic

Table 4.4 Distribution of the trips per Month

	Mes	Viajes
0	1	405169
1	2	388993
2	3	405169
3	4	361883
4	5	412001
5	6	375834
6	7	376822
7	8	398916
8	9	385766
9	10	423640
10	11	379046
11	12	318538

4.4.2. Clean Data

For the process of cleaning the data was need it first read the 12 files that corresponds the 12 months of the year after that was created one consolidated file, it worth to mention that was need it to rename some files because none of all match a standard format so for in order to make more agile the process.

To the consolidated file needed to change to date time the fields that apply (Fecha Inicio Viaje, Hora Inicio Viaje and Fecha Fin Viaje, Hora Fin Viaje), deleted the NaN values and added a new field:

- Viaje: The sum of trips that occur every day
- Tiempo: The difference between Hora Fin Viaje and Hora Inicio Viaje

Also, it was needed to create classifications for the fields

- Temporada, that goes from 1 - winter, 2 – spring, 3 – summer, 4 – fall

- Tiempo, that goes from 0 - 3 and is the time that the trip took from the beginning to the end
- Viajes, that goes from 0 – 2 and is the number of trips per day

Detailed categorical values created for:

- Fecha Inicio Viaje: were group by month 0-12
- Fecha Fin Viaje: were group by month 0-12
- Tiempo: were group in 4 categories
    - Class 0 lees or 4 minutes
    - Class 1 more than 4 less than 6
    - Class 2 more than 6 less than 8
    - Class 3 more than 8

- Viajes: were group in 3 categories
    - Class 0 lees or 2616 trips
    - Class 1 more than 2616 less than 10685
    - Class 2 more than 10685

4.4.3. Implementation of Algorithms

With the data set complete a decision tree and K-NN was built with the following characteristics and in the following chapters, the results will be discussed.

Table 4.5 Characteristics of the algorithms used

Decision Tree	KNN
Y = Viajes	Y = Viajes
Max_depth = 2	K = 4
Min_samples_leaft = 5	

For further details, refer to git repository [9]

It is important to mention that it created a correlation matrix in order to try understanding the relation between variables with the purpose of utilizing the variables that will attribute more meaning to the investigation.

The implementation of the algorithms follows the rule of 80% for training and 20% for test the model, this means, that 80% of the data set will be used for training the algorithm and 20% of it will be used for test how good the implementation was. Please note that before doing the split all data were randomly sorted in order not to hinder the investigation.

For both Decision Tree and K-NN, the best option for parameters was chosen following different methods that applied accordingly.

4.5. Stage 2 (KMeans)

For the second stage I considered taking the data from December 2014 to December 2020. For this was need it to once again to read all the files per month in order to make a consolidated file with all the records from this period, with the originals fields as mentioned in a step before.

4.5.1. Prepare Data Set

The first step was step was to identify the records with null data. The following list was created:

```
Missing Viaje_Id data : 850845
Missing Usuario_Id data : 850845
Missing Genero data : 949000
Missing Año_de_nacimiento data:  98395
Missing Inicio_del_viaje data : 850845
Missing Fin_del_viaje data : 850845
Missing Origen_Id data : 850845
Missing Destino_Id data : 850845
```

Figure 4.4 Records with null data in the DataSet

For gender was decided to change null values for undefined and for 'Año_de_nacimiento' changed to -1 value. All the other rows with null values dropped.

Once the data was free of null data the next step was to separate de fields 'Inicio_del_viaje' and 'Fin_del_viaje' to date and time accordingly the new fields created were:

- Fecha_inicio_del_viaje: date in format yyyy-mm-dd when the ride starts
- Hora_inicio_del_viaje: time in format hh:mm:ss when the ride starts
- Fecha_fin_del_viaje: date in format yyyy-mm-dd when the ride end
- Hora_fin_del_viaje: time in format hh:mm:ss when the ride end

The next step was to create a new field called 'Tiempo' that refer to the difference between the fields 'Hora_fin_del_viaje' and 'Hora_inicio_del_viaje' with the purpose to have a record of the time spend in every ride.

Figure 4.5 is a representation of the final data, which left 14,150,129 rows

	Viaje_Id	Usuario_Id	Genero	Año_de_nacimiento	Fecha_inicio_del_Viaje	Hora_inicio_del_viaje	Fecha_fin_del_viaje	Hora_fin_del_viaje	Origen_Id	Destino_Id	Tiempo
0	4601.0	1436.0	undefined	-1	2014-12-01	00:33:47	2014-12-01	00:36:54	47.0	47.0	0 days 00:03:07
1	4604.0	1436.0	undefined	-1	2014-12-01	01:06:54	2014-12-01	01:08:45	5.0	5.0	0 days 00:01:51
2	4628.0	1.0	M	1990	2014-12-01	09:47:20	2014-12-01	09:47:47	79.0	79.0	0 days 00:00:27
3	4631.0	22.0	M	1982	2014-12-01	09:48:23	2014-12-01	09:48:37	79.0	79.0	0 days 00:00:14
4	4632.0	1.0	M	1990	2014-12-01	09:48:46	2014-12-01	09:48:57	79.0	79.0	0 days 00:00:11
...	...	...	...	...	...	...	...	...	...	...	...
14150124	18023308.0	46055.0	M	1992	2020-12-31	13:16:26	2020-12-31	13:25:07	266.0	265.0	0 days 00:08:41
14150125	18022940.0	59898.0	M	1991	2020-12-31	12:20:16	2020-12-31	12:48:48	75.0	216.0	0 days 00:28:32
14150126	18024152.0	70096.0	F	1991	2020-12-31	14:57:24	2020-12-31	15:04:32	13.0	33.0	0 days 00:07:08
14150127	18024749.0	70096.0	F	1991	2020-12-31	16:10:01	2020-12-31	16:13:36	16.0	13.0	0 days 00:03:35
14150128	18026137.0	70481.0	M	2001	2020-12-31	20:54:41	2020-12-31	21:07:35	168.0	28.0	0 days 00:12:54

14150129 rows × 11 columns

Figure 4.5 Final DataSet

To the final dataset four different variations were applied for the analysis as listed below:

- Origen and Destiny
- Only Origen
- Origen and Destiny Difference
- Only Destiny

4.5.2. Implement Algorithm

All this with the approach to evaluate the impact when using different fields to test the algorithm. To all of the different analysis first where applied “The elbow Method” to identify the best number of clusters to use. Once the number of clusters defined the K-Means algorithm applied to the data that were set on the steps before.

4.6. Stage 3

The third stage implemented using AWS and figure 4.6 is a diagram of the implementation of this solution

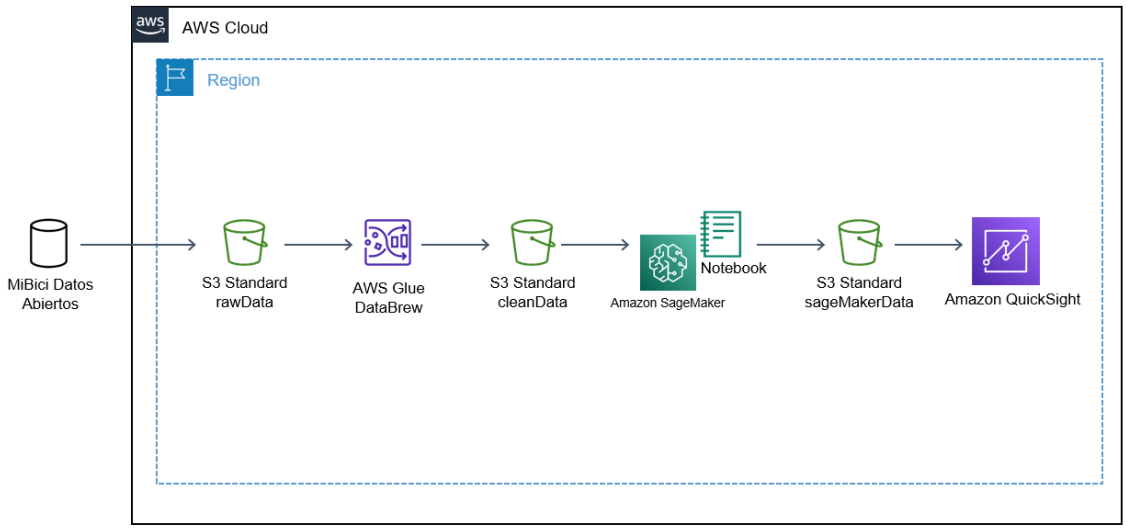


Figure 4.6 Diagram of the implementation at AWS

Basically, in this implementation there were 3 fundamental steps where:

- First was, need it to upload to a bucked all the raw data that was collected from MiBici Data Base to the data was need to applied conversion using AWS Glue DataBrew and the result from the transformation was applied to another bucket.
- Second to the clean data of the second bucked was consumed by a notebook in the service of Amazon SageMaker the result was applied to another bucked.
- Third, the bucked with the result from the model was consumed and interpreted by Amazon Quick Sight.

4.6.1. Prepare Data

For the first step as mentioned before it was needed to create a bucket and upload the files as listed in figure 4.7

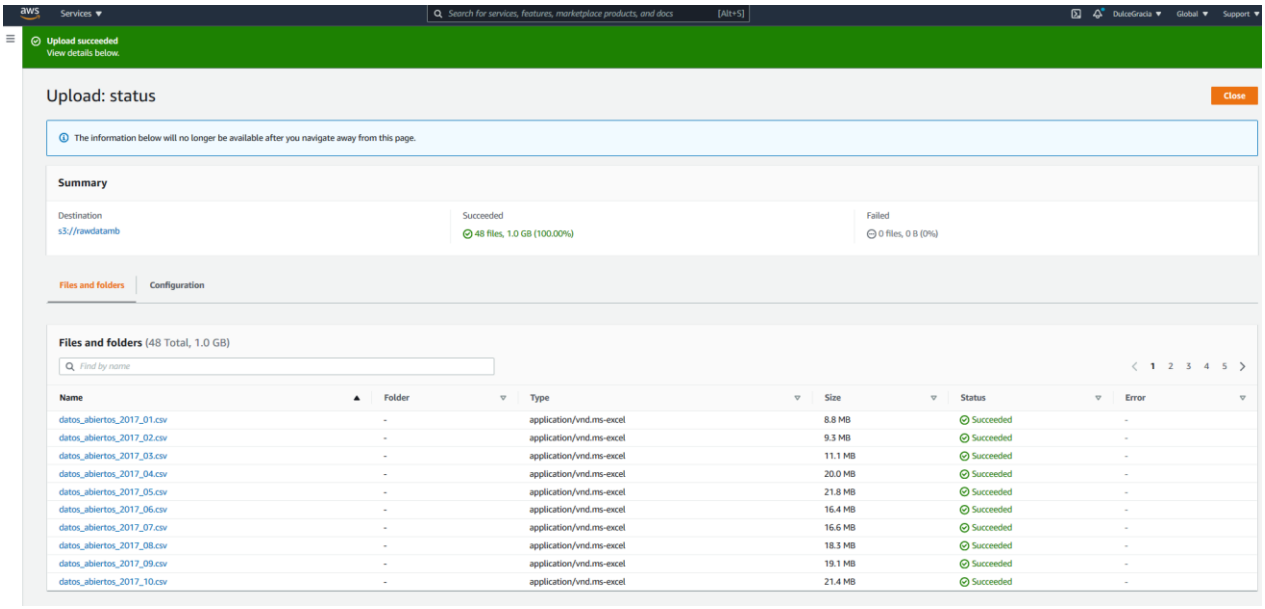


Figure 4.7 Data part of the Bucked Raw Data

With the files in the raw data bucket was need it to start to analyze using AWS Glue Data Brew, for this was need it to create a new project and applied the transformations listed in the recipe that can be found at the git repository [9].

Figure 4.8 is an example of the clean data after the steps from the recipe were applied.

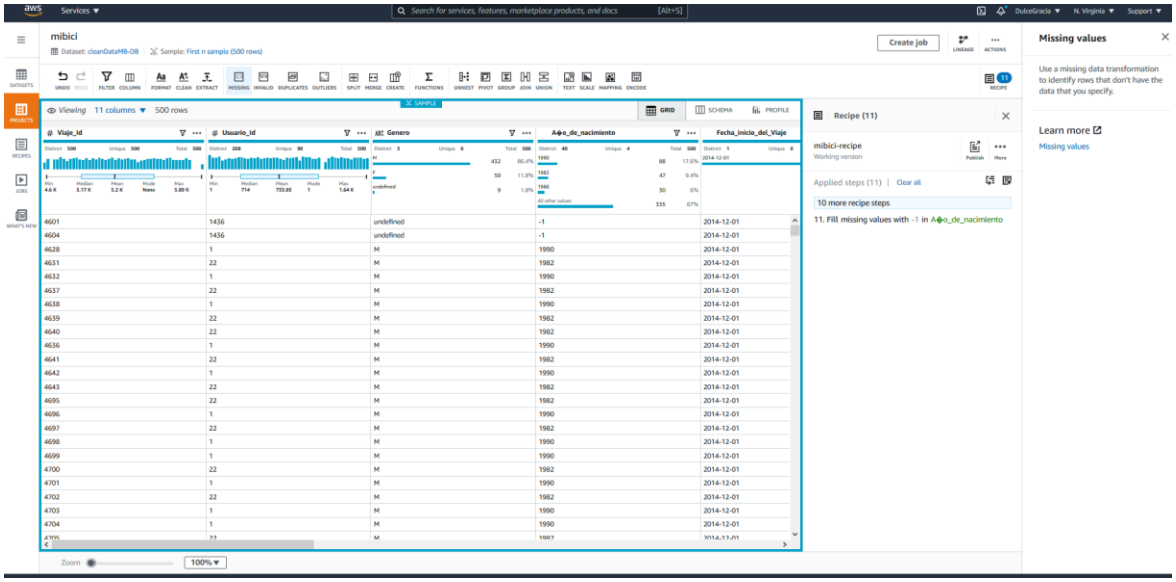


Figure 4.8 DataSet to be clean using AWS Glue Data Brew

Once the data is clean it is needed to be passed it to a new bucket using the wizard from the tool to create a job to do this step as listed in figure 4.9

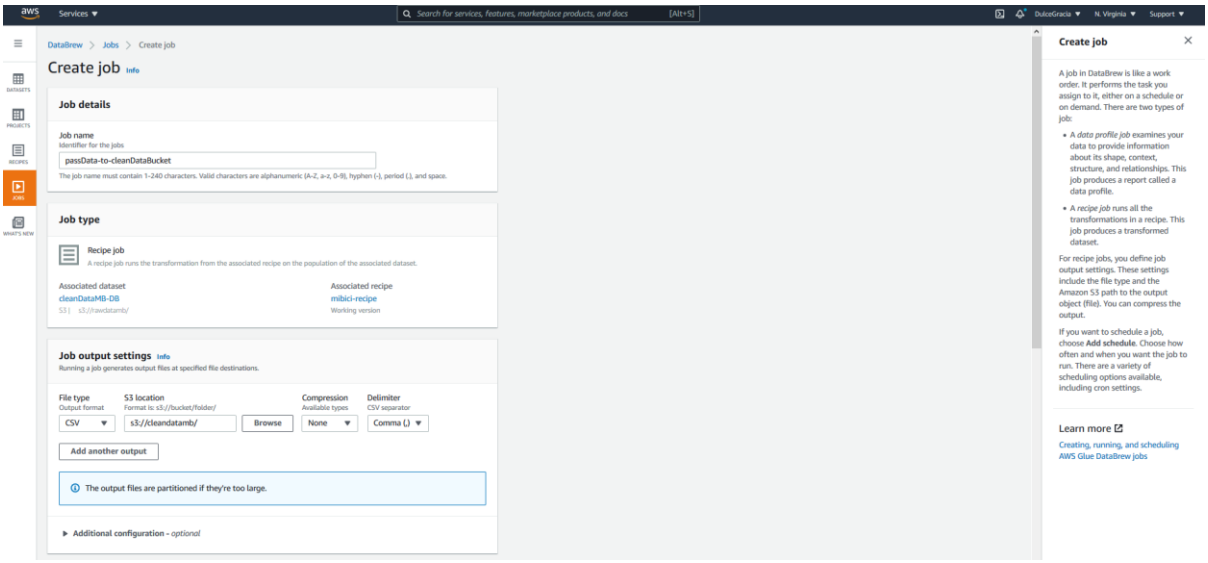


Figure 4.9 Job created for export the clean data

When the job is complete run and when complete check that the code the data is where is supposed to be. Figure 4.9 is an example on how this looks like when it is complete

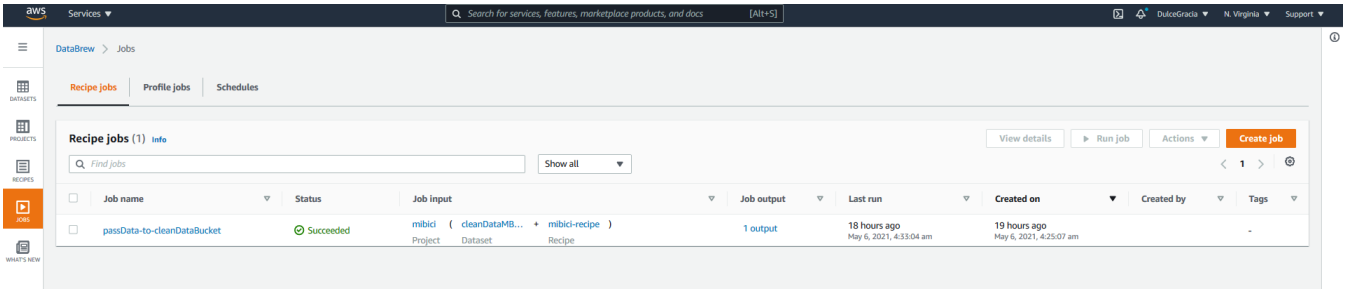


Figure 4.9 Recipe Jobs

4.6.2. Applied Classification Algorithm

For the second step that is the one related to Sage Maker service was needed to first create a new notebook for this implementation was used the one with origin and destiny file, for further details of the notebook please review git repository [9]. It is important to mention that when this is complete it is important to review that the new bucket contains the information with the results from this step.

Figure 4.11 is the file with the K-Means implementation at Sage Maker

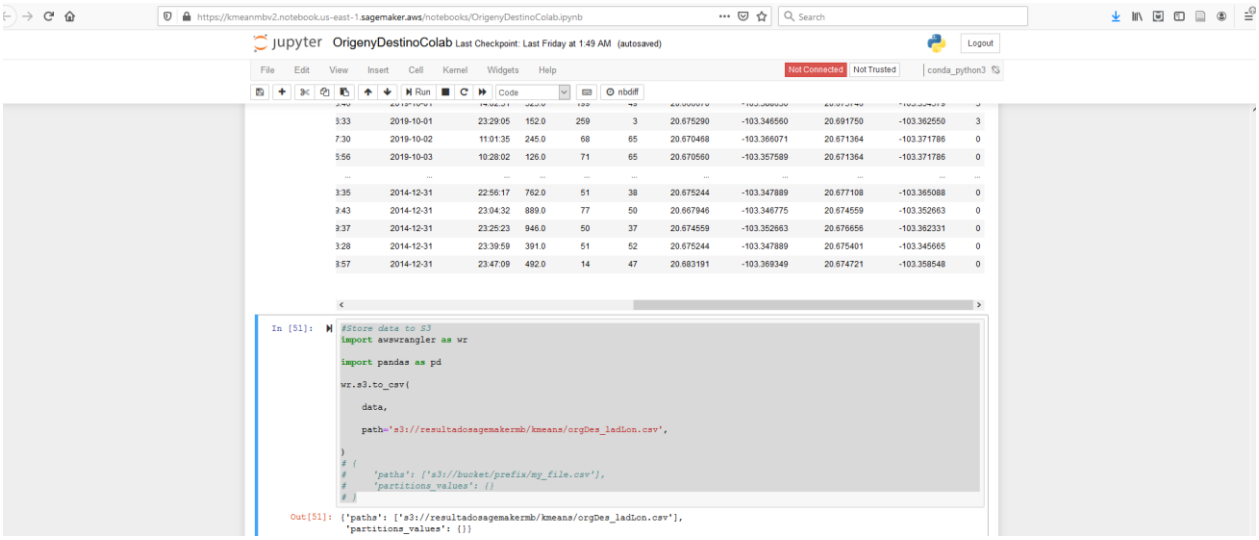


Figure 4.11 Notebook to be run at AWS Sage Maker

4.6.3. Analyze Results

Whit the third step first was need it to configure Quick Sight for analyze the results got it from the step before. Figure 4.12 is a representation of the dataset exported with the results

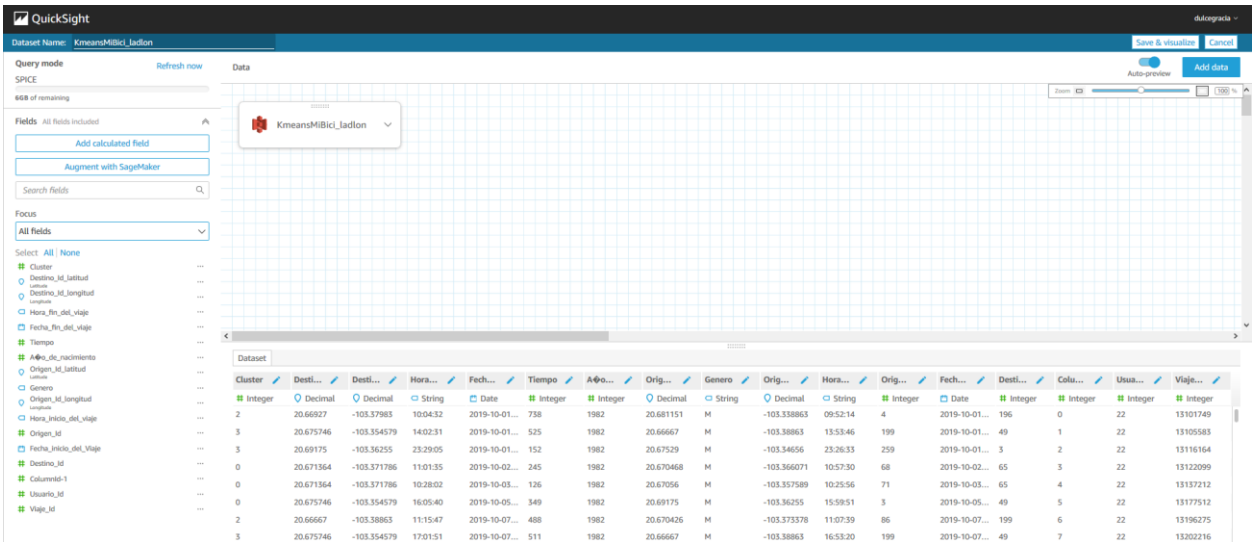


Figure 4.12 Results from the classification exported to Quick Sight

Figure 4.13 is the analysis of the results exported before

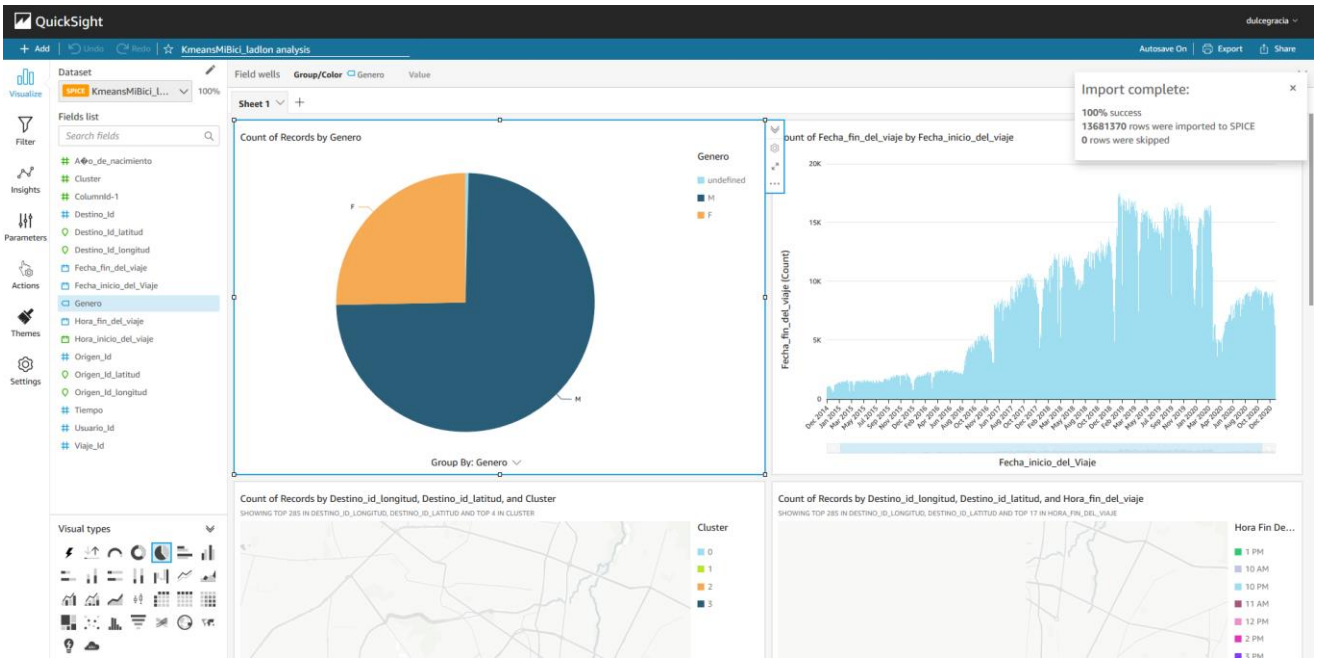


Figure 4.13 Results interpretation

The three stages of this project were all important for understanding the data. In the first approach, we were allowed to play with the data and test hypotheses about the classification using algorithms described in the previous steps. In the second step, taking as a premise the results of the previous step, which will be discussed later, it was interesting to see how the classification was disbursed differently by changing the algorithm and parameters.

The third step was important to better understand all the results obtained and test a cloud-focused implementation for future development.

5. RESULTS AND DISCUSION

In the following chapter, the results obtained during the investigation will be shown and discussed in order to test the hypotheses created from the beginning and will allow creating new ones based on the results obtained.

5.1. Result Stage 1

Different results were produced during stage one. Without a doubt, the most important were the results obtained from the implementation of the classification algorithms, which were decision trees and K-NN.

As a reminder for this part of the investigation was only consider the period of January 2019 – December 2019.

Correlation matrix was created in order to understand better which data set provided better correlation as figure 5.1 and figure 5.2 shown as continue

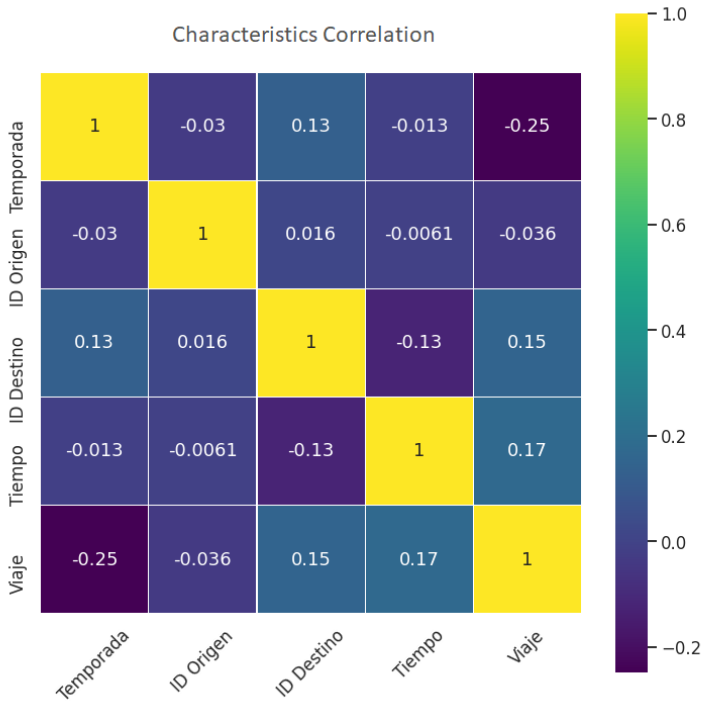


Figure 5.1 Characteristics Correlation 1st DataSet

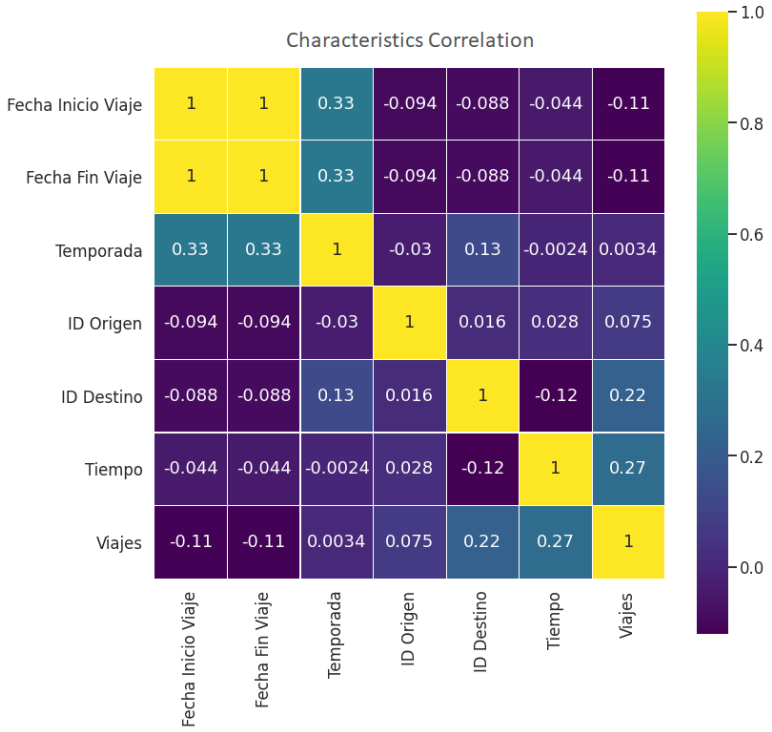


Figure 5.2 Characteristics Correlation 2nd DataSet

The correlation is not that good; however, the relationship from figure 5.1 to figure 5.2 improved, so it was decided to use the dataset of the second experiment.

5.1.1. Decision Tree

80% of the data set was used to train the model. To define the value of the maximum depth of the tree, the maximum number of attributes of the dataset was used. The figure 5.3 shows the average accuracy with its corresponding depth of the tree. Of the seven subgroups, the second is the one that have more accueracy with 80%. So that was the max depth used for the tree in order to also keep it simple.

Decision Tree Results

Max Depth	Average Accuracy
1	0.436275
2	0.807754
3	0.735740
4	0.729768
5	0.724866
6	0.703654
7	0.668360

Figure 5.3 Decision Tree Results

Once defined the depth of the tree, it was needed to train the model with the set deep. The result tree was the following figure 5.4

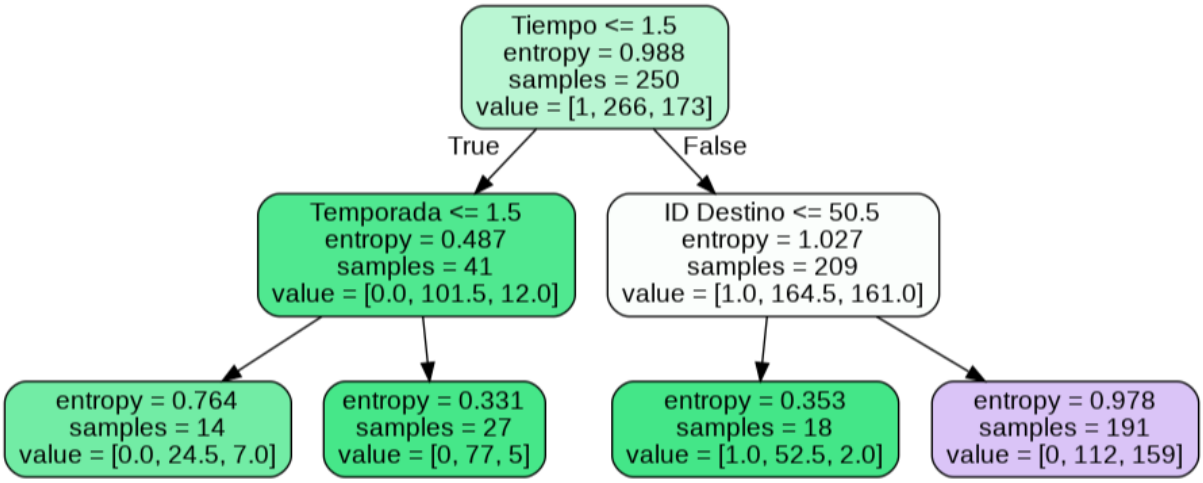


Figure 5.4 Decision Tree max deep 2

The precision of the decision tree was 81.20% in the training stage. However, since it was evaluated with an example of real data, its probability of success lowered to 58.67%. For more details, review the git repository [9].

5.1.2. K-NN

For the implementation of K-NN, took the independent variables as X: Temporada, ID Origen, ID Destino and Tiempo. For depend on variable as Y: Viajes.

In figure, 5.5 could be found the options of the values that K could take with the precision rate

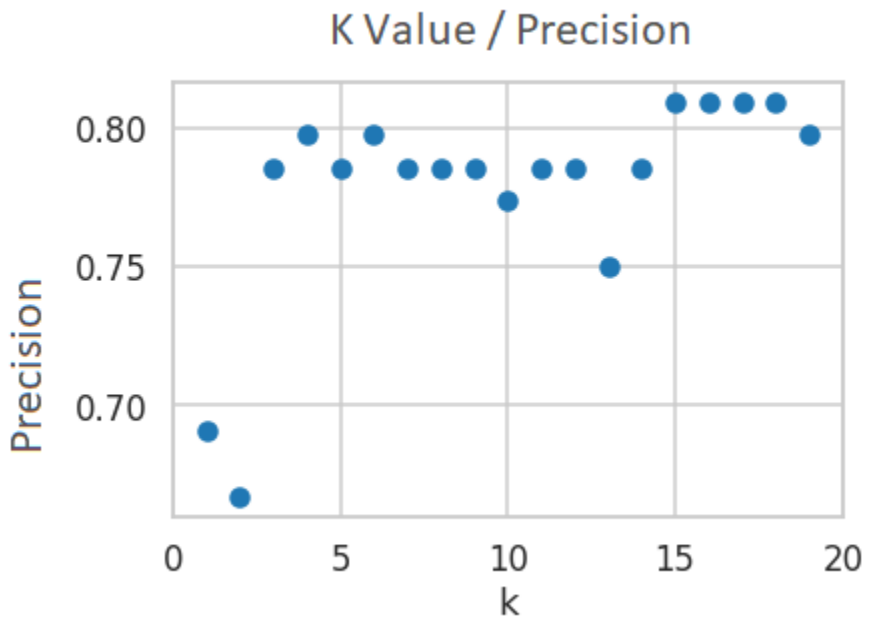


Figure 5.5 K value and precision rate accordingly

The best value for K would be in the range of 15 – 19. In order to keep the simplicity in the implementation the value of k would be four that also provide a good precision rate.

The results of the implementation using K value of four were:

K-NN Accuracy	
Training	0.82
Test	0.80

Table 5.1 K-NN Accuracy

With the purpose of exploring a bit more the implementation, proposed to run the algorithm for the month of January 2019 where an accuracy of 70% obtained.

5.2. Result Stage 2

At Stage 2 were created four different variations for the implementation of K-Means in order for try to understand better the data and to determinate which variation will be the most manful for the investigation.

5.2.1. Only Origin

For only origin have the purpose to analyze the data where x was equals to the field Origen. The elbow method was used to finding the best value of K.

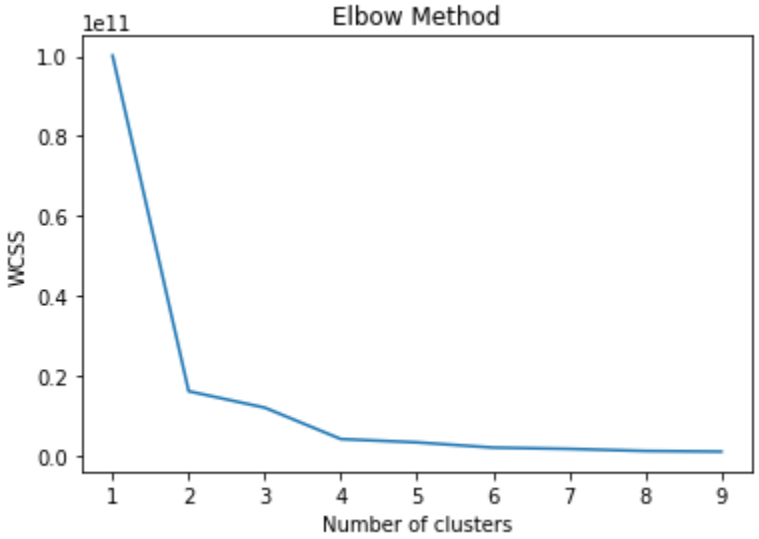


Figure 5.6 Elbow Method for Origin

As could be seen in figure 5.6 the best value for K would be 4, using this value K-Means algorithm applied and the result of the clustering shown in figure 5.7

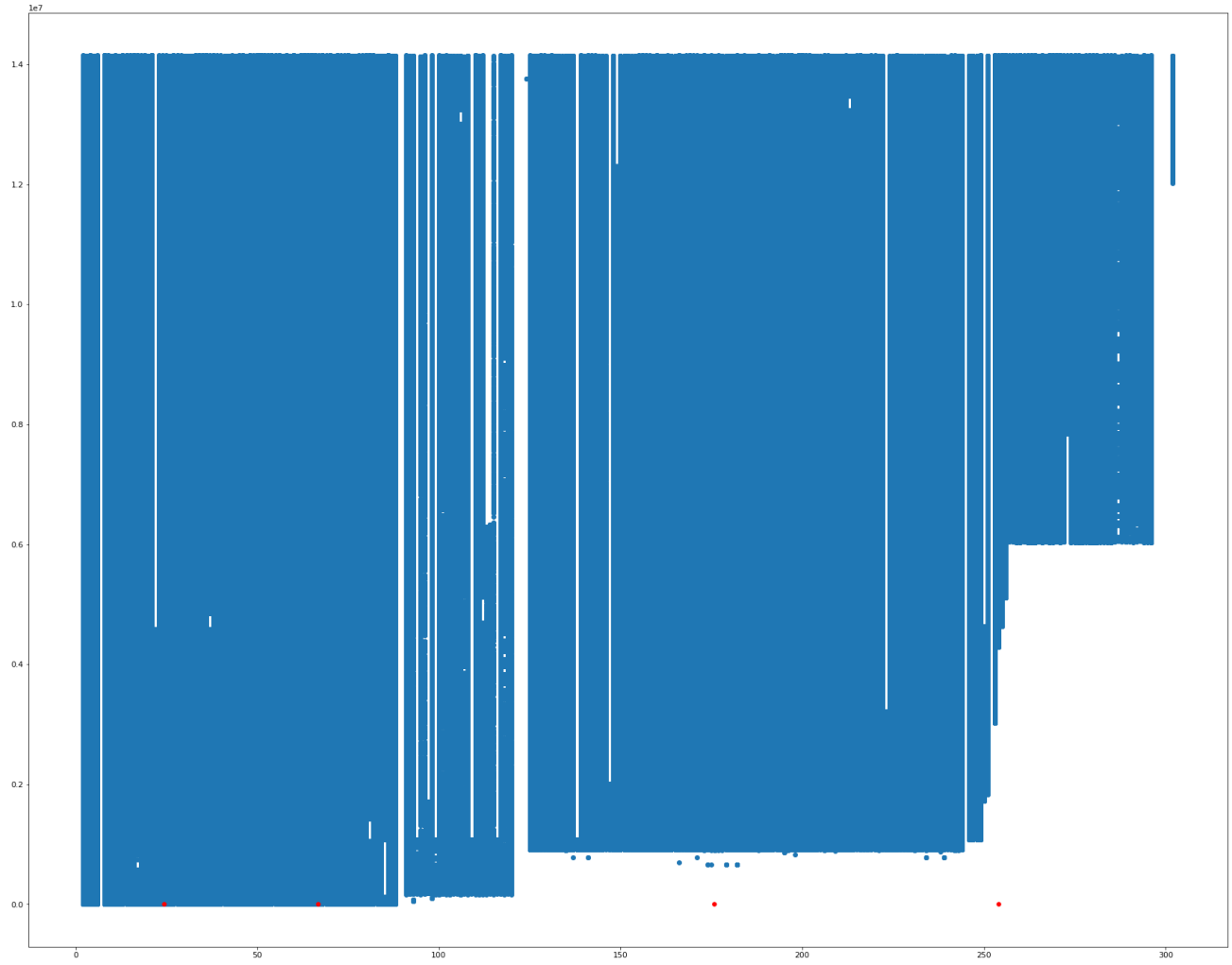


Figure 5.7 K-Means Origen Cluster

5.2.2. Only Destiny

For only origin have the purpose to analyze the data where x was equals to the field Destino. The elbow method was used to finding the best value of K.

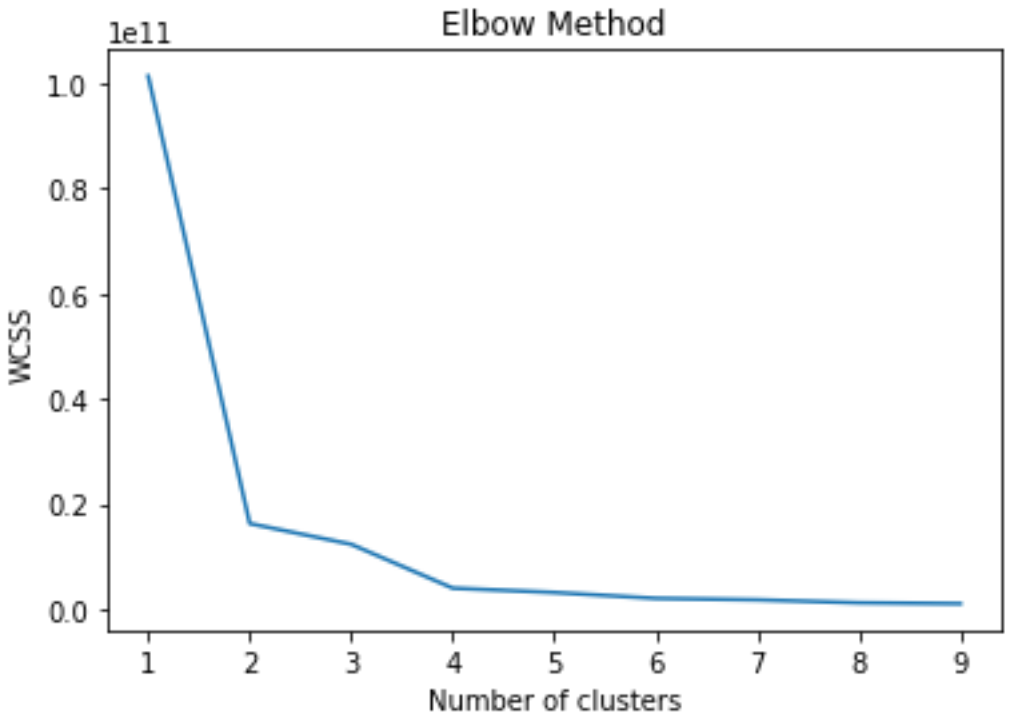


Figure 5.8 Elbow Method for Destiny

As could be seen in figure 5.8 the best value for K would be 4, using this value K-Means algorithm applied and the result of the clustering shown in figure 5.9

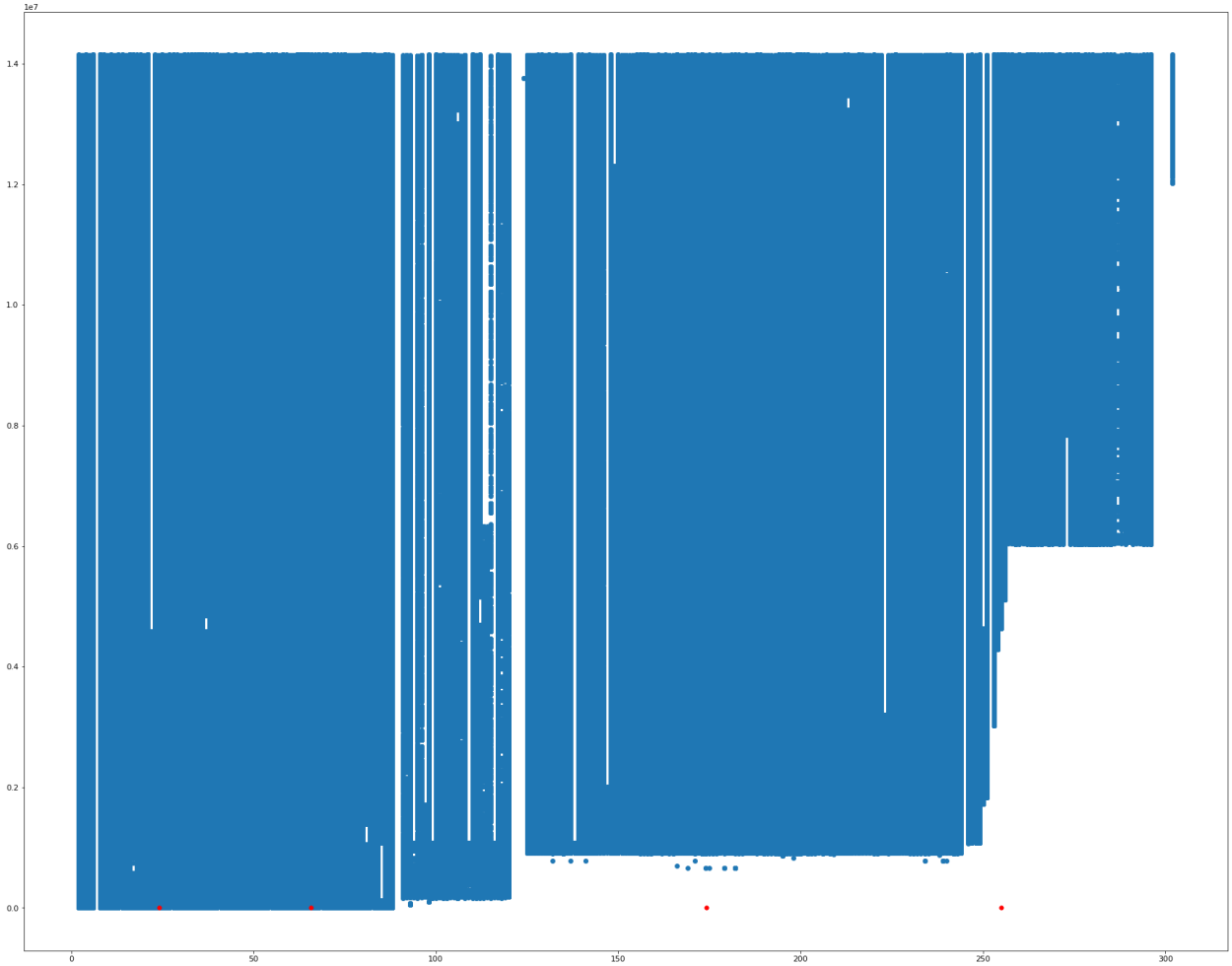


Figure 5.9 K-Means Destino Cluster

Also was created for k value two,

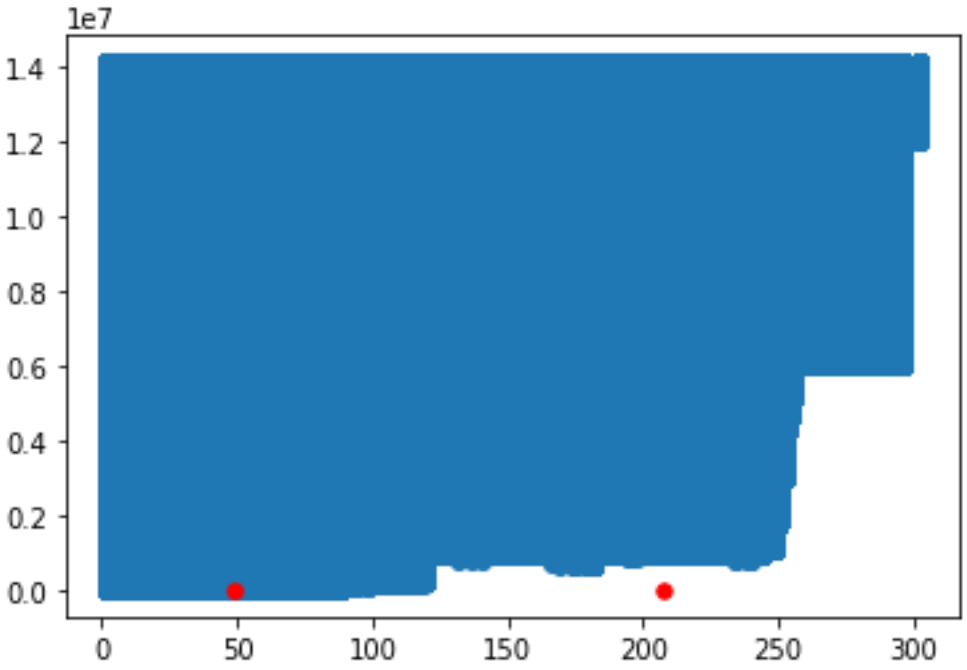


Figure 5.10 K-Means Destino 2 clusters

5.2.3. Origin and Destiny

For only origin have the purpose to analyze the data where x was equals to the fields Destino and Origen. The elbow method was used to finding the best value of K.

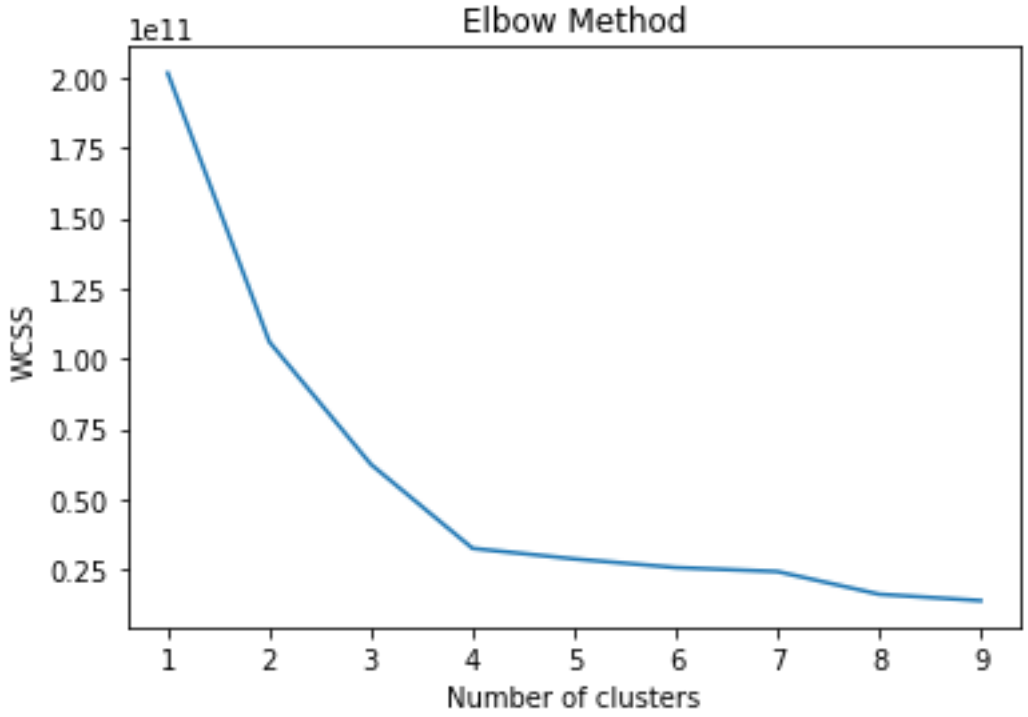


Figure 5.11 Elbow Method for Origen and Destino

As could be seen in figure 5.10 the best value for K would be 4, using this value K-Means algorithm applied and the result of the clustering shown in figure 5.12

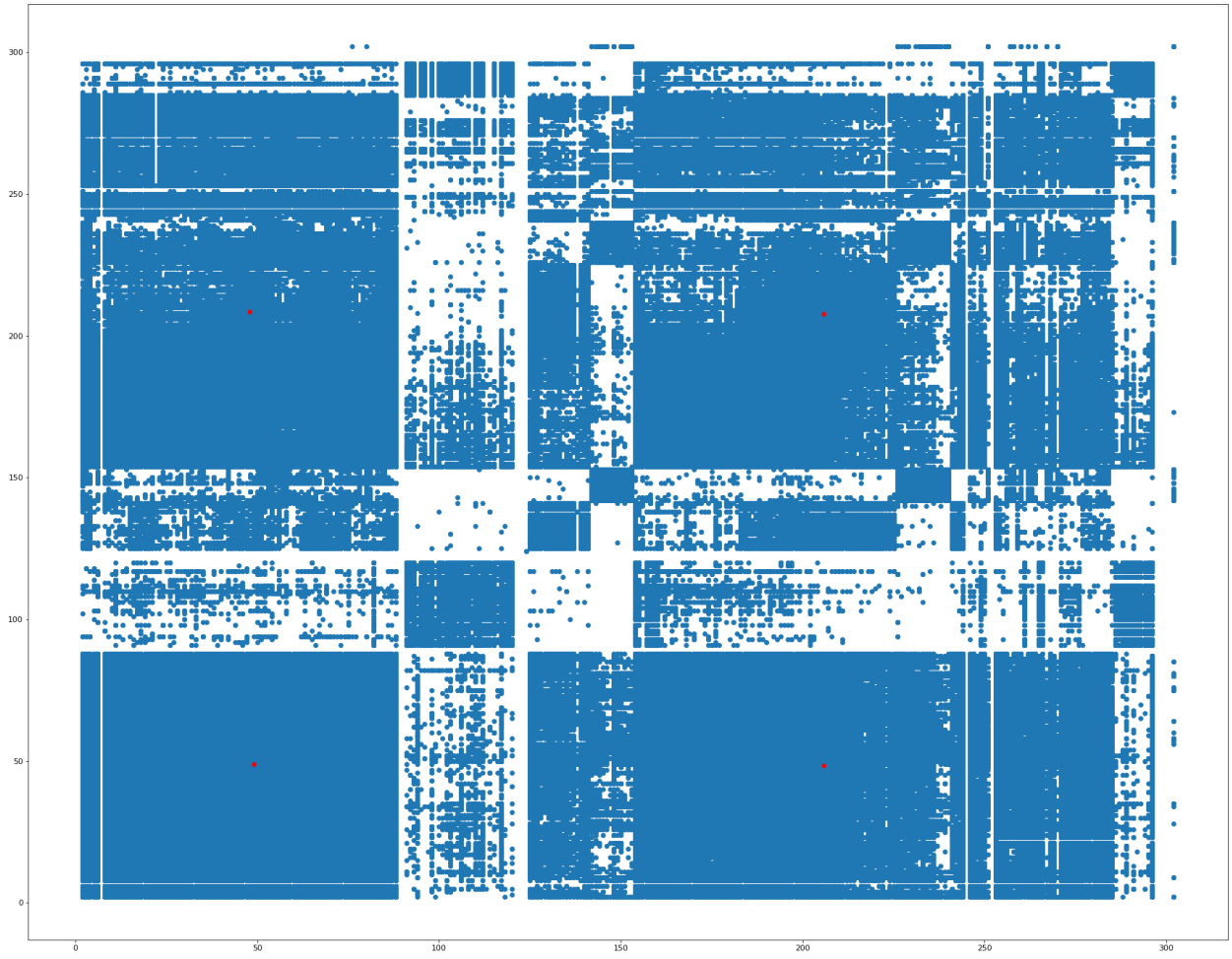


Figure 5.12 K-Means Origen and Destino

5.2.4. Origin and Destiny Difference

For only origin have the purpose to analyze the data where x was equals to the fields Destino and Origen but only took as valid row those whose destiny and origin is different. The elbow method was used to find the best value of K.

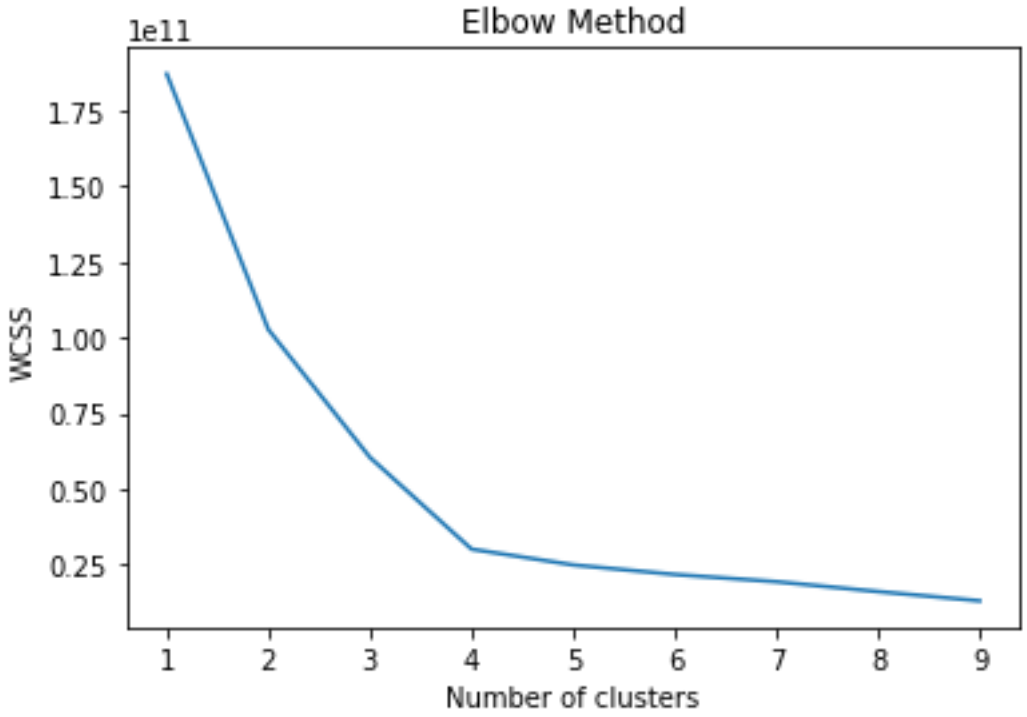


Figure 5.13 Elbow Method for Origen and Destino Diferentes

As could be seen in figure 5.13 the best value for K would be 4, using this value K-Means algorithm applied and the result of the clustering shown in figure 5.14

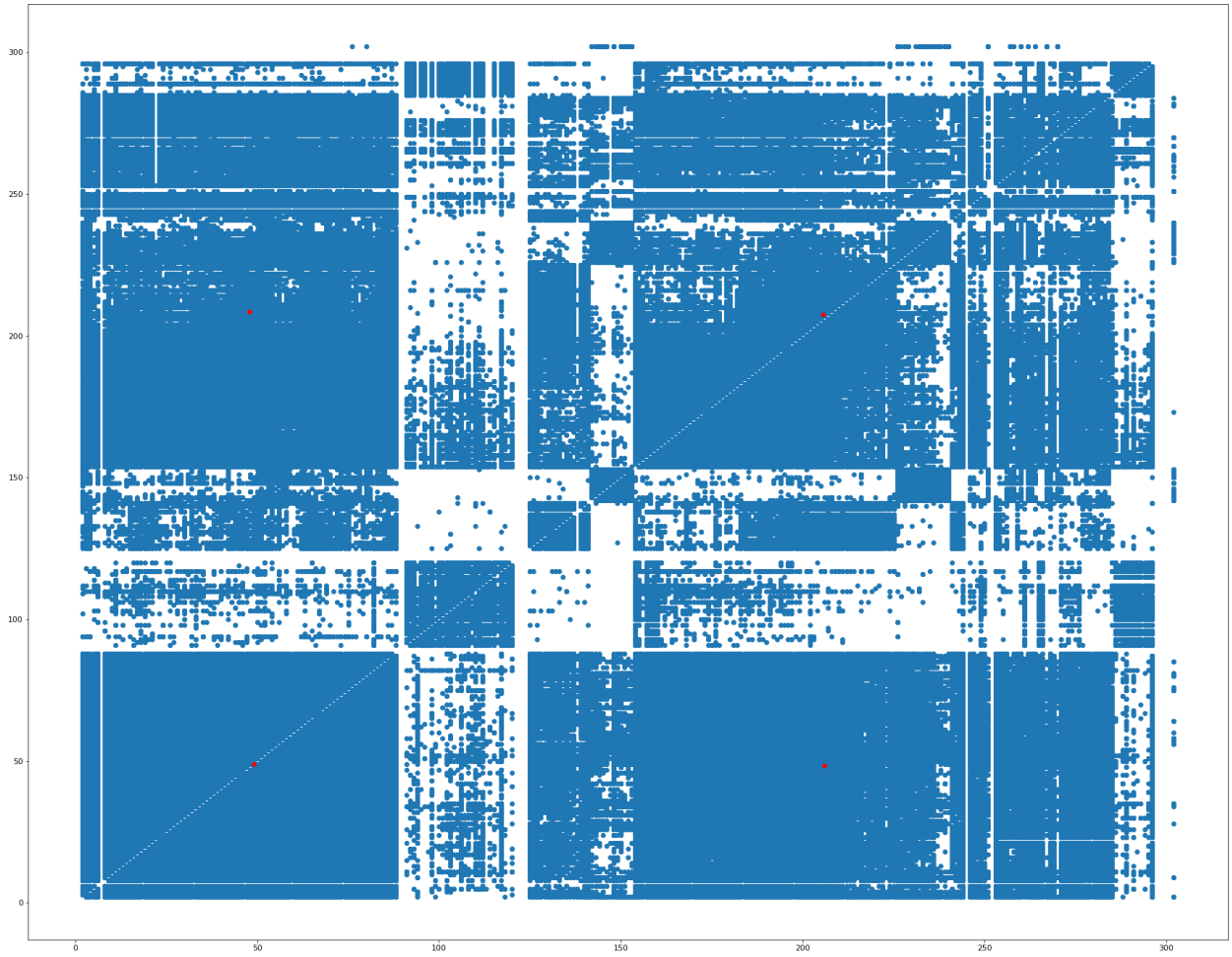


Figure 5.14 K-Means Origen and Destino Different

6. CONCLUSION

6.1. Conclusions

Social media data is already playing a key role in many fields, as sentiment analysis, urbans, emergency response, environmental change and so on … so, it is not weird to start using this treasure of information that contains a lot of attributes that with a good analysis can help us to understand better the world around us. Nowadays every person with access to a phone, internet, bicycle sharing data, social media user leaves rich content, such geographic coordinates, photographs, videos, resulting into significant sources of big data that normally does not publish with that goal but generally carry a lot.

During the investigation there were some challenges that are worth to be mentioned, one of them is to make sure that the data source that you are going to use is actually data that makes sense with the area that you want to study, and it would be extremely helpful if all the data have a format. A recommendation for MiBici will be to keep open data with a format so it is not difficult the analysis of the data for the users who desired to work with it.

As general conclusion, the investigation achieved the objective of creating new open-source oriented method for analysis urban dynamics and also complemented with a cloud-oriented solution, that hopefully could help as foundation for further investigations, because something that was found reading the literature was that most of the technologies used were paid. In addition, the results obtained from the proposed process demonstrated that it is possible to use bicycle-sharing data for map urban dynamic and finding land-characteristics.

6.2. Future Work

As work in the future, it is desired to add the use of social networks to be able to complement the research with the text that users get to share and see if there exists any relationship with the place where they are using social media. Also, we would like to use another source data that contains images and see if adding another data source (like any image data provider) helps to improve the behavior of the algorithm or just hinder the results. All this with the purpose of understanding better the human behavior that uses the applications and seeing if there could exist a better way to suit the hot spots/ bike stations or even help map better the road of the city.

APENDIX A

The following section will address most of the concepts related to the state of art with the purpose of deepening more in the concepts related to better understanding.

Satellite

A moon, planet, or machine that orbits a planet or star is known as a satellite. Because it orbits the sun, the Earth, for example, is a satellite. Because it orbits Earth, the moon is also a satellite. The term "satellite" usually refers to a mechanism that is launched into space and orbits Earth or another celestial body [56].

Thousands of artificial, or man-made, satellites orbit Earth. Some take pictures of the planet that helps meteorologists predict weather and track hurricanes. Some take pictures of other planets, the sun, black holes, dark matter, or faraway galaxies. These pictures help scientists better understand the solar system and universe. In the context of this research document, the referred satellites are used for GPS communication and geo-localization [56].

Remote sensing

The acquisition of data from a distance is known as remote sensing. NASA uses remote sensors, for example, on satellites and aircraft to view Earth and other planetary bodies, detecting and recording reflected or emitted radiation. Remote sensors, which provide a global perspective and a large amount of data about Earth systems, allow for data-driven decision making based on our planet's current and future state [57].

Sensors on satellites and aircraft measure the energy reflected by the Sun as a source of illumination or providing their own source of illumination. Sensors that rely on natural energy from the Sun are known as passive sensors, whereas those that generate their own energy are known as active sensors.

Radiometers (instruments that quantitatively measure the strength of electromagnetic radiation in specific bands) and spectrometers are examples of passive sensors (devices that are designed to detect, measure, and analyze the spectral content of reflected electromagnetic radiation). The visible, infrared, thermal infrared, and microwave regions of the electromagnetic spectrum are used by most passive systems used in remote sensing applications. These sensors [57] also measure Land and sea surface temperatures, vegetation qualities, cloud and aerosol properties, and other physical parameters.

Radio detection and range (radar) sensors, altimeters, and scatterometers are examples of active sensors. Most active sensors work in the microwave region of the electromagnetic spectrum, allowing them to penetrate the environment under almost all circumstances. Aerosol vertical profiles, forest structure, precipitation and winds, sea surface topography, and ice, among other things, can all be measured using these sensors [57].

SAR

For more than 30 years, synthetic aperture radar (SAR) has been widely employed for Earth remote sensing. It provides high-resolution, day-and-night, and weather-independent images for a multitude of applications ranging from geoscience and climate change research, environmental and Earth system

monitoring, 2-D, and 3-D mapping, change detection, 4-D mapping (space and time), security-related applications up to planetary exploration. In the 1990s, breakthroughs in radar technology and geo/bio-physical parameter inversion modeling, employing data from a variety of airborne and spaceborne systems, resulted in a paradigm change from technology-driven development to user-driven demand. Today, more than 15 spaceborne SAR systems are in use for a variety of purposes [58].

SAR's imaging capacity is unique in that it produces high-resolution two-dimensional images regardless of daylight, cloud cover, or weather conditions. It's built to keep track of dynamic processes on the Earth's surface in a reliable, continuous, and global manner. SAR systems use a pulsed radar placed on a forward-moving platform and have a side-looking imaging geometry [58].

Multi-Temporal InSAR

Because of the great spatial resolution achieved and the ability to acquire data remotely, spaceborne interferometric synthetic aperture radar (InSAR) is a powerful tool for assessing surface deformation. However, limitations with InSAR's applicability due to variations in scattering qualities of the Earth's surface with time and gaze direction limit its use [59].

Multi-temporal InSAR (MT-InSAR) approaches, which involve the processing of numerous acquisitions in real time, are one solution to these problems. MT-InSAR techniques are now divided into two categories: persistent scatterer (PS) methods, which identify pixels largely based on their phase variation in time, and small baseline (SB) methods, which identify pixels primarily based on their phase correlation in space. The categories' names are incongruent since 'persistent scatterer' relates to the sort of pixel that is detected, whereas'small baseline' refers to the interferogram production approach [59].

Google Earth

Google Earth is a geo browser that retrieves satellite and aerial photography, topography, ocean bathymetry, and other geographic data to create a three-dimensional globe of the Earth. Virtual globes or Earth browsers are other names for geo browsers. Google sometimes refers to Google Earth as a “geographic browser”. Other examples of Geo browsers are NASA's World Wind, ESRI's Explorer for ArcGIS, and GeoFusions's Geo Player [60].

ArcGIS Software

ArcGIS is a geospatial application that allows you to view, edit, manage, and analyze geographic data. Esri creates ArcGIS for desktop, mobile, and online mapping. "Science of Where" is their motto. As a result, ArcGIS is concentrating on location information and analytics. ArcGIS software is created by Esri (Environmental Systems Research Institute). The corporation was created in 1969 with the primary goal of land development [61].

Density-based spatial clustering of applications with noise (DBSCAN)

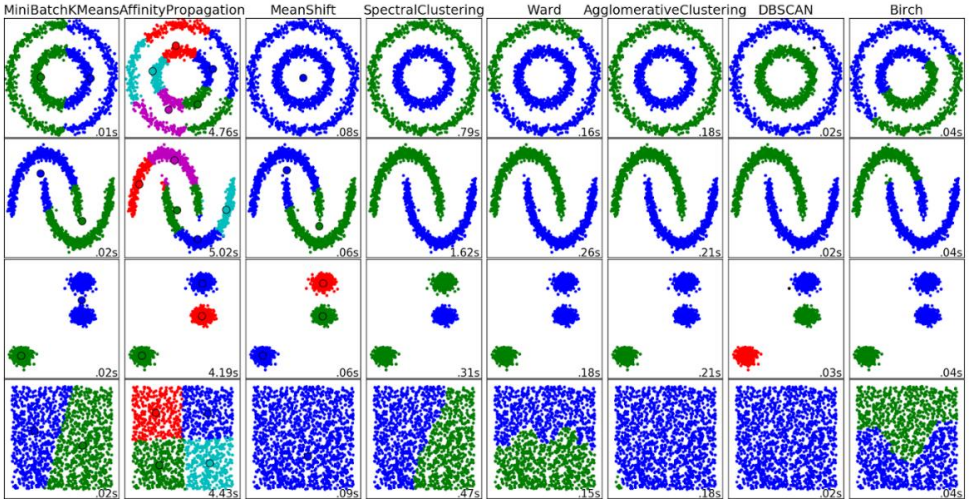
DBSCAN is a well-known data clustering algorithm that is commonly used in data mining and machine learning.

DBSCAN combines points that are close to each other based on a distance measurement (typically Euclidean distance) and a minimal number of points based on a set of points (let's consider in a

bidimensional space as demonstrated in the picture). It also identifies sites in low-density areas as outliers [62].

The DBSCAN algorithm should be used to reveal correlations and structures in data that are difficult to find manually but are meaningful and valuable in predicting patterns and trends.

In biology, health, social sciences, archaeology, marketing, character recognition, management systems, and other fields, clustering algorithms are commonly utilized.



A.1 Plot Cluster Comparation

Convolution Neuronal Network (CNN)

Deep learning techniques are based on neural networks, which are a subset of machine learning. They're made up of node levels, each of which has an input layer, one or more hidden layers, and an output layer. Each node is connected to the others and has a weight and threshold assigned to it. If a node's output exceeds a certain threshold value, the node is activated, and data is sent to the next tier of the network. Otherwise, no data is sent to the network's next tier [63].

The higher performance of convolutional neural networks with picture, speech, or audio signal inputs sets them apart from conventional neural networks. They are divided into three types of layers:

- Convolutional layer
- Pooling layer
- Fully connected (FC) layer

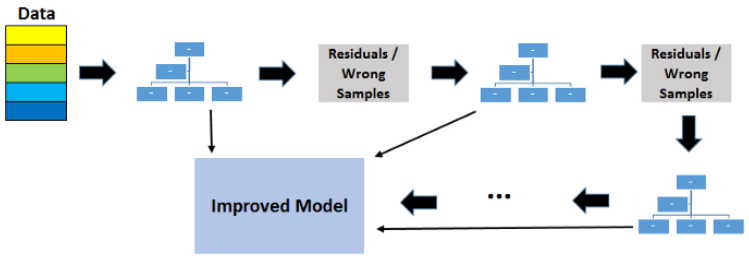
A convolutional network's first layer is the convolutional layer. While further convolutional layers or pooling layers can be added after convolutional layers, the fully connected layer is the last layer. CNN becomes more complicated with each layer, detecting larger areas of the image. Earlier layers concentrate on basic elements like colors and borders. As the visual data goes through the CNN layers, it begins to distinguish larger elements or features of the item, eventually identifying the target object [63].

Boos Trees

Boosting is the process of successively combining learning algorithms to produce a strong learner from a group of weak learners. The weak learners in the gradient boosted decision trees algorithm are decision trees [64].

Each tree tries to decrease the mistakes of the one before it. Boosting trees are slow learners, but by stacking them in a row and focusing on the flaws of the prior one, boosting becomes a very efficient and accurate model. Unlike bagging, boosting does not involve bootstrap sampling. Every time a new tree is added, it fits on a modified version of initial dataset [64].

Boosting algorithms take a long time to learn since trees are added in a sequential order. Models that learn slowly do better in statistical learning



A.2 Gradient Boosted Decision Trees

One Class Support Vector Machine

The unsupervised one-class SVM algorithm trains a decision function for novelty detection by categorizing fresh data as similar or dissimilar to the training set [65].

In a high or infinite dimensional space, a support vector machine creates a hyper-plane or group of hyper-planes that may be used for classification, regression, or other tasks. Intuitively, the hyper-plane with the greatest distance to the nearest training data points of any class (so-called functional margin) achieves a decent separation, because the higher the margin, the lower the classifier's generalization error [66].

Matlab

MATLAB is a desktop environment that is optimized for iterative analysis and design processes, as well as a programming language that directly represents matrix and array mathematics. It comes with a Live Editor, which allows you to create scripts that integrate code, output, and formatted text in a single executable notebook [67].

Deep Neural Network (DNN)

A deep neural network (DNN), or deep net for short, is a neural network with a certain amount of complexity, usually at least two layers. Deep nets use advanced math modeling to process data in complex ways. Deep neural networks have a lot of potential for statisticians, especially when it comes to improving the accuracy of a machine learning model [68].

Big Data

Big data refers to large, diversified amounts of data that continue to grow at an exponential rate. The volume of data, the velocity or pace at which it is created and collected, and the variety or scope of data points covered (known as the "three v's" of big data) are all factors to consider. Big data is frequently generated by data mining algorithms and is available in various of formats [69].

BIBLIOGRAPHY

[1] A. Leichter, D. Wittich, F. Rottensteiner, M. Werner and M. Sester, "IMPROVED CLASSIFICATION OF SATELLITE IMAGERY USING SPATIAL FEATURE MAPS EXTRACTED FROM SOCIAL MEDIA", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. -4, pp. 335-342, 2018. Available: 10.5194/isprs-archives-xlii-4-335-2018 [Accessed 25 October 2021].

[2] M. Jendryke, T. Balz, S. McClure and M. Liao, "Putting people in the picture: Combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in Shanghai", Computers, Environment and Urban Systems, vol. 62, pp. 99-112, 2017. Available: 10.1016/j.compenvurbsys.2016.10.004 [Accessed 25 October 2021].

[3] M. Li, Z. Shen, and X. Hao, "Revealing the relationship between spatio-temporal distribution of population and urban function with social media data", GeoJournal, vol. 81, no. 6, pp. 919-935, 2016. Available: 10.1007/s10708-016-9738-7 [Accessed 25 October 2021].

[4] J. Zhao, W. Fan and X. Zhai, "Identification of land-use characteristics using bicycle sharing data: A deep learning approach", Journal of Transport Geography, vol. 82, p. 102562, 2020. Available: 10.1016/j.jtrangeo.2019.102562 [Accessed 25 October 2021].

[5] Z. Miao, L. Wu, W. Shi, P. Gamba and M. Jiang, "Towards an Automatic Urban Settlement Mapping from Multi-Temporal InSAR Trained by Social Media", 2018 Progress in Electromagnetics Research Symposium (PIERS-Toyama), 2018. Available: 10.23919/piers.2018.8597712 [Accessed 25 October 2021].

[6] Z. Miao, G. Iannelli and P. Gamba, "Using social media data to map urban areas: ideas and limits", IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, 2019. Available: 10.1109/igarss.2019.8898361 [Accessed 25 October 2021].

[7] M. Qiao, Y. Wang, S. Wu, A. Luo, S. Ruan and Y. Gu, "Fine-Grained Subjective Partitioning of Urban Space Using Human Interactions from Social Media Data", IEEE Access, vol. 7, pp. 52085-52094, 2019. Available: 10.1109/access.2019.2911664 [Accessed 25 October 2021].

[9] "DulceGracia/MiBici", GitHub, 2021. [Online]. Available: https://github.com/DulceGracia/MiBici. [Accessed: 24- Oct- 2021].

[10] "MiBici |Datos abiertos", MiBici, 2021. [Online]. Available: https://www.mibici.net/es/datos-abiertos/. [Accessed: 24- Oct- 2021].

[11] "MIBICI | Frequently asked questions", MiBici, 2021. [Online]. Available: https://www.mibici.net/en/faq/. [Accessed: 12- May- 2021].

[12] "Guadalajara Travel Information for North American Travelers | Guadalajara, Mexico", Visitguadalajara.com, 2021. [Online]. Available: https://visitguadalajara.com/. [Accessed: 21- Oct- 2021].

[13] "What does "georeferenced" mean?", Usgs.gov, 2021. [Online]. Available: https://www.usgs.gov/faqs/what-does-georeferenced-mean?qt-news_science_products=0#qt-news_science_products. [Accessed: 23- Oct- 2021].

[14] "trip", TheFreeDictionary.com, 2021. [Online]. Available: https://www.thefreedictionary.com/trip. [Accessed: 24- Oct- 2021].

[15] h. behavior et al., "Human behaviour - Latest research and news | Nature", Nature.com, 2021. [Online]. Available: https://www.nature.com/subjects/human-behaviour. [Accessed: 23- Oct- 2021].

[16] "Definition of LBS (Location-based Services) - Gartner Information Technology Glossary", Gartner, 2021. [Online]. Available: https://www.gartner.com/en/information-technology/glossary/lbs-location-based-services. [Accessed: 29- Oct- 2021].

[17] "Point of Interest (POI) - Are these points really necessary for Mapping?? | Ceinsys", Ceinsys.com, 2021. [Online]. Available: https://www.ceinsys.com/blog/point-of-interest-really-necessary-for-mapping/. [Accessed: 23- Oct- 2021].

[18] "What is Urban Dynamics?", Osborneclarke.com, 2021. [Online]. Available: https://www.osborneclarke.com/insights/what-is-urban-dynamics. [Accessed: 23- Oct- 2021].

[19] "Introduction to Population Demographics | Learn Science at Scitable", Nature.com, 2021. [Online]. Available: https://www.nature.com/scitable/knowledge/library/introduction-to-population-demographics-83032908/. [Accessed: 23- Oct- 2021].

[20] "Population Distribution Definition", Worldpopulationreview.com, 2021. [Online]. Available: https://worldpopulationreview.com/articles/population-distribution-definition. [Accessed: 23- Oct- 2021].

[21] "What is Spatial Temporal? Definition and Related FAQs | OmniSci", Omnisci.com, 2021. [Online]. Available: https://www.omnisci.com/technical-glossary/spatial-temporal. [Accessed: 23- Oct- 2021].

[22] I. Education, "What is Machine Learning?", Ibm.com, 2021. [Online]. Available: https://www.ibm.com/cloud/learn/machine-learning. [Accessed: 24- Oct- 2021].

[23] "What is machine learning?", MIT Technology Review, 2021. [Online]. Available: https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/. [Accessed: 24- Oct- 2021].

[24] Mr.Y.Madhusekhar, Dr.G.Nanda Kishore kumar, Dr.K.Umapavankumar, Mr.R.Mantru naik, “Frequently used classification algorithms in Machine Learning with comparative analysis of various parameters”, IJAST, vol. 29, no. 7, pp. 11528-11529, Jun. 2020.

[25] "Machine Learning Basics with the K-Nearest Neighbors Algorithm", Medium, 2021. [Online]. Available: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761. [Accessed: 24- Oct- 2021].

[26] "KNN Algorithm | What is KNN Algorithm | How does KNN Function", Analytics Vidhya, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/. [Accessed: 24- Oct- 2021].

[27] M. Ibrahim, "The art of Data Analysis", Journal of Allied Health Sciences Pakistan, vol. 1, no. 1, pp. 98-104, 2015. [Accessed 21 October 2021].

[28] "Zone Mapping with Landsat Imagery", Usgs.gov, 2021. [Online]. Available: https://www.usgs.gov/centers/fort/science/zone-mapping-landsat-imagery?qt-science_center_objects=0#qt-science_center_objects. [Accessed: 22- Oct- 2021].

[29] "Statistics for Machine Learning", O'Reilly Online Learning, 2021. [Online]. Available: https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/c71ea970-0f3c-4973-8d3a-b09a7a6553c1.xhtml. [Accessed: 23- Oct- 2021].

[30] "Decision Tree Introduction with example - GeeksforGeeks", GeeksforGeeks, 2021. [Online]. Available: https://www.geeksforgeeks.org/decision-tree-introduction-example/. [Accessed: 26- Oct- 2021].

[31] "1.10. Decision Trees", scikit-learn, 2021. [Online]. Available: https://scikit-learn.org/stable/modules/tree.html. [Accessed: 26- Oct- 2021].

[32] "How K-Means Clustering Works", Docs.aws.amazon.com, 2021. [Online]. Available: https://docs.aws.amazon.com/es_es/sagemaker/latest/dg/algo-kmeans-tech-notes.html. [Accessed: 12- May- 2021].

[33] "Data cleaning: The benefits and steps to creating and using clean data", Tableau, 2021. [Online]. Available: https://www.tableau.com/learn/articles/what-is-data-cleaning. [Accessed: 12- May- 2021].

[34] "Aprendizaje no supervisado — Aprendizaje automático — DATA SCIENCE", DATA SCIENCE, 2021. [Online]. Available: https://datascience.eu/es/aprendizaje-automatico/aprendizaje-automatico-no-supervisado/. [Accessed: 12- May- 2021].

[35] "What is Git?", Educative: Interactive Courses for Software Developers, 2021. [Online]. Available: https://www.educative.io/edpresso/what-is-git. [Accessed: 21- Oct- 2021].

[36] "What is Python? Executive Summary", Python.org, 2021. [Online]. Available: https://www.python.org/doc/essays/blurb/. [Accessed: 12- May- 2021].

[37] "What is NumPy? — NumPy v1.21 Manual", Numpy.org, 2021. [Online]. Available: https://numpy.org/doc/stable/user/whatisnumpy.html. [Accessed: 21- Oct- 2021].

[38] M. Heller, "What is Keras? The deep neural network API explained", InfoWorld, 2021. [Online]. Available: https://www.infoworld.com/article/3336192/what-is-keras-the-deep-neural-network-api-explained.html. [Accessed: 21- Oct- 2021].

[39] S. Yegulalp, "What is TensorFlow? The machine learning library explained", InfoWorld, 2021. [Online]. Available: https://www.infoworld.com/article/3278008/what-is-tensorflow-the-machine-learning-library-explained.html. [Accessed: 21- Oct- 2021].

[40] "What Is Pandas in Python? Everything You Need to Know", ActiveState, 2021. [Online]. Available: https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/. [Accessed: 21- Oct- 2021].

[41] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python", Jmlr.org, 2021. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html. [Accessed: 21- Oct- 2021].

[42] "Jupyter Notebook Viewer", Nbviewer.jupyter.org, 2021. [Online]. Available: https://nbviewer.jupyter.org/github/jupyter/notebook/blob/master/docs/source/examples/Notebook/What%20is%20the%20Jupyter%20Notebook.ipynb#. [Accessed: 12- May- 2021].

[43] "What is AWS?", Amazon Web Services, Inc., 2021. [Online]. Available: https://aws.amazon.com/what-is-aws/. [Accessed: 12- May- 2021].

[44] "Getting started with Amazon S3 - Amazon Simple Storage Service", Docs.aws.amazon.com, 2021. [Online]. Available: https://docs.aws.amazon.com/AmazonS3/latest/userguide/GetStartedWithS3.html. [Accessed: 23- Oct- 2021].

[45] "Amazon Quick Sight – Business Intelligence Service - Amazon Web Services", Amazon Web Services, Inc., 2021. [Online]. Available: https://aws.amazon.com/quicksight/. [Accessed: 19- Apr- 2021].

[46] "AWS | Amazon S3", Amazon Web Services, Inc., 2021. [Online]. Available: https://aws.amazon.com/s3/. [Accessed: 02- May- 2021].

[47] "What is AWS Glue DataBrew? - AWS Glue DataBrew", Docs.aws.amazon.com, 2021. [Online]. Available: https://docs.aws.amazon.com /databrew/latest/dg/what-is.html. [Accessed: 12- May- 2021].

[48] "FAQs about Amazon SageMaker – Amazon Web Services (AWS)", Amazon Web Services, Inc., 2021. [Online]. Available: https://aws.amazon.com/sagemaker/faqs/. [Accessed: 12- May- 2021].

[49] Working with ML Insights. [Online]. Available: https://docs.aws.amazon.com/es_es/quicksight/latest/user/making-data-driven-decisions-with-ml-in-quicksight.html.[Accessed: 20-Apr-2021].

[50] "Importing Data into SPICE - Amazon Quick Sight", Docs.aws.amazon.com, 2021. [Online]. Available: https://docs.aws.amazon.com/quicksight/latest/user/spice.html. [Accessed: 12- May- 2021].

[51] "Social Media Definition", Investopedia, 2021. [Online]. Available: https://www.investopedia.com/terms/s/social-media.asp. [Accessed: 23- Oct- 2021].

[52] "What is Twitter? - Definition from WhatIs.com", WhatIs.com, 2021. [Online]. Available: https://whatis.techtarget.com/definition/Twitter. [Accessed: 23- Oct- 2021].

[53] "What is a Wizard? - Definition from Techopedia", Techopedia.com, 2021. [Online]. Available: https://www.techopedia.com/definition/32108/wizard-software. [Accessed: 24- Oct- 2021].

[54] "Jobs in SQL Server", C-sharpcorner.com, 2021. [Online]. Available: https://www.c-sharpcorner.com/UploadFile/ff0d0f/jobs-in-sql-server/. [Accessed: 23- Oct- 2021].

[55] "IBM Docs", Ibm.com, 2021. [Online]. Available: https://www.ibm.com/docs/en/zos-basic-skills?topic=more-what-is-data-set. [Accessed: 23- Oct- 2021].

[56] "What Is a Satellite?", NASA, 2021. [Online]. Available: https://www.nasa.gov/audience/forstudents/5-8/features/nasa-knows/what-is-a-satellite-58.html. [Accessed: 13- Nov- 2021].

[57] "What is Remote Sensing? | Earthdata", Earthdata.nasa.gov, 2021. [Online]. Available: https://earthdata.nasa.gov/learn/backgrounders/remote-sensing. [Accessed: 13- Nov- 2021].

[58] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek and K. Papathanassiou, "A tutorial on synthetic aperture radar", IEEE Geoscience and Remote Sensing Magazine, vol. 1, no. 1, pp. 6-43, 2013. Available: 10.1109/mgrs.2013.2248301. [Accessed: 13-Nov-2021]

[59] A. Hooper, "A multi-temporal InSAR method incorporating both persistent scatterer and small baseline approaches", Geophysical Research Letters, vol. 35, no. 16, 2008. Available: 10.1029/2008gl034654 [Accessed: 13 November 2021].

[60] "What is Google Earth?", Teaching with Google Earth, 2021. [Online]. Available: https://serc.carleton.edu/introgeo/google_earth/what.html. [Accessed: 13- Nov- 2021].

[61] "What is ArcGIS? - GIS Geography", GIS Geography, 2021. [Online]. Available: https://gisgeography.com/what-is-arcgis/. [Accessed: 13- Nov- 2021].

[62] "How DBSCAN works and why should we use it?", Medium, 2021. [Online]. Available: https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80. [Accessed: 23- Oct- 2021].

[63] I. Education, "What are Convolutional Neural Networks?", *Ibm.com*, 2021. [Online]. Available: https://www.ibm.com/cloud/learn/convolutional-neural-networks. [Accessed: 13- Nov- 2021].

[64] "Gradient Boosted Decision Trees-Explained", *Medium*, 2021. [Online]. Available: https://towardsdatascience.com/gradient-boosted-decision-trees-explained-9259bd8205af. [Accessed: 13- Nov- 2021].

[65] "One-class SVM with non-linear kernel (RBF)", scikit-learn, 2021. [Online]. Available: https://scikit-learn.org/stable/auto_examples/svm/plot_oneclass.html. [Accessed: 13- Nov- 2021].

[66] "1.4. Support Vector Machines", *scikit-learn*, 2021. [Online]. Available: https://scikit-learn.org/stable/modules/svm.html#svm-outlier-detection. [Accessed: 13- Nov- 2021].

[67] 2021. [Online]. Available: https://www.mathworks.com/products/matlab.html. [Accessed: 18- Nov- 2021].

[68] "What's a Deep Neural Network? Deep Nets Explained", BMC Blogs, 2021. [Online]. Available: https://www.bmc.com/blogs/deep-neural-network/. [Accessed: 13- Nov- 2021].

[69] "Big Data", Investopedia, 2021. [Online]. Available: https://www.investopedia.com/terms/b/big-data.asp. [Accessed: 13- Nov- 2021].