National College of Ireland

Prediction of Crowdedness at Campus Fitness Center

Domain Applications of Predictive Analysis - Project Design
Data Analytics

Hitesh Patil
Student ID: x19147996

School of Computing
National College of Ireland

Supervisor: Prof. Vikas Sahani

.

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Hitesh Patil
Student ID:	x19147996
Programme:	Data Analytics
Year:	2020
Module:	Domain Applications of Predictive Analysis - Project Design
Supervisor:	Prof. Vikas Sahani
Submission Due Date:	28/06/2020
Project Title:	Prediction of Crowdedness at Campus Fitness Center
Word Count:	2218
Page Count:	8

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	28th June 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Crowdedness at Campus Fitness Center

Hitesh Madhukar Patil
School of Computing
National College of Ireland
Dublin, Ireland
x19147996@student.ncirl.ie

***Abstract*—In this paper, the intention is to explain the different demographics present for the number of people attending the gym and a summary of what methods have been used to predict the crowdedness is presented. From a business' viewpoint, the crowd density prediction can be helpful for people going to the gym and fitness centres, shopping malls, theatres and restaurants. The overall study also consider various ethical concerns. The techniques used to predict crowdedness are Decision Tree, Random Forest and Support Vector Machine.**

***Index Terms*—crowdedness, visualization, machine learning, prediction.**

I. Introduction

A. Background

The world hasn't seen a pandemic like a Covid-19 in many decades where everyone's life is at risk. Many countries have imposed a lockdown, rules like Social Distancing, Quarantine and Isolation are implemented for the better safety of people [1]. It is important to note that after lockdown, people are likely to swarm to the theatres, marketplaces, ballparks and other places such as barbershop, gymnasium, restaurants, etc. where chances of getting infected are high [2]. Monitoring and handling of the crowd in such places become much more important [3].

Many fitness enthusiast, especially teenagers, go to the gymnasium to stay healthy and fit. These people need to know how they can avoid the infection by just changing simple things such as their usual routine. For this situation, it is necessary to study the different factors. This lets us create an algorithm which can help to predict the crowdedness at a gymnasium.

B. An overview of data

The data analysed for predicting the crowdedness is detailed data of the number of people (students from university) who went to the gym last year. The data-set has 26,000 records. Besides that, data has some extra information such as weather and semester-specific information is also available that might help to look for different patterns. This data is collected from Kaggle and have been recognised as our baseline for all oprations.

The data about the people who go to gym is provided in Fig. 1. Data contain factors like date, month, timestamp, weekday records, temperature, and hours spent in the gym along with other factors such as weekends and semester records.

Sr. No.	Column	Datatype
1	number of people	Integer
2	date	Object
3	timestamp	Integer
4	day of week	Integer
5	is weekwnd	Integer
6	is holiday	Float
7	is start of semester	Integer
8	is during semester	Integer
9	Month	Integer
10	hour	Integer

Fig. 1. An Overview of Data

C. Scope

The use cases for predicting crowdedness at a particular place can be useful not only in gymnasiums and fitness studios but also various places like shopping malls, restaurants and stadiums. In this particular case, after analyzing the data, the external features such as semester records of students and weekend records can help to understand a prediction pattern.

II. The Proposed Goal

The inclusion of machine learning into the health sector holds promise for considerably improving the safety of people. The proposed objective of this project is to create an algorithm that will predict the crowdedness at the fitness centre considering all the factors from the data.

III. Ethical Concerns

As per paper [4], many ethical concerns need to consider while doing an analysis. Some of them such as privacy, informed consent, confidentiality, and risk of harm are explained below with some concerns if they are available inside the data for a prediction that can help us to avoid ethical concerns [5].

A. Cataloguing the Stakeholders

The main stakeholders in this study are the gymnasium and fitness studios where fitness enthusiasts go to stay fit. This is the main source of all the information [4].

B. Voluntary Participation and Consent

The dataset about people count went gym contain no private data about any individual being or organisation. As the data is present in tables contains no classified information and publicly accessible for research. So for this particular dataset, there is no need to consider voluntary Participation and consent from the stakeholders [5].

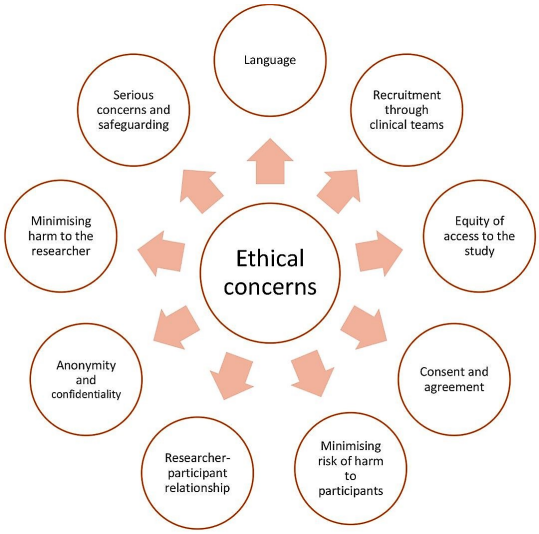

Fig. 2. Ethical Concerns

C. Risk of Harm

Considering the data does not contain any personal information. so we can assume that the data does not cause any harm for the research work [4].

D. Justice

In this dataset, the goal is to predict crowdedness and it is not directed towards any particular organisation or social club and so it does not damage to any individual or organisation [4].

E. Public interest

The data is originally obtained from the Kaggle website and it is available for public use. To dodge all the ethical issues related to data, all necessary precautions are taken [5].

F. Safeguards

Before proceeding with the analysis of data, to protect the identity of the contributors, their names and other details were removed from the data. The visual analysis will help to learn if biasing is available to recognise patterns of crowd density throughout the year [5].

IV. Strategy

For this dataset, Cross-Industry Standard Process for Data Mining (CRISP-DM) is used which is a data mining technique. In CRISP-DM, various steps are used as shown in figure 2 to get accurate result and this result can help to support a business strategy. After recognising the goal to accomplish a business strategy(Business Understanding Phase), the data requires to be analyzed (Data Understanding Phase) and processed (Data Preparation Phase). After that a various models are performed and evaluated to get accurate results(Modelling Phase and Evaluation Phase) [6].-

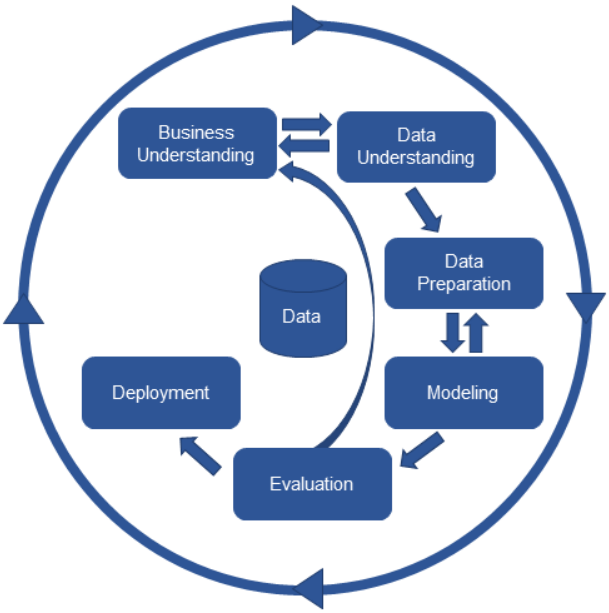

Fig. 3. CRISP-DM

A. Business Understanding Phase

This phase concentrates mainly on the business idea. All other requirements of the project are fulfilled at business level.

B. Data Understanding Phase

This phase involves the identifying the structure of data and recognising any features. This process involves recognising internal and external factors.

C. Data Preparation Phase

In this phase, the data is modified from raw data and transformed it into the final dataset to build a machine learning model.

D. Modelling Phase

This phase builds a model that represents the analysed data(for example, using a model predict the target value). Different models are performed and assessed on the given data. Depending on the type of data, sometimes the data preparation phase might not work for data modelling phase. In such a case, the data preparation phase is modified as per requirement and then modelling phase started again.

E. Evaluation Phase

In the evaluation phase, a model assessed by its performance and efficiency.

F. Deployment Phase

If the executed model is not satisfactory for use to support the business idea, the alternative to the model is defined. Otherwise, the model is built in a real-time environment (Developement phase).
Using all the phases above can help the organization to execute the business strategy.

V. Visualizations

Data Visualization is the first part of analysing the data set in the preparatory stage. The dataset have multiple internal and external factors to take into consideration for the visualizations. Every factor may or may not have some relationships. The relationship and correlation between the factors are analysed through different visualizations to gain better knowledge. Many correlations can be performed based on the factors in the dataset such as weekday and the start of the semester. The relation between the factors with the crowdedness has been performed in the form of many graphs.

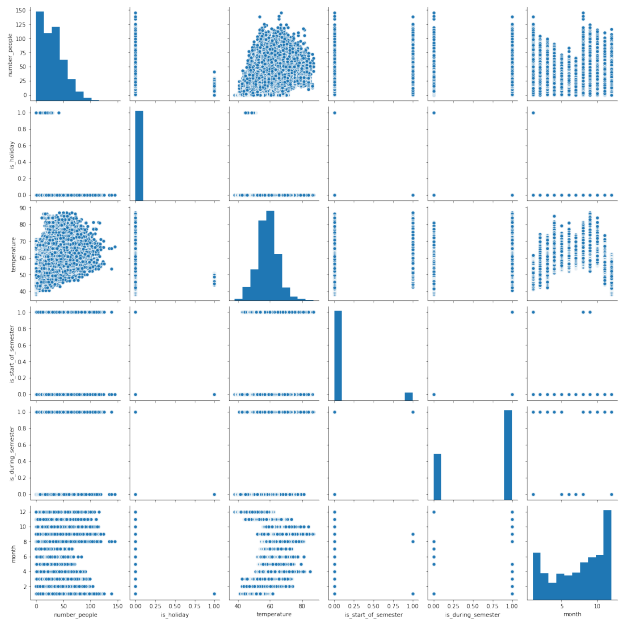

Fig. 4. Pairplot

A. Pairplot and Heatmap of Data

Below we have pair-plot in fig 6, used to get a complete overview of the data. The graphs show the histograms of the each factor and likewise shows the scatter plot with every other factor.

Let's observe how data is correlated with the help of heatmap. There is a positive relationship between the number of people attending the gym and time they are going to the gym.

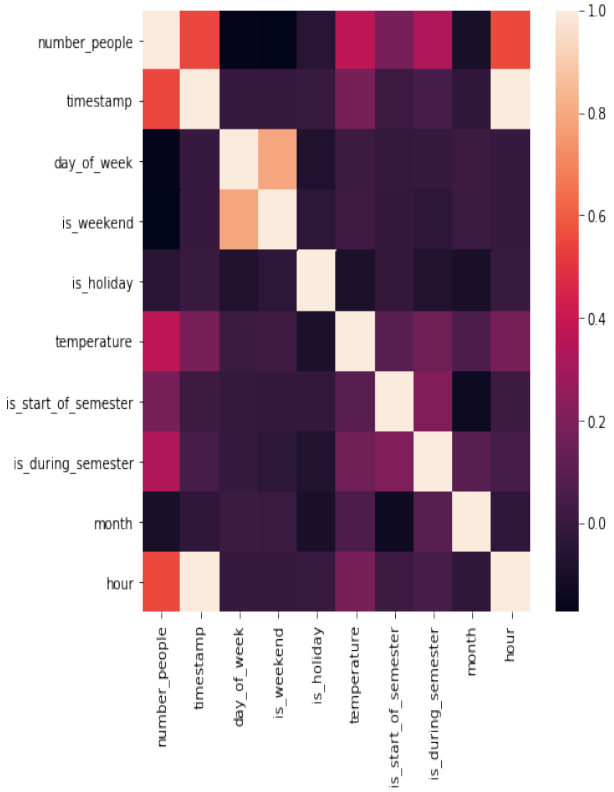

Fig. 5. Heatmap

From the above graph, few things can be understood

- The target variable 'number of people attending the gym' (crowdedness) and temperature shows Gaussian distribution.
- There is a positive correlation present between semester and the number of people attending the gym.
- There is a strong relationship present between apparent temperature and temperature which helps to predict a model, we need to eliminate one of them.

B. At what time people go to the gym

Below graph shows us a pattern in which more number of people are going to the gym at the higher timestamp. Meaning that more number of people prefer to go to the gym in the evening or late in the day than early morning.

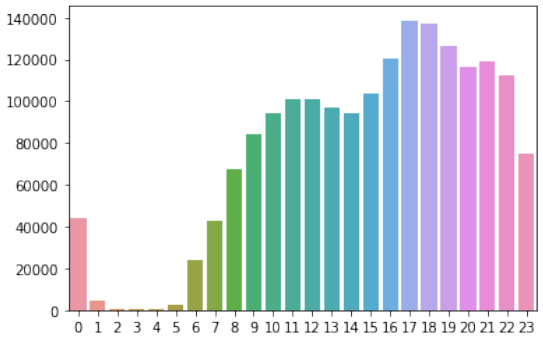

Fig. 6. Barplot: Timestamp Vs NOP

C. When people prefer to go to the gym

There is a negative correlation present between weekday and the number of people attending the gym as we can clearly see trend in the heatmap. We can say that during the weekdays more number of people prefer to go to the gym and the graph steadily decreases till Sunday.

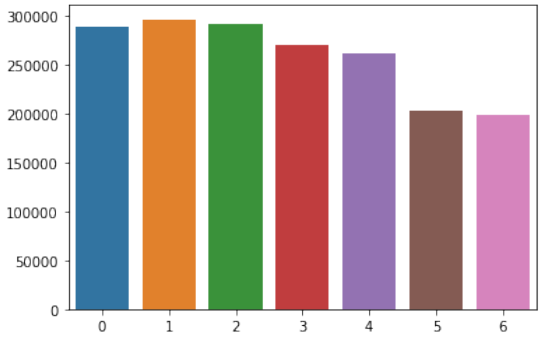

Fig. 7. Barplot: Weekdays Vs NOP

D. What temperature people prefer to go to the gym

There is a positive correlation present between the number of people going to the gym and temperature.

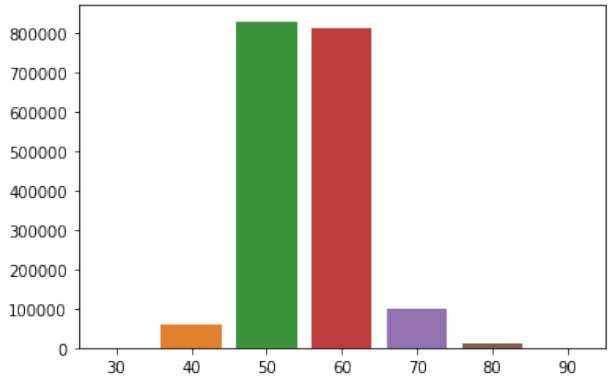

Fig. 8. Barplot: Temperature Vs NOP

Even though from graphs it appear that more people go to the gym in warm weather, this factor can be misleading for analysis as the average weather remains in range 50 to 70 Fahrenheit. So the fact that more people go to the gym between 50 to 70 Fahrenheit temperature is not totally true.

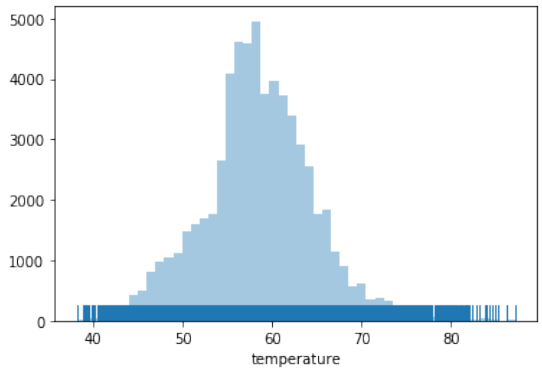

Fig. 9. Histogram: Temperature Vs NOP

E. When in semester people prefer to go to the gym

From Pairplot and Heatmap, it can be observed that more people go to the gym at the start of the semester. From calendar heatmap, it can be observed that more people do not prefer to go to the gym at the end of the semester.

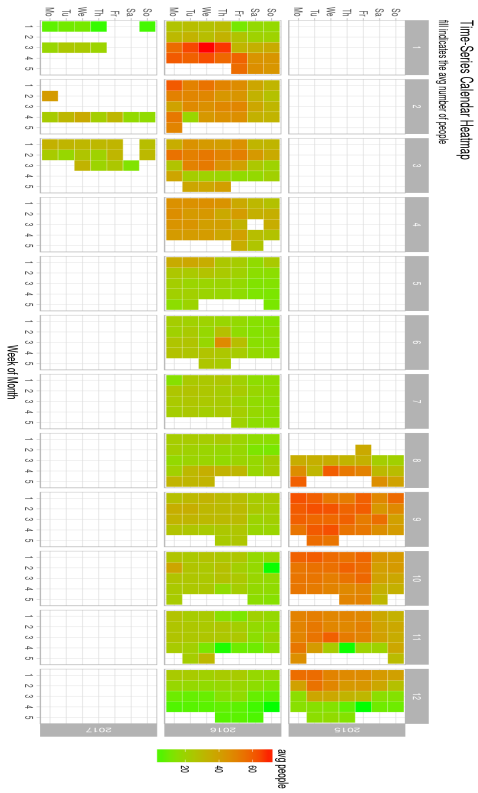

Fig. 10. Calender Heatmap

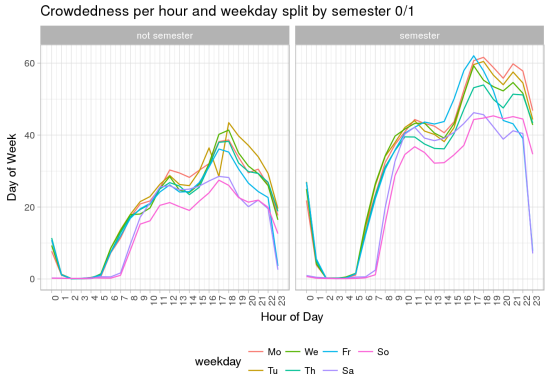

Fig. 11. Non-Semester Day and semester Days

VI. Applicable Techniques

For prediction using machine learning algorithms, the independent variable considered is 'Number of people'(i.e. crowdedness). Three machine learning models are used to get predictive value. Detailed research is done on the three models to generate the best results [7]. Root Mean Squared Error (RMSE) is used to compare the performance and differentiate between expected and predicted results. The three different machine learning models used are Support Vector Machine Model, Decision Tree Model and Random Forest Model [8].

A. Decision Tree Model

The Decision tree model can be a good fit for the prediction of the number of people attending the gym. This decision tree model can affect only on few factors to get the best predictive result and with the help of RMSE, we can decide whether to use or not use this model. If the predicted result is the same as the expected result than the model can be used for the prediction of crowdedness [8].

B. Random Forest Model

The Random Forest Model is nothing but a revised model of a decision tree model where a group of trees attached to nodes. This model can assist in overfitting problem of the model while making a choice. This model takes input at the first node and till the last nodes of the tree. As a random forest is a revised model of a decision tree that gives a fitter result compare to decision tree model [8].

C. Support Vector Machine Model

A support vector machine model is nothing but a supervised machine learning model that uses classification algorithms. After providing training data for each class, A SVM can categorise to a new class. Support Vector Machine is capable of deciding the boundary to categorise classes and maximize the margin [9].

References

[1] A. Pan, L. Liu, C. Wang, H. Guo, X. Hao, Q. Wang, J. Huang, N. He, H. Yu, X. Lin *et al.*, "Association of public health interventions with the epidemiology of the covid-19 outbreak in wuhan, china," *Jama*, vol. 323, no. 19, pp. 1915–1923, 2020.

[2] X. Wu, G. Liang, K. K. Lee, and Y. Xu, "Crowd density estimation using texture analysis and learning," in *2006 IEEE international conference on robotics and biomimetics*. IEEE, 2006, pp. 214–219.

[3] W.-L. Hsu, K.-F. Lin, and C.-L. Tsai, "Crowd density estimation based on frequency analysis," in *2011 Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. IEEE, 2011, pp. 348–351.

[4] D. R. Thomas, S. Pastrana, A. Hutchings, R. Clayton, and A. R. Beresford, "Ethical issues in research using datasets of illicit origin," in *Proceedings of the 2017 Internet Measurement Conference*, 2017, pp. 445–462.

[5] P. Henderson, K. Sinha, N. Angelard-Gontier, N. R. Ke, G. Fried, R. Lowe, and J. Pineau, "Ethical challenges in data-driven systems," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 123–129.

[6] S. Moro, R. Laureano, and P. Cortez, "Using data mining for an application of the crisp-dm methodology," 2011.

[7] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Modelling crowd scenes for event detection," in *18th international conference on pattern recognition (ICPR'06)*, vol. 1. IEEE, 2006, pp. 175–178.

[8] N. Runge, N. Kilian, J. Smeddinck, and M. Krause, "Predicting crowd-based translation quality with language-independent feature vectors," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[9] X. Feng, X. Ling, H. Zheng, Z. Chen, and Y. Xu, "Adaptive multi-kernel svm with spatial–temporal correlation for short-term traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2001–2013, 2018.