# Introduction:

## Background Information

Manzo is a local restaurant chain that was started in 2017, their team is composed of young Emiratis with diverse range of talents and interests such as finance, human resource management, IT, architecture, design and photography. The restaurant operates in the UAE with 5 branches in Abu Dhabi and A1 Ain, where it specialises in international cuisines. Its operations involve both delivery of food products to its customers or pick up services, depending on the customer’s preferred service (Arunkumar and Kannimuthu, 2020).

This project allows the companies to monitor the evolution of their branding and investment, search for groups of branding and investment with similar behaviours and predict the final grade that each of them could have. When predicting the restaurant’s performance, the system is based on a set of data collected in previous courses. In this way, the objective of our system is to facilitate investment and take advantage of the enormous amount of data collected from previous years (Blikstein and Worsley, 2016). Currently, we can see that there are different types of projects that use statistical techniques or techniques based on machine learning to give users a global view of the system and predict situations or data that help them make decisions. An example would be financial systems analysis applications, these show statistical graphs and predict possible economic results. Based on this data, a management team would be able to make better decisions. This makes a huge impact and over time has become a fundamental part of the industry (Arunkumar and Kannimuthu, 2020).

However, these techniques can not only be applied to the economic sector. They are also used in projects such as: search engines, antivirus, text comprehension applications or image recognition, among others. Despite how well these types of techniques have worked in these sectors, today there are few lines of work that use these techniques focused on the business environment (Blikstein and Worsley, 2016). If we observe the social context of the business environment that we have just mentioned, it can be seen that for a long time many studies have been carried out that investigate the causes and factors that lead branding and investment to abandon their respective degrees. After these investigations, we can highlight some factors such as the socioeconomic problems of the individual, the lack of knowledge of the degree before entering it or even the lack of adequate prior knowledge (Blikstein and Worsley, 2016).

However, one of the factors with the highest rate of recurrence that we can find is the lack of adequate monitoring of the fiscal year, which is usually due to different causes. Among them we highlight:

* Previous ignorance of the pace or particularities of a specific subject by the restaurants. This leads to poor time management.
* The degree of complexity that the restaurant is presented with when carrying out a business follow-up of their branding and investment (Blikstein and Worsley, 2016).

All these causes and factors, as well as the social and technological context in which we have found ourselves, lead us to the conclusion that the use of a statistical/predictive system applied to business monitoring would improve performance on both sides. The business results of a restaurants correspond to a series of socioeconomic variables and a behavioral pattern of customers in the UAE (Blikstein and Worsley, 2016). In most cases, the previously mentioned variables are usually unknown to the teacher, but the performance in the restaurant industry is not. A restaurant's behavior and attitude towards the subject can be predicted based on previous experiences, to the point of knowing, roughly, what a restaurants 's performance is in the UAE based on the data analytics (Blikstein and Worsley, 2016).

## Problem Statement

Business development is turning out to be more conspicuous for hopeful organizations and new companies and more endeavours are being placed into expanding the development rates. In any case, a specific breaking point for development might be arrived at after which it becomes more diligently to extend a business and arrive at a pre-determined target. Manzo restaurant has been working beginning around 2017, and there have been different branches that were shut because of not gathering the designated profit from investment of 33% and capital turnover in a half year. Accordingly, they are searching for the best area to grow their business and guarantee it is finished in the most productive manner. This undertaking intends to find intriguing bits of knowledge from concentrating on the information connected with a developing nearby restaurant chain that could be additionally broke down utilizing information examination approaches as opposed to applying the normally utilized measurable methodology. These experiences can assist with figuring out the business issue and distinguish open doors for investments which target growing the local restaurant chain and increment the quantity of clients it serves every week.

## Research Objectives

The project presented in this document aims to create a prototype of a tool that allows the teacher to monitor his class. This monitoring will be done by viewing statistics, grouping branding and investment with similar behaviours and predicting grades. Although the development will focus on the specific subject of Interactive Systems Development (ISD), the system will consist of a modular architecture that allows it to be easily expanded (Ding and Guo, 2018). In this way, in the future work section you can adapt to new subjects and add new features. For this, the project will be based on the visualization of the data in a pleasant format for the user, the creation of behavior models of the branding and investment who study a certain subject and the use of automatic learning techniques to try to predict the performance of the restaurant in the UAE (Blikstein and Worsley, 2016).

In this version of the project, we focus on the development of the internal architecture of the system and the functionality dedicated to the teacher. Despite this, we will take special care so that, in future work, it is easy to support the restaurants functionality (Ding and Guo, 2018). For the implementation of the system, some objectives have been set, which we will detail below:

* The tool must be able to obtain a series of data from a series of excel files or with a '.csv' extension with a previously defined format and transform them into information that is easy to use and relevant to the system.
* The system should provide an important section dedicated to the visualization of statistical graphs. It will be done from the data that the teacher provides to the system about the course he teaches (Ding and Guo, 2018).
* Detection of groups of branding and investment with similar behaviours and visualization of these groups through graphs. It will also show a list that indicates in which the restaurant operates (Ding and Guo, 2018).

A simple prediction system will be provided. This must show, for local restaurants in the class, a prediction of their possible final grade based on their performance in the subject and comparing it with the performance of branding and investment in previous courses. This mark will be between one of the following values: failing, approved, remarkable and outstanding (Ding and Guo, 2018).

This last point that we have just mentioned, aims to investigate the viability of predictive algorithms in this environment and, although the project will consist of a simple prediction system, this part of the project will remain for study and improvement in future work. With these objectives fulfilled, the system offers a way of monitoring the business year that aims to obtain an improvement in its performance (Goodnight, 2011).

# Research Methods

## Research Method – CRISP- CM

CRISP-DM (Cross Industry Standard Process for Data Mining) gives a normalized depiction of the existence pattern of a standard information examination project, similar to computer programming with improvement life cycle models of programming. The Fresh DM model covers the periods of a venture, their particular undertakings, and the connections between these errands. At this degree of portrayal, it is beyond the realm of possibilities to expect to recognize all connections; Connections could exist between any undertaking relying upon the objectives, the specific situation, and the client's advantage in the information (Qin et al., 2018).

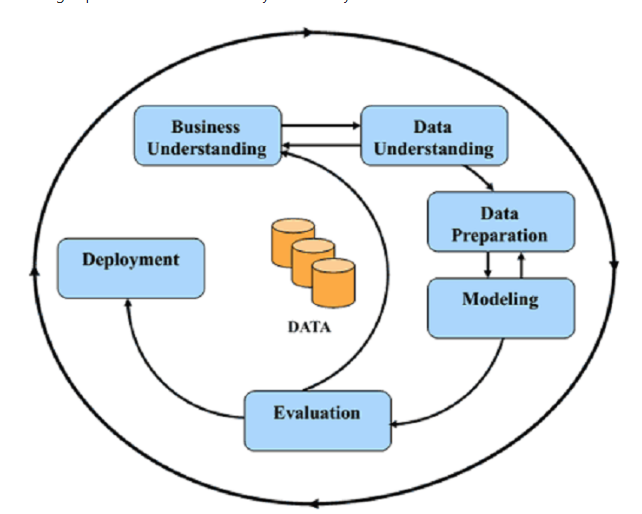


Figure 7: (Qin et al., 2018)

The Fresh DM procedure sees the information examination process as an expert undertaking, hence laying out a lot more extravagant setting that impacts displaying. This setting considers the presence of a client that isn't important for the improvement group, as well as the way that the task not in the least doesn't end once the ideal model is found (since later organization and support are required) , rather it is connected with different undertakings, and should be totally reported so other improvement groups can utilize and expand on the information acquired (Qin et al., 2018).

It briefly describes each of the phases.

***Phase I. Business Understanding. Definition of customer needs (understanding of the business)***

This underlying stage centers around grasping the undertaking goals. This information on the information is then transformed into the meaning of an information mining issue and a starter plan intended to accomplish the targets (Qin et al., 2018).

***Phase II. Data Understanding. Study and understanding of the data***

The information understanding stage starts with the underlying information assortment and go on with exercises to get comfortable with the information, distinguish quality issues, find starter information about the information, or potentially find intriguing subsets to frame speculations about. the secret data (Rahman and Reza, 2022).

***Phase III. Data Preparation. Data analysis and feature selection***

The data preparation phase covers all the activities necessary to build the final data set (the data that will be used in the modeling tools) from the initial raw data. Tasks include table, record, and attribute selection, as well as data transformation and cleansing for modeling tools (Rahman and Reza, 2022).

***Phase IV. Modeling. Modelled***

In this stage, the demonstrating methods that are pertinent to the issue are chosen and applied (the more the better), and their boundaries are aligned to ideal qualities. There are commonly a few procedures for similar sort of information mining issue. A few procedures have explicit prerequisites on the type of the information. In this way, quite often in any venture you wind up returning to the information arrangement stage (Rahman and Reza, 2022).

***Phase V. Evaluation. Evaluation (obtaining results)***

At this stage in the venture, at least one demonstrates have been fabricated that have all the earmarks of being of adequate quality from an information examination point of view. Prior to continuing to the last arrangement of the model, it is critical to completely assess it and survey the means executed to make it, contrast the got model and the business goals (Rahman, 2018). A key goal is to decide whether there are any significant business gives that poor person been adequately thought of. Toward the finish of this stage, a choice on the utilization of the consequences of the information examination interaction ought to be gotten (Rahman, 2018).

***Phase VI. Deployment. Deployment (put into production)***

By and large, the production of the model isn't the finish of the task. Regardless of whether the goal of the model is to expand the information on the information, the information acquired should be coordinated and introduced so the client can utilize it (rojanala, 2021). Contingent upon the prerequisites, the improvement stage can be basically as straightforward as creating a report or as mind boggling as playing out an occasional and maybe robotized information examination process in the association (Rahman, 2018).

## Data Sources

One of the most important factors that determines the outcome of the data analysis is the data source and the quality of the data being studied. As the machine learning model is studied to detect patterns from the data being fed to it, the data accuracy and representativeness determine the effectiveness of the model training and in turn will affect the overall accuracy of the predictions and the reliability of the machine learning model in general. The data used for this study has been collected from the local restaurant and it includes the revenues for all five branches in different areas, the data for sales and delivery as well as competitor’s analysis (rojanala, 2021).

## Software Used

The libraries that will be discussed in this section are "Pandas", dedicated to data analysis, and "Scikit-learn", dedicated to data mining and analysis, as will be seen below.

***Pandas –*** Itis a free software library implemented for the Python language that is dedicated to data analysis. It was initially developed with the aim of managing financial data. However, its development has been escalating over time, leaving us with a multitude of useful functions not only in that sector (rojanala, 2021). This library has a "BSD" free software license and implements a multitude of operations that facilitate data processing. Currently, it is one of the most widely used language extensions in many projects that require efficient information processing. This package contains a series of features that make it one of the most used packages in data analysis (rojanala, 2021). Some of these features are:

* It includes new data structures that make it easier to handle large amounts of data. The most important data structures included for our project are the Data Frame and the Series.
* More efficient versions of structures already implemented for the storage of information (Sriraam, 2022).
* Multilevel indexing that allows powerful control over table groupings and large data collections.
* Simple support for input/output operations from files such as Excel or “.csv”
* Advanced sequence range generator that provides a wide range of possibilities for loops with a greater degree of control (Sriraam, 2022).

Thanks to the characteristics that we have just mentioned, this library adapts perfectly to the present project.

***Scikit-learn*** - It is a library implemented for the Python language and has a large community that supports its development. It is currently the most used library in this language for projects that include machine learning techniques. It provides a large number of functionalities in data mining and analysis, and is also perfectly structured in its fields of action (Yang, Yuan and Li, 2020). These fields, as we can see on their website, are the following:

* Classification
* Regression
* clustering
* Data Dimension Reduction
* Model selection and generation
* pre-processing

In this way, this library allows us to implement the entire part of the project dedicated to grade prediction and clustering (Yang, Yuan and Li, 2020).

## Data Exploration

Exploratory data analysis will be one of the first steps in the analysis and will help in making the decisions for the data modelling step that precedes the application of machine learning techniques. Python is a powerful tool that will help visualising the data even in the very first steps of data pre-processing, as it can help in providing insights on the distribution of missing data using a heatmap. The Python Pandas package can also help in providing summary statistics and descriptive information on the data which includes the data types for each feature and the different values in each column (Zulkipli, Satari and Wan Yusoff, 2020). Univariate Analysis can be performed to better understand the values for each column and to explore the distribution of outliers using boxplots for example. Bivariate Analysis can be performed using the Seaborn Pair plot to compare each two features separately and find possible correlations (Yang, Yuan and Li, 2020).

## Data Mining

Correlations between different variables can be implemented using a correlation matrix and visualising it in a heatmap. This can help in finding the most correlated features with our target variable (Zulkipli, Satari and Wan Yusoff, 2020). Other techniques could be used like applying association rules to indicate the strength of the relation between variables. Another approach would be to use a tree-based model to identify the most important features for making a prediction, the model then has to be evaluated on accuracy and performance (Zulkipli, Satari and Wan Yusoff, 2020).

Various ways can be used to evaluate the accuracy and performance of the model, including: Accuracy, Precision, Recall, Specificity and F1 score (Zulkipli, Satari and Wan Yusoff, 2020). For Example, a cancer study to identify patients who are predicted to have cancer and those who actually have cancer there are some common terms that are explained below for the sake of simplicity and demonstration:

* True positives (TP): Predicted positive and are actually positive.
* False positives (FP): Predicted positive and are actually negative.
* True negatives (TN): Predicted negative and are actually negative.
* False negatives (FN): Predicted negative and are actually positive.