CCT College Dublin

Assessment Cover Page

Module Title:	**Programming for Data Analytics** **Statistics for Data Analytics** **Machine Learning** **DataPreparation and Visualization**
Assessment Title:	**Data Analytics CA2- Comparison of the Five Years Apple Producer Price Index of the Countries of Ireland and Sweden**
Lecturer Name:	**Sam Weiss** **David Mcquaid** **Dr. Muhammad Iqubal** **John O'Sullivan**
Student Full Name:	**Tayfun Tekin**
Student Number:	**2022323**
Assessment Due Date:	**06.01.2023**
Date of Submission:	**06.01.2023**

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

Within this project's scope, datasets consisting of various data columns belonging to the countries of Sweden and Ireland were used. Using these datasets, the Apple value in the Item column was tried to be estimated using the Value column. The algorithms used during this study are Linear Regression and Decision Tree Algorithms.

As a result of the operations, the Decision Tree Algorithm R2 score is 0.98, and the Linear Regression R2 score is 0.99. In light of these values, it can be said that the prediction ability of the algorithms is high.

1. Method Used in the Project

During the writing of the report, the CRISP-DM method was followed. In determining the titles, each stage of this method was chosen. The operations performed under the relevant headings will be summarized, and the parameters used will be mentioned.

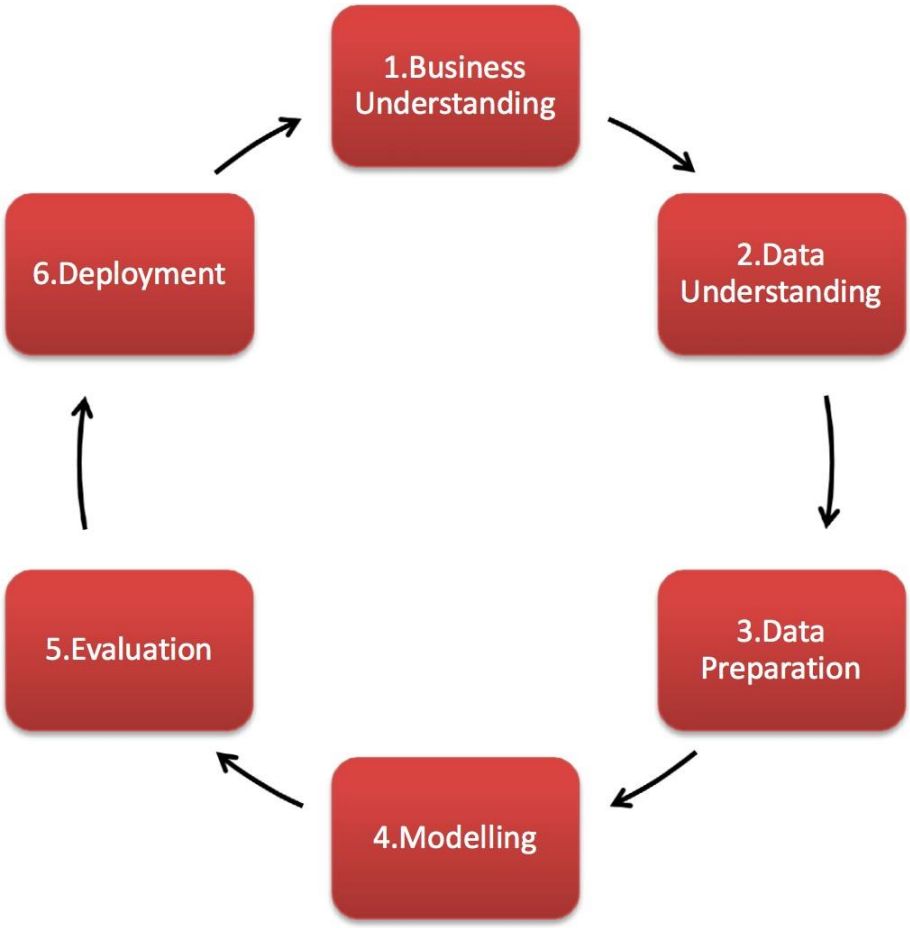

Figure 1: CRISP-DM stage

2. Stages of CRISP-DM in Project

As seen in the figure above, CRISP-DM stages:

- Business Understanding,
- Data Understanding,
- Prepare Data,
- Model Data
- Evaluation and
- Deployment

In this study, we will talk about the processes carried out in related fields.

2.1. Business Understanding

This section will focus on the main idea of the study. This main idea, which is also the subject of the study, is to estimate the Apples variable in the Item column through the variables in the Valuecolumn, using the Decision Tree Algorithm and Linear Regression algorithms, and to evaluate theperformance of the algorithms using the R2 metric

2.2. Data Understanding

The Understanding the data section has added the relevant libraries to the work environment. Thedata sets it will use later are included in our working environment with the names df1 and df2. Totake a quick look at the added datasets, the head() function that comes with the Python language is used, and the first five variables of the dataset are printed on the screen thanks to this function.The data sets it will use later on are included in our working environment with the names df1 anddf2.

The variable to be estimated, the Apple variable, was printed on the screen by sorting the Item column according to the number of elements it contains. While doing this, the value_counts() function, another function that comes with the Python language, is used

After all these operations, the for loop is created, and a circle is set up over the variables. Thanks to this loop, the number of non-unique variables is printed on the screen.

In the next stage, to get more information, the info() function, which is another function that comeswith the Python language, is applied to the datasets added to the working environment, and more information about the data is obtained.

In line with the information obtained, it was noticed that there were missing data in the datasets; the IsNull () function, which came with the Python language, was used to learn the exact number of missing data. The sum() function was used to show the total number.

After all these processes, it is time to understand the statistical datasets. In line with this need, thedescribe() function, which comes with the Python language, is applied to both datasets and the basic statistical values of the data are printed on the screen.

As a result of basic statistical operations, advanced statistical calculations were made. These calculations are,

- Wilcoxon Test,
- One-Sample Hypothesis Test,
- T-Test,
- Analysis of Variance,
- Mann Whitney U Test.

2.3. Data Preparition

At this stage, the relevant data has been converted into numerical values using Label Encoder. In this way, the correlation values of the transformed variables were found with the help of the corr()function that comes with the Python library.

In order to obtain statistical summaries of these transformed variables, the describe() function, which was previously used in this project, is used.

In order to better understand the data, “Year” and “Value” expressions in the data are visualized with the sns function that comes with the seaborn library. The “Value” variable, which will be used as a variable later, was used with boxplot and outliers in the data were expressed.

All numerical data are visualized with a created for loop, and the seaborn library's countplot function is used during this visualization process.

2.4. Data Modelling

The expressions to be used as dependent and independent variables here are standardized with thehelp of a standard scaler. Then, with the train_test_split function, test_size=0.3 is split to random_state=0. After these processes, the data sets are ready for the algorithms to be used.

The first algorithm applied to the dataset is the Decision Tree Regression algorithm. When this algorithm was applied to the dataset, the R2 score was 0.98 and Train Accuracy was 1. From thispoint of view, it was revealed that the algorithm had learned too much, and one of the biggest problems of Decision Tree Regression was encountered. After applying GridSearchCV to this algorithm, the best score would be 0.95 and the parameters required for this score were printed onthe screen

The second algorithm applied to the dataset is Linear Regression. When Linear Regression was first applied, the R2 score was 0.40, and the Train Accuracy result was relatively low at 0.36. Afterthe GridSearchCV process was applied, the score of the algorithm increased.

3. Information about the Algorithms Used

3.1. The Decision Tree Algorithm.

The Decision Tree algorithm is an algorithm that allows the nodes to be built from top to bottom, starting from the root node, and the data with similar values (homogeneous) to be divided into subsets. It can also be used as a classification method. In this project, this algorithm was used as a Regressor.

3.2. Linear Regression

Linear regression is a data analysis technique that estimates the value of unknown data using another relevant and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation.

4. Evaluate

In this section, the algorithms applied to the dataset within the scope of the project areevaluated, and information about the algorithms is given. At the same time, information about the R2 parameter, which is used as an evaluation scale, is given. Machine Learning algorithms are used for a specific situation and for similar situationsunknowns are also used to predict outcome.

Regression problem is considered in this project. Regression is basically a statistical approach to find the relationship between variables. It is used in machine learning to predict the outcome of an event based on the relationship between the variables obtained from the dataset.

4.1. Evaluation of Regression Models

There are many methods for evaluating regression models. These :

1. R2,
2. Adjusted R2,
3. Mean Absolute Error (MAE),
4. Mean Squared Error (MSE),
5. Confusion Matrix,
6. F1 Score,
7. AUC-ROC.

In this study, we used the R2 evaluation metric

4.2. R2.

R2 is a statistical measure of how close the data are to the fitted regression line. It is also known as coefficient of determination or multiple coefficient of determination for multiple regression.The fact that R2 = 1 is proof that the experimental data provides a perfect linear curve. This isa value that is not expected to be very possible in the real world. If we have such an R2 valuein the algorithm, it is likely that there will be problems such as "overfitting"

in the algorithm. On the other hand, the more data points there are, the higher the reliability of R2. To make R2more understandable, an example can be given as follows. For example, if R2=085, 85% of the total variation in the y variable can be explained, while 15% cannot.

Two questions can be asked here. Are these respectively low R2 values always a problem? And Are High R2 values always great? The answer to both questions is no. Some fields of study have more unexplained variation in nature. In these areas, your R2 values are bound to be lower. For example, studies trying to explain human behavior. Fortunately, if you have a low R-squared value but the independent variables are statistically significant, important conclusions can still be drawn about the relationships between the variables. Statistically significant coefficients continue to represent the mean change in the dependent variable, given

a one-unit shift in the independent variable. Obviously, being able to draw such conclusionsis vital.

5. Discussion of Result

Within the scope of this project, datasets from Ireland and Sweden were analyzed using Linear Regression and Decision Tree algorithms. The results were evaluated on the R2 scale, and excessive learning problems were encountered in Decision Tree and Linear Regression algorithms. Although the GridSearchCV method is applied, the over-learning problem in the algorithms has not been solved.

5. **Source**

- **Steele, J. and Iliinsky, N.P.N. (2010). Beautiful visualization : [looking at data througt the eyes of experts]. [online] Sebastopol, California: O'reilly. Available at: https://www.goodreads.com/book/show/7405941-beautiful-visualization [Accessed 4 Jun. 2022].**
- **George, N. (2021). Practical Data Science with Python. Packt Publishing Ltd.**
- **Mckinney, W. (2018). Python for data analysis : data wrangling with pandas, NumPy, and IPython. Sebastopol, Ca: O'reilly Media, Inc., October.**
- **Fenner, M. (2019). Machine Learning with Python for Everyone. Sydney: Addison Wesley.**
- **Bruce, P.C., Bruce, A. and Gedeck, P. (2020). Practical statistics for data scientists : 50+ essential concepts using R and Python. Sebastopol, Ca: O'reilly Media, Inc.**
- **McDaniel, E. and McDaniel, S. (2012). The Accidental Analyst. Createspace Independent Pub.**