**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Programming for DA  Statistics for Data Analytics  Machine Learning for Data Analysis  Data Preparation & Visualisation |
| **Assessment Title:** | MSC\_DA\_Repeat |
| **Lecturer Name:** | Marina Iantorno/Taufique Ahmed  Sam Weiss,  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Batuhan CIMEN |
| **Student Number:** | 2023172 |
| **Assessment Due Date:** | 27.06.2024 |
| **Date of Submission:** | 07.08.2024 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

ABSTRACT

**SAN FRANCISCO & DUBLIN BIKE USAGE ANALYSIS AND VISUALIZATION**

**CCT COLLEGE DUBLIN**

**DATA ANALYTICS**

**LECTURER**

**David McQuaid**

**Muhammad Iqbal**

**DUBLIN, 2024**

This study aims to compare the cycling use and infrastructure of the cities of Dublin and San Francisco. Criteria such as monthly shared bicycle usage numbers of cities, bicycle preferences (Classic or E-Bike) and community satisfaction were examined. These data aim to guide other cities that want to develop sustainable transportation policies by revealing the similarities and differences between the two cities. As a result of the analysis, significant differences and similarities in the bicycle use of both cities were identified, and the effects of these differences and similarities on the transportation policies of the cities were evaluated.

**Keywords:** Transportation System, Biycle, San Francisco, Dublin, Database, Data Analysis, Sentiment Analysis, Data Visualization

CONTENTS

**Page**

[ABSTRACT ii](#_Toc173963911)

[**1.** **INTRODUCTION** 4](#_Toc173963912)

[**2.** **CONCEPTS** 5](#_Toc173963913)

[2.1 Bike & E-Bike 5](#_Toc173963914)

[2.2 Machine Learning 5](#_Toc173963915)

[2.3 Data Analysis 6](#_Toc173963916)

[2.4 CSV and JSON File 7](#_Toc173963917)

[**3.** **MATERIALS and METHODS** 7](#_Toc173963918)

[3.1 Selecting a Topic and Finding Data 7](#_Toc173963919)

[3.2 Starting to Write Code in Jupyter Notebook 8](#_Toc173963920)

[**4. RESULTS** 12](#_Toc173963921)

[REFERENCES 12](#_Toc173963922)

# **INTRODUCTION**

The bicycle paths and sharing projects is an important step to reduce traffic congestion in cities, protect the environment and promote a healthy lifestyle. Micromobility vehicles are a low-cost and environmentally friendly means of transportation and also provide many health benefits. The bicycle path and sharing project increases bicycle use and reduces automobile use by providing a safe road network for cyclists; By facilitating individuals access to bicycle use, it appeals to a wider user base and helps to popularize bicycle use. While this system provides a more sustainable transportation option for people living in cities, it also contributes to protecting the environment, reducing traffic congestion and helping people live a healthy life.

Modern cities are making cycling an increasingly important issue in the search for sustainable transport solutions. There are significant changes in urban mobility with the use of internet of things technology in different areas. City planners and local governments are investing in cycling infrastructure to reduce traffic congestion, promote environmentally friendly transportation options, and improve overall public health. In this context, examining cycling use and infrastructure in different cities can provide valuable information for developing effective policies and practices. This study aims to compare cycling data between Dublin, one of the dynamic capitals of Europe, and San Francisco, the technology center of the United States. Criteria such as monthly shared bicycle usage numbers, bicycle preferences (Classic or E-Bike) and community satisfaction in the two cities will be examined and similarities and differences in bicycle use will be revealed. These data can also be a guide for other cities that want to develop sustainable transportation policies.

The data used for Dublin in the project was taken from Dublin City Council (https://data.gov.ie/). Especially the moby bike dataset (https://data.gov.ie/dataset/moby-bikes) was chosen. It is thought that it will contribute more to the progress of the project. The months of May, June and July of different years were examined. For San Francisco, DataSF (https://data.sfgov.org/) was examined. The same periods as Dublin data were used to analyze the common points, differences and improvability of the two cities.

Pie chart, bar plot, line plot in the visualization sections; Bernoulli Distribution, Binomial Distribution, mode, median, variance in mathematics; In machine learning, methods such as Confusion Matrix, K-Fold Cross Validation, Principal Component Analysis (PCA) , K-means Clustering, Logistic Regression Classifier, K-Nearest Neighbors Classifier (KNN), Support Vector Machine Classifier (SVC), Naive Bayes Classifier, Decision Tree Classifier, Random Forest Classifier have been used.

# **CONCEPTS**

1. **Bike & E-Bike**

Bicycles are used as a part of transportation in more than 200 cities around the world. Especially Europe Serious steps have been taken in this regard. Cities with the highest bicycle usage in Europe The Netherlands, Denmark, Germany and Belgium.

In Japan, 15% of people commute to work by bicycle and 10 million bicycles are sold annually.

There are 5.2 million bicycles in Belgium, which has a population of 10.83 million. 84% of people living in the Netherlands own one or more bicycles (Ministry, 2019). There are 23 million bicycles in the Netherlands, which has a population of 17 million. The number of bicycles in the Netherlands is three times that of cars, accounting for 48% of vehicle traffic in urban areas. The Netherlands ranks first among European countries with an average daily cycling distance of 3 km per person. In addition to having a safe and well-run bicycle road network, the adoption of bicycles as a means of transportation by all segments of society stands out.

Among the cycling policies in England; There are applications such as adding bicycle parks to bus parks and driving areas, arranging unused railways and canals for bicycles, and developing strategic passages known as "green roads".

More than 100 thousand people use bicycles every day in New York, and this number is increasing day by day. In order to increase the use of bicycles in the city, May has been declared "Bicycle Month" and various events are organized throughout New York. It has been stated that in New York, where bicycles are used as a means of transportation as well as a sports tool, there are more bicycle commuters than other cities in the USA. More than 50 kilometers of bicycle paths have been created in the Manhattan area to encourage and popularize cycling (Ministry, 2019).

1. **Machine Learning**

A developing subfield of computing algorithms called "machine learning" aims to mimic human intelligence by taking in information from its surroundings. In the brand-new era of "big data," they are regarded as the workhorse. Machine learning techniques have been effectively used in a variety of sectors, including computational biology, finance, entertainment, pattern recognition, computer vision, aerospace engineering, and biomedical and medical applications (Murphy).

* **Decision Tree**

Decision tree is a graph to represent choices and their results in form of a tree. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions. Each tree consists of nodes and branches. Each node represents attributes in a group that is to be classified and each branch represents a value that the node can take.

* **Naive Bayes**

It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes mainly targets the text classification industry. It is mainly used for clustering and classification purpose depends on the conditional probability of happening

* **Support Vector Machine**

Another most widely used state-of-the-art machine learning technique is Support Vector Machine (SVM). In machine learning, support-vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. It's basically draw margins between the classes. The margins are drawn in such a fashion that the distance between the margin and the classes is maximum and hence, minimizing the classification error.

* **K-Means Clustering**

K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them is much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as bary center of the clusters resulting from the previous step.

* **K Nearest Neighbor**

The K-Nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows (Mahesh, 2019).

1. **Data Analysis**

Data analysis is the process of obtaining meaningful information and conclusions through the collection, cleaning, processing and examination of raw data. This process involves examining and interpreting data using a variety of statistical techniques, data mining methods, and visualization tools. Data analysis is used for various purposes such as evaluating the performance of businesses, developing marketing strategies, improving operational efficiency and conducting scientific research. The analysis process begins with data collection and continues with cleaning the data from errors and converting it into a suitable format. The data is then examined in detail and the results are presented in the form of graphs, tables and reports. These results provide strategic insights to decision makers and stakeholders, thus helping to make more informed and effective decisions. Data analysis plays a critical role in many different industries and fields today.

1. **CSV and JSON File**

CSV (Comma-Separated Values) and JSON (JavaScript Object Notation) are two widely used formats for data storage and exchange.

* **Comma-Separated Values (CSV)**

When storing tabular data in a spreadsheet or database, the CSV file format is a convenient option. All of the fields in a row that are represented by a comma in the CSV file are also represented by rows in the table.

* **JavaScript Object Notation (JSON)**

JSON is a lightweight data interchange format that is easy for humans to read and write, and for machines to parse and create. It is based on a subset of the JavaScript Programming Language.

# **MATERIALS and METHODS**

**3.1 Selecting a Topic and Finding Data**

First of all, there needs to be data to start the project. While doing my research on the subject given to us, I saw that many countries are actively trying to popularize the bicycle; but these countries did not share their data. Those who were open to sharing preferred to sell their data. After a long search, I saw that San Francisco data was publicly available and I decided to choose this city (https://datasf.org/opendata/).

The Dublin side was a little easier. I tried to find a data set suitable for the San Francisco data set I chose from “https://data.gov.ie”, the official database sharing site of the government. In order to make a comparison between the two cities, I needed to examine similar features. When I looked at the features of the Moby dataset, I saw that it was suitable and decided to choose this dataset.

First, I noticed that these two datasets have different dates. For example, San Francisco data started from January 2018, while Dublin Moby data started from September 2020. I tried to get equal spacing so the results would be comparable. I later realized that the Dublin data was not kept during the corona period, so I couldn’t make a comparison there either. Finally, I started working with May-June 2021, May-July 2022, and May-June 2024.

Dublin Dataset

A screenshot of a computer

Description automatically generated

San Francisco Dataset

A close-up of a number

Description automatically generated

**3.2 Starting to Write Code in Jupyter Notebook**

As for the data, there were missing data in both datasets. In the Dublin data set, there were 122 null values ​​in the “station\_id” column, and in the San Francisco data set, there were many null values ​​in the “start\_station\_name, start\_station\_id, end\_station\_name, end\_station\_id” columns.

A screen shot of a computer code

Description automatically generated A screen shot of a computer

Description automatically generated

Rows with NaN or NULL values were deleted with this code:

“dub\_BikeDF.dropna(inplace=True)”

If we wanted to ignore it, we would use “dub\_BikeDF[dub\_BikeDF["station\_id"].notna()]”. When we ran the code again and checked, we saw that the null values ​​were deleted.

Afterwards, we learned about data with functions such as “describe” and “unique”.

Pie chart was first used in the visualization part. Classic bike and e-bike rates were examined one by one for cities. While e-bikes are quite common in San Francisco, we saw that the rate is very low in Dublin.

San Francisco,

A pie chart with a number of percentages

Description automatically generated

For Dublin,

A blue circle with a number of percentages

Description automatically generated

Total results visualized with bar plot,

**A graph of different colored squares

Description automatically generated**

Only May and June of 2024 were used in the charts. The computer had difficulty operating it because the data set size was large. But of course, I was wondering about general use cases. That's why I focused on the numbers rather than processing all of the data.

A table with numbers and a number of months

Description automatically generated

When we look at the months of May and June of 2021, May and July of 2022, and finally May and June of 2024, the table is as follows.

A graph of a bar chart

Description automatically generated with medium confidence

A graph with a line drawn on it

Description automatically generated

# **4. RESULTS**

Various analyzes have been conducted based on statistics on shared bike use in Dublin and San Francisco. Some results were achieved in this direction. Bicycle use in San Francisco has become more common over the years. We can say that it is growing at the right rate. But there's a difference in Dublin. While Dublin was equivalent to the number of bicycle usage in San Francisco in 2021, it experienced a significant decrease in 2022. When we look at the year 2024, an extreme increase is evident. We can attribute this situation to the state's shared bicycle policies and the support it gives to bicycle use.

Another issue is e-bike and classic bike usage. In San Francisco, 76.8% bicycle use is seen. This rate is only 2.1% in Dublin. San Francisco's mild climate and generally bike-friendly weather conditions encourage cycling. On the other hand, Dublin has a wetter and windier climate, which may make e-bike use particularly difficult.

Looking at the results of the algorithms used in the project, it was revealed that the accuracy rate of Logistic Regression Classifier was 25%. Support Vector Machine is the same way. From here we can conclude that it should not be used. On the other hand, K-Nearest Neighbors Classifier and Random Forest Classifier had 50% accuracy. While Naive Bayes Classifier had 75% accuracy, Decision Tree Classifier reached 100% accuracy. It shows that the instances of this model is perfectly classified. We do not use algorithms with an accuracy of 50% or less because our risk of making mistakes is very high. Naive Bayes should not be used even though it has an accuracy rate of 75%. We know that in addition to rapid and efficient growth, it is more successful in problems such as text classification. The success rate in our analysis may be misleading.

REFERENCES

Ahmed Oussous, F. Z. (2018). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences, 30*(4), 431-448. doi:https://doi.org/10.1016/j.jksuci.2017.06.001

Andreas C. Müller, S. G. (2016). *Introduction to Machine Learning with Python.* O'Reilly Media, Inc.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd Edition ed.). O'Reilly Media, Inc.

Mahesh, B. (2019). *Machine Learning Algorithms.* doi:10.21275/ART20203995

Ministry, T. E. (2019, 12 23). Retrieved from T.C. Çevre ve Şehircilik Bakanlığı. (2017). Şehir içi Bisiklet Yolları Kılavuzu. https://webdosya.csb.gov.tr/db/meslekihizmetler/ustmenu/ustmenu1010.pdf

Murphy, I. E. (n.d.). *What Is Machine Learning?*

Rigatti, S. J. (2017, January). Random Forest. *47*(1), 31-39.