A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name: Cristhian Elson Pereira Macedo* |  |
| *Student Number: 2024104* |  |
| *Module Title: Statistical Techniques for Data Analysis* |  |
| *Assessment Title: CA2* |  |
| *Assessment Due Date: Sunday, 19 May 2024, 11:59 PM* |  |
| *Date of Submission: Sunday, 19 May 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.



**Exam Scores Dublin Secondary School, Diamonds, PlantGrowth and Trees Dataset**

**Cristhian Elson Pereira Macedo**

**2024104**

Higher Diploma in Science in Data Analytics for Business

Statistical Techniques for Data Analysis (StatsDA)

Lecturer: Kayoum Khbuli

CCT College Dublin

Dublin, Ireland

2024



**SUMMARY**

[**1.** **Introduction** 5](#_Toc167055630)

[**2.** **Objectives** 5](#_Toc167055631)

[**Task 1** 5](#_Toc167055632)

[**Task 2** 6](#_Toc167055633)

[**Task 3** 6](#_Toc167055634)

[**Task 4** 6](#_Toc167055635)

[**3.** **TASK 1 – Exploratory Data Analysis** 7](#_Toc167055636)

[**4.** **TASK 1 – Next Steps** 8](#_Toc167055637)

[**5.** **Task 1 – Exploring the data** 8](#_Toc167055638)

[**a)** **Machine Learning - DecisionTreeClassifier** 10](#_Toc167055639)

[**6.** **TASK 2 – Exploratory Data Analysis** 10](#_Toc167055640)

[**7.** **TASK 2 – Next Steps** 11](#_Toc167055641)

[**8.** **Task 2 – Exploring the data** 12](#_Toc167055642)

[**a)** **LDA - (Linear discriminant analysis)** 12](#_Toc167055643)

[**9.** **TASK 3 – Exploratory Data Analysis (Steps A, b, c)** 13](#_Toc167055644)

[**10.** **Task 3 – Exploring the data** 15](#_Toc167055645)

[**a)** **MACHINE LEARNING – Linear Regression** 15](#_Toc167055646)

[**b)** **MACHINE LEARNING – DECISIONTREECLASSIFIER** 15](#_Toc167055647)

[**c)** **MACHINE LEARNING – RANDOM FOREST** 16](#_Toc167055648)

[**d)** **MACHINE LEARNING – Logistic regression** 16](#_Toc167055649)

[**11.** **TASK 4 – Exploratory Data Analysis (Steps A, b, c)** 16](#_Toc167055650)

[**12.** **Task 4 – Exploring the data** 17](#_Toc167055651)

[**a)** **MACHINE LEARNING – Linear Regression (girth and volume)** 17](#_Toc167055652)

[**b)** **MACHINE LEARNING – Linear Regression (Height and volume)** 18](#_Toc167055653)

[**c)** **MACHINE LEARNING – Linear Regression (Girth and height)** 18](#_Toc167055654)

[**d)** **MACHINE LEARNING – KNeighborsregressor** 19](#_Toc167055655)

[**13.** **Conclusion** 19](#_Toc167055656)

[**References** 21](#_Toc167055657)

# **Introduction**

This project aims to complete the second Assessment Task for the module Statistical Techniques for Data Analysis from the course Higher Diploma in Science in Data Analytics for Business by CCT College Dublin, this is a document to describe the project itself whose name is “CristhianMacedo\_StatisticalTecHDip\_CA2.ipynb” and should be always with this document. This project for Statistical Techniques uses the programming language Python, the environment of Anaconda Navigator with Jupyter Notebook, with CRISP-DM methodology as project management.

Also using a free AI Writing Assistance Grammarly (Grammarly, 2009) to help with English grammar while typing. A GitHub Repository was created for this project and is available on [this link](https://github.com/CristhianMacedo2024104/Statistical-Techniques-CA2) and the complete URL is available on References (Macedo, 2024).

This assessment task consists of completing four tasks that are described in the next topic name as “OBJECTIVES” with four different datasets, the first one is about the exam scores (in percentages) of 50 students from a Dublin Secondary School, the second is about the Diamonds dataset, third is about “PlantGrowth” dataset from the “pydaset” library and fourth is about “Trees” dataset also from “pydaset” library.

The tasks are focused on statistics analyses with Hypothesis Testing using T-test, ANOVA tests, Chi-Square tests and more, with a short Exploratory Data Analysis, plots, graphics, tables and deploying Machine Learning.

# **Objectives**

Below are described all the requirements and instructions to be followed and applied in each task.

# **Task 1**

Load the dataset Q1.csv. It contains the exam scores (in percentages) of a sample of 50 students from a Dublin secondary school.

* 1. Find and comment on important summary statistics and produce an appropriate plot to summarise the dataset.
  2. One of the teachers is concerned about the performance of the students in the school. She suspects that their performance may be below the reported national average of 70%. Does the data show that her concerns are justified? Use a significance level of alpha = 0.05.
  3. Produce and comment on an appropriate plot to illustrate your findings.

# **Task 2**

Load the diamonds dataset, and print the first 5 rows. The color variable refers to the colour of the diamond, with categories from “D” to “J”. Colourless diamonds are considered better than diamonds with a yellow tint. Diamonds from “D” to “F” are considered colourless, and diamonds from colour “G” to “J” are not considered colourless (that is, they have a very faint colour).

1. Create a new binary variable in the dataframe called “colourless” which records 1 in rows with colourless diamonds and 0 otherwise.
2. Perform an appropriate hypothesis test to determine whether there is any association between the clarity of a diamond and whether it is colourless or not. Use a significance level of alpha = 0.01.
3. Produce and comment on an appropriate plot to illustrate your findings.
4. Find and interpret 90% confidence intervals for both the mean price of colourless diamonds and the mean price of non-colourless diamonds.

# **Task 3**

Load the PlantGrowth dataset from the pydataset library. It contains the results of a small study comparing the yields of plants obtained under a control and under two different treatment conditions.

1. Find and comment on important summary statistics by treatment and produce an appropriate plot to summarise the dataset.
2. Conduct an appropriate hypothesis test to see if there is evidence of a difference between the three means (that is, the control and the two treatments). Use a significance level of alpha = 0.05.
3. If there is evidence of a difference between the three means, find and comment on where this difference may be.

# **Task 4**

Load the trees dataset from the pydataset library. It contains measurements of the diameter, height and volume of timber in 31 felled black cherry trees. Note that the diameter (in inches) is labelled girth in the dataset. It is measured at 4 foot 6 inches above the ground.

1. Perform a correlation analysis between all numerical variables. Include and comment on the results of hypothesis tests for the population correlation coefficients between all three pairs of variables (you can use the pearsonr function from the scipy.stats library).
2. There is interest in estimating the volume of timber from trees using either the girth or the height of the trees, or both. Perform a regression analysis to decide which of the three possible models you would recommend using. Interpret your results and provide a short conclusion of your findings.

# **TASK 1 – Exploratory Data Analysis**

First, it was imported some important libraries to be used in all Assessment Tasks such as Pandas, NumPy, Seaborn, MatPlotLib and Scipy, also used a filter to ignore warnings, next used important commands to understand the Q1 dataset, with the student's exam scores. Used “Head” to get the first 5 rows of the dataset, “shape” to get the size of the dataset in this case 50 observations (Rows) and 1 feature(s) (Column(s)), “info” to get a concise summary, “describe” to get statistical information of the numerical feature whose name is “exam\_score”.

Next typed some commands to get statistical values with the results: Length: 50.00, Mean: 68.74, Standard Deviation: 9.75, Min: 48.73, Percentile 25%: 62.04, Percentile 50%: 68.39, Percentile 75%: 74.63, Max: 87.36, Sum: 3436.78, Median: 68.39, Mode(s): 48.73, Range: 38.63, Variance: 93.16, Interquartile Range (IQR): 12.59 and Coefficient of Variation (CV): 14.04.

Checked, duplicate observations in this case, and there are none, next plotted a Distribution graphic of Exam Scores to check how the distribution is, in this case, the distribution is "considered a normal distribution" and is symmetric data due to the values Mean/Median being practically the same values and being in the centre of the graph, with the Statistical Values: Skew: 0.01, Mean: 68.74, Median: 68.39 and Standard Deviation: 9.75.

Also, used a boxplot, violin plot, hist plot, kde plot with three different parameters, to understand better how the distribution of Exam Score is, also, used an example learned from the website "towardsdatascience.com" in the topic "Hypothesis Testing with Python: Step by step hands-on tutorial with practical examples" (Polat, 2022), to check the distribution of the feature to confirm the distribution and after calculations, a result of p-value: **0.6557** and Fail to reject null hypothesis >> The data is normally distributed, it was obtained.

# **TASK 1 – Next Steps**

Started the Null Hypothesis and alternative hypothesis followed by step **b**, Sample (Mean): 68.74, Population: 70, Standard Deviation: 9.76 and Sample Size: 50, after the calculations, a result of t-statistic value **-0.92,** p-value **0.36**, with **49** degrees of freedom, it was obtained, in this case, if the p-value is greater than alpha-value we reject the hypothesis if the p-value is smaller than alpha-value then we fail to reject the hypothesis, in this case, p-value 0.36 <= alpha 0.05 = False, so we fail to reject the Null Hypothesis.

Next, it was plotted a histogram graphic to illustrate the findings and the Alternative Hypothesis with a One-Tailed Test, after the calculations, a result of t-statistic **-0.92,** and a p-value of **0.82**, with **49** degrees of freedom, was obtained, confirming that this is a right-tailed test.

# **Task 1 – Exploring the data**

After completing the tasks, an exploratory analysis was carried out on the dataset. It was created a new feature named “performance” with 1 as a value for students who got more than 70 per cent in the exam score and 0 otherwise, checked with “value\_counts” in this case, 26 did not get the score equal or greater than 70 and 24 students achieved this result.

Also plotted in a bar plot graphic, next, getting the percentage from the total student performance, in this case, 26 or 52% of the students did not get equal or more than 70% in the grades and 24 or 48% of the students got it, and next plotting using a pie graphic.

Next, created a new feature named "classification" to add a classification of the grades achieved, for this case the Classifications are:

* 0 – 34% Unsatisfactory (Fail).
* 35 – 39% Limited (Fail (pass by compensation, where applicable)).
* 40 – 49% Acceptable (Pass).
* 50 – 59% Good (Lower Second Class Honours or Merit One).
* 60 – 69% Very Good (Upper Second Class Honours or Merit Two).
* 70 – 79% Excellent (First Class Honours or Distinction).
* 80 – 89% Outstanding (First Class Honours or Distinction).
* 90 – 100% Exceptional (First Class Honours or Distinction).

Next, created a new feature named "grade\_band\_standard" to add a Grade Band and Standard of the grades achieved, for this case, the Grade Band and Standard are:

* 0 – 34% Unsatisfactory (Fail).
* 35 – 39% Limited (Fail (pass by compensation, where applicable)).
* 40 – 49% Acceptable (Pass).
* 50 – 59% Good (Lower Second Class Honours or Merit One).
* 60 – 69% Very Good (Upper Second Class Honours or Merit Two).
* 70 – 79% Excellent (First Class Honours or Distinction).
* 80 – 89% Outstanding (First Class Honours or Distinction).
* 90 – 100% Exceptional (First Class Honours or Distinction).

Classification results below also plotted using a count plot:

* 17 students: Very Good.
* 15 students: Excellent.
* 9 students: Outstanding.
* 7 students: Good.
* 2 students: Acceptable.

Grade Band Standard results below, also plotted using a count plot:

* 24 students: First Class Honours or Distinction.
* 17 students: Upper Second Class Honours or Merit Two.
* 7 students: Lower Second Class Honours or Merit One.
* 2 students: Pass.

Next, it was encoded the feature values, got a correlation matrix table and a heat map graphic, also used a for loop to get a dist plot and kde plot of each feature with Skew, mean, Median and Standard Deviation values and marks in their respective graphic and next used a pair plot graphic.

# **Machine Learning - DecisionTreeClassifier**

Used a Decision Tree Classifier, to predict the performance of the students, and split the data into **X** and **y**, it was chosen 30% of the data to test and 70% of the training, with the parameters: criterion = 'gini', max\_depth = 5 and random\_state = 43, got the results: Accuracy: 1.00, Precision: 1.00 and Recall: 1.00. Also used a Confusion Matrix and classification report for more details and used a plot tree to get a tree with max depth 5 as value, but in this case got jut with 1.

# **TASK 2 – Exploratory Data Analysis**

Imported the Diamonds dataset and got the “Head” to understand how the dataset looks like with the first 5 rows and check the name of the features that in this case, are “carat, cut, color, clarity, depth, table, price, x, y and z”, “shape” to get the size of the dataset, 53940 observations (Rows) and 10 feature(s) (Column(s)), “info” to get a concise summary in this data set have 7 numerical features and 3 object, “describe” to get statistical information of the numerical features.

Next plotted a boxplot of all numerical features, looks like the feature price has some outliers, and next a violin plot of the feature “price”, to check the distribution and the box plot individually itself.

Checked, duplicate observations, in this case, 146 observations were found and dropped, next, used a for loop to get a dist plot and kde plot of each numerical feature with Skew, Mean, Median and Standard Deviation values and marks in their respective graphic, also created a count plot for each categorical feature with a count of their respective values and next used a pair plot graphic.

# **TASK 2 – Next Steps**

Created a new binary feature called “colourless” which records 1 in rows with colourless diamonds and 0 otherwise, the Diamonds colours variable from "D" to "F", where "D" to "F" are considered colourless and “G” to “J” are not considered colourless. Next, Started the Null Hypothesis to determine whether there is any association between the clarity of a diamond and whether it is colourless or not followed by step **b**.

First, create a crosstab with the features “Clarity” and “Color” to get the proportions to understand better, also get a marginal proportion of colours and clarity for each value 0 and 1 from the table.

Next used the chi2\_contingency to calculate the expected contingency table, after the calculations, a result of chi2 value **2038.45,** p-value **0.0**, with **42** degrees of freedom, was obtained, in this case, if the p-value is smaller than the alpha-value we reject the hypothesis if the p-value is greater than alpha-value then we fail to reject the hypothesis, in this case, p-value **0.0** <= alpha **0.01** = **True**, so we fail to reject the Null Hypothesis, there is an association between the clarity of a diamond and colours.

Also create a table using the method Pearson, to understand more about the correlation between the features and plotted in a heatmap graphic.

Next, checking the correlation between "Clarity" and "Color" in a "crosstab" and using chi2\_contingency to get the answer, after the calculations, a result of chi2 **2038.45,** and a p-value of **0.0**, with **42** degrees of freedom, was obtained, where the p-value is **0.0** and hence we reject the null Hypothesis with **42** degrees of freedom.

Next, checking the association between the clarity of a diamond with colourless or not with a “crosstab” and using chi2\_contingency to get the answer, after the calculations, a result of chi2 **483.42,** and a p-value of **2.93**, with **7** degrees of freedom, was obtained, where the p-value is **2.93** and hence we reject the null Hypothesis with **7** degrees of freedom.

In this case, p-value **2.93** < alpha **0.01** = **True**, so we fail to reject the Null Hypothesis, there is an association between the clarity of a diamond with colourless.

Next, testing to check the values individual first "colourless\_yes" and next "colourless\_no", first colourless yes equals 1 in a "crosstab" and using chi2\_contingency to get the answer, after the calculations, a result of chi2 **0.0,** and a p-value of **1.0**, with **0** degrees of freedom, was obtained, where the p-value is **1.0** and hence we accept the null Hypothesis with **0** degrees of freedom.

Next, colourless no equals 0 in a "crosstab" and using chi2\_contingency to get the answer, after the calculations, a result of chi2 **0.0,** and a p-value of **1.0**, with **0** degrees of freedom, was obtained, where the p-value is **1.0** and hence we accept the null Hypothesis with **0** degrees of freedom.

Next created a copy from the dataset to a new one to do a test, encoded the object features and tested using a Pearson correlation, after the calculations, a result of statistic **-0.028,** and a p-value of **8.265**, was obtained, where the p-value is **8.265** and hence we reject the null Hypothesis with **-0.028** personr statistic.

In this case, p-value **8.265** < alpha **0.01** = **True**, so we fail to reject the Null Hypothesis, there is an association between the features.

As a step **c** it was plotted appropriate graphics about the finds before, First, a bar plot with “Clarity” and “Color” to get the correlations in the bars graphic, Next a Contingency table of “Clarity” and “Color” using a "heatmap". Next, a bar plot with “Clarity” and “Colourless” to get the correlations in the bars graphic and a Contingency table of Clarity and Colourless using a "heatmap".

As a step **d** to interpret 90% of confidence intervals for both the mean price of colourless diamonds and the mean price of non-colourless diamonds, it was got the follow results: True Colourless Mean Price: 3340.687612759587 and Non Colourless Mean Price: 4489.314421655913.

For 90% of confidence intervals for the mean price of colourless diamonds: 3304.720900142728, 3376.6543253764457 and 90% of confidence intervals for the mean price of non-colourless diamonds: 4446.830351975592, 4531.798491336234.

# **Task 2 – Exploring the data**

After completing the tasks, an exploratory analysis was carried out on the dataset. It was created a violin plot to get the distribution of cut in colors, also checked the “Info” and “Isnull” to see if have null values, next got the “Unique” and “Value Counts” of cut feature, and replaced from 0 to 4 values and encoded.

# **LDA - (Linear discriminant analysis)**

Used an LDA, to predict if a Diamond is 'Ideal', 'Premium', 'Good', 'Very Good' or 'Fair' from the cut feature in the dataset, and split the data into **X** and **y**, after fit in the LinearDiscriminantAnalysis model, got new 4 specific features with a strong signal from the 11 as before, plotted in a 3D graphic.

Next used a Decision Tree Classifier, to predict the performance of the new features, and split the data into **X** and **y**, it was chosen 30% of the data to test and 70% of the training with a random state of 43, with the parameters: basics, got the results: Accuracy: **0.60**. Also used a Confusion Matrix and classification report for more details of Recall, f1-score and support and plotted a heatmap confusion matrix.

Next, got the 4 features in a Data Frame, the index of the original dataset was reset, once deleted duplicated values were, the index got without fewer numbers less, created a new feature as a Target from the original dataset, using the cut as it, Split **X** and **y** once again, and split train and text with 30% of test random state 43.

Tested next in different algorithms and got the following results, where the first values are the accuracy and the standard deviation:

* LR - LogisticRegression: 0.602629 (0.004594).
* LDA - LinearDiscriminantAnalysis: 0.610251 (0.003833).
* KNN - KNeighborsClassifier: 0.677573 (0.008197).
* CART - DecisionTreeClassifier: 0.599920 (0.006859).
* NB - GaussianNB: 0.601142 (0.009030).
* SVM - SVC: 0.708379 (0.006077).
* RFC - RandomForestClassifier: 0.700890 (0.006173).
* ANN - MLPClassifier: 0.707024 (0.005153).

Also plotted in a boxplot of Algorithm Comparison, after tests chosen the SVC Machine Learning algorithm due to his performance and distribution of comparison, tested and got values s Accuracy **0.71**, Classification Report with Precision, Recall, f1-score and support of it the values, a Confusion Matrix and plotted in a heatmap.

# **TASK 3 – Exploratory Data Analysis (Steps A, b, c)**

Imported the PlantGrowth dataset from the “pydataset” library and got the “Head” to understand how the dataset looks like with the first 5 rows and check the name of the features that in this case, are “weight and group”, “shape” to get the size of the dataset, 30 observations (Rows) and 2 feature(s) (Column(s)), “info” to get a concise summary in this data set have 2 numerical features, “describe” to get statistical information of the numerical feature weight, also check the “Unique” and “Value Counts” of group feature.

Next, checked a Distribution of Weight in a dist plot graphic with Skew, Mean, and Median marks, typed some commands to get statistical values with the results: Length: 30.00, Mean: 5.07, Standard Deviation: 0.70, Min: 3.59, Percentile 25%: 4.55, Percentile 50%: 5.15, Percentile 75%: 5.53, Max: 6.31, Sum: 152.19, Median: 5.15, Mode(s): 4.17, Range: 2.72, Variance: 0.48, Interquartile Range (IQR): 0.98 and Coefficient of Variation (CV): 13.59.

Plotted a count plot of group feature, next got a Statistical Summary with Mean, Median, Standard Deviation, Min, Max and Count for each value in the group feature, also plotted in a box plot to get better, and looks like the trt1 has some outliers values, also used a cat plot, point plot and cat plot with different paraments of each value group, next plotted in a histogram graphic to compare the values of Control (ctrl), Treatment 1 (trt1) and Treatment 2 (trt2).

As a step **b** it was conducted an appropriate hypothesis test to see if there is evidence of a difference between the three means (that is, the control and the two treatments), first create a "crosstab" and using chi2\_contingency to get the answer, after the calculations, a result of chi2 **57.00**, and a p-value of **0.43**, with **56** degrees of freedom, was obtained, where the p-value is **0.43** and hence we accept the null Hypothesis with **56** degrees of freedom.

Next, get the mean of each value of the group feature, in this case, ctrl\_mean 5.032, trt1\_mean 4.661 and trt2\_mean 5.526. Next using ANOVA to the null hypothesis, after the calculations, a result of statistic **4.84**, p-value **0.015**, where p-value 0.015 < alpha 0.05 equals True, so there is evidence of a difference between the three means (that is, the control and the two treatments).

Next, performed a T-test in “weight ~ group” got the results: params.Intercept equals 5.031, params["group[T.trt1]"] -0.370 and params["group[T.trt2]"] 0.493 and printed the respective OLS Regression Results summary.

Next, a regression graphic it was created, after created an ANOVA table and checked the Normal distribution of residuals, with Sample Quantiles as y and Theoretical Quantiles as x, with the following ShapiroResult: statistic 4.846, p-value 0.015.

Next, created an ols model with "weight ~ group" and got params.Intercept 5.031, and printed the respective OLS Regression Results summary individually, first to ctrl values in the group feature, next a normal Q-Q plot to test this assumption, with the following ShapiroResult: statistic 0.95, p-value 0.74.

Next, checked individually to trt1 values in the group feature, next a normal Q-Q plot to test this assumption, with the following ShapiroResult: statistic 0.93, p-value 0.45 and next, checked individually to trt2 values in the group feature, next a normal Q-Q plot to test this assumption, with the following ShapiroResult: statistic 0.94, p-value 0.56.

As step **c** checked if there is evidence of a difference between the three means, in this case, it was used the “pairwise\_tukeyhsd” to use Multiple Comparisons of Means with Tukey HSD and got FWER **0.05**, next in the summary got as result ctrl and trt1 reject: False, ctrl and trt2 reject: False and trt1 and trt2 reject: True, so checking this summary the difference may be between trt1 and trt2.

# **Task 3 – Exploring the data**

After completing the tasks, an exploratory analysis was carried out on the dataset. It was created to replace the group values ctrl, trt1 and trt2 from 0 to 2 values.

# **MACHINE LEARNING – Linear Regression**

Used Linear Regression, to predict values from the group feature in the dataset, and split the data into **X** and **y**, with test size as 20% and random state 0, next created an get\_cv\_scores function to get CV Mean and STD, after fit the LinearRegression in this case got CV Mean: -0.034 and STD: 0.209, intercept\_ -0.996 and coef\_ 0.37, used the predict in X\_test and plotted in a Scatter plot with Data Points and Linear Regression Line, next created a new df Data frame with the Actual and Predicted values as features, and got the followed values: Mean Absolute Error: 0.79, Mean Squared Error: 0.72 and Root Mean Squared Error: 0.84.

# **MACHINE LEARNING – DECISIONTREECLASSIFIER**

Used a Decision Tree Classifier, to predict ctrl, trt1 and trt2, and split the data into X and y, it was chosen 30% of the data to test and 70% of the training, random state 43, with the parameters: criterion = 'gini', max\_depth = 5 and random\_state = 43, got the results: Accuracy: **0.78**, next plot a tree visualiser, also used a Confusion Matrix and Classification Report for more details in Precision, Recall, f1-score and support and plotted a heatmap.

# **MACHINE LEARNING – RANDOM FOREST**

Used a Random Forest Classifier, to predict ctrl, trt1 and trt2, and split the data into X and y, it was chosen 30% of the data to test and 70% of the training, random state 43, with the parameters: n\_estimators 100, got the results: Accuracy: **0.78**, next plot a tree visualiser, also used a Confusion Matrix and Classification Report for more details in Precision, Recall, f1-score and support and plotted a heatmap.

# **MACHINE LEARNING – Logistic regression**

Used a Logistic Regression, to predict ctrl, trt1 and trt2, and split the data into X and y, it was chosen 30% of the data to test and 70% of the training, random state 43, with the parameters: random state 43, got the results: Accuracy: **0.56**, next used a Confusion Matrix and Classification Report for more details in Precision, Recall, f1-score and support and plotted a heatmap, also plotted a reg plot of the results, and left commented on a ROCAUC visualiser of the previous result, opted not to keep once change all the next Task 4 graphics in the yellowbrick pattern.

After testing the Machine Learning algorithms, the Decision Tree Classifier and Random Forest Classifier performed the best results of it.

# **TASK 4 – Exploratory Data Analysis (Steps A, b, c)**

Imported the Tree dataset from the “pydataset” library and got the “Head” to understand how the dataset looks like with the first 5 rows and check the name of the features that in this case, are “Girth, Height and Volume”, plotted a graphic to see how these values are, “shape” to get the size of the dataset, 31 observations (Rows) and 3 feature(s) (Column(s)).

As a step **a** checking the correlation analysis between all numerical variables, first checking the column name, next getting a table with correlation using the command “corr”, next using “pearsonr”, first between the features “Girth and Height” PearsonRResult: statistic 0.51 and p-value 0.0027, in this case the p-value is 0.0027 and hence we reject the null Hypothesis with 0.51 personr statistic.

Nex between the features “Girth and Volume” PearsonRResult: statistic 0.97 and p-value 8.64, in this case, the p-value is 8.64 and hence we reject the null Hypothesis with 0.96 personr statistic.

Nex between the features “Height and Volume” PearsonRResult: statistic 0.59 and p-value 0.00037, in this case, the p-value is 0.00037 and hence we reject the null Hypothesis with 0.59 personr statistic.

As a step **b** to perform it, a regression analysis was created to decide which of the three possible models to recommend using, first created an ols model with "Volume ~ Girth" and plotted the respective OLS Regression Results summary individually, first with the "Volume" feature and "Girth", next a normal Q-Q plot regression to test this assumption, with the following ShapiroResult: statistic 0.83, p-value 0.00027.

Next an ols model with "Volume ~ Height" it was created and plotted the respective OLS Regression Results summary, also, a normal Q-Q plot regression to test this assumption was plotted, with the following ShapiroResult: statistic 0.86, p-value 0.0013.

Next an ols model with "Volume ~ Girth and Height" it was created and plotted the respective OLS Regression Results summary, also, a normal Q-Q plot regression to test this assumption was plotted, with the following ShapiroResult: statistic 0.98, p-value 0.95.

The results are kind of similar, but after comparing the regression analysis of each of them, and testing the three possible models, the best to recommend would be the last one, the ols model with Volume ~ Girth and Height once the Data Points and Linear Regression Line are most of the time near to each other, and also the values got ShapiroResult: statistic 0.98, p-value 0.95, shows a better performance.

# **Task 4 – Exploring the data**

After completing the tasks, an exploratory analysis was carried out on the dataset, check the “Head” next plot the feature values in a plot, and looks like the Volume feature has some outliers, plotted a distribution of the values between Girth and Height, next Girth and Volume, and Height and Volume, and the only one that seems a better distribution in the eyes of a linear line is Girth and Volume.

# **MACHINE LEARNING – Linear Regression (girth and volume)**

After choosing the feature Girth and Volume, used Linear Regression, to predict values from the group feature in the dataset, and split the data into X and y, with test size as 20% and random state 43, next created an get\_cv\_scores function to get CV Mean and STD, after fit the LinearRegression in this case got CV Mean: 0.42 and STD: 0.62, intercept\_ -36.68 and coef\_ 5.023, used the predict in X\_test and plotted in a Scatter plot with Data Points and Linear Regression Line, next created a new df Data frame with the Actual and Predicted values as features, and got the followed values: Mean Absolute Error: 2.54, Mean Squared Error: 9.51 and Root Mean Squared Error: 3.84.

# **MACHINE LEARNING – Linear Regression (Height and volume)**

But also tested with the feature Height and Volume, used Linear Regression, to predict values from the group feature in the dataset, and split the data into X and y, with test size as 20% and random state 43, next created an get\_cv\_scores function to get CV Mean and STD, after fit the LinearRegression in this case got CV Mean: -8.27 and STD: 18.72, intercept\_ -86.54 and coef\_ 1.54, used the predict in X\_test and plotted in a Scatter plot with Data Points and Linear Regression Line, next created a new df Data frame with the Actual and Predicted values as features, and got the followed values: Mean Absolute Error: 9.28, Mean Squared Error: 121.37 and Root Mean Squared Error: 11.01.

# **MACHINE LEARNING – Linear Regression (Girth and height)**

But also tested with the feature Girth and Height, used Linear Regression, to predict values from the group feature in the dataset, and split the data into X and y, with test size as 20% and random state 43, next created an get\_cv\_scores function to get CV Mean and STD, after fit the LinearRegression in this case got CV Mean: -6.087 and STD: 9.76, intercept\_ 61.80 and coef\_ 1.06, used the predict in X\_test and plotted in a Scatter plot with Data Points and Linear Regression Line, next created a new df Data frame with the Actual and Predicted values as features, and got the followed values: Mean Absolute Error: 4.09, Mean Squared Error: 22.66 and Root Mean Squared Error: 4.76.

As checked before, columns "Girth" and "Volume" have the best performance as a linear Regression.

# **MACHINE LEARNING – KNeighborsregressor**

Next used K Neighbors Regressor with the feature Girth and Height, to predict the nearest neighbour from the values of the feature in the dataset, and split the data into X and y, with a test size of 20% and random state of 43, chosen after tests 3 neighbors to the model, next getting the Test set predictions, Test set R^2: **0.96** and plotting three different graphics with 1 neighbor(s) as a test, next 3 neighbor(s) and 9 neighbor(s).

# **Conclusion**

All tasks are concluded with their respective required steps, task 1 was created and commented with statistical summaries of the dataset, and it was made a Hypothesis Testing based on the teachers concerned, after calculating the test statistics with Null and Alternative hypotheses it was as a result that we fail to reject her concerned as a Null Hypothesis, next plotted the findings and test once again with One-Tailed Test considering now a right-tailed test. Created also an exploration of the data with feature creation, more graphics and tests and used this new data set in a Decision Tree Classifier Machine Learning to predict the performance of the students.

In task 2 created a new binary feature with the colourless values where the Diamonds from the colours “D” to “F” are considered colourless, and diamonds from colour “G” to “J” are not considered colourless, it was performed appropriately hypotheses to test and determined if there any association between the features using “crosstab”, “chi2\_contingency” and “pearsonr”, plotted graphics to illustrate and also test and interpreted 90% of confidence intervals between the mean of the colourless feature, also created an exploration of the data with Feature Extraction using LDA (Linear Discriminant Analysis) testing in the Machine Learning: LogisticRegression, LinearDiscriminantAnalysis, KNeighborsClassifier, DecisionTreeClassifier, GaussianNB, SVC, RandomForestClassifier and MLPClassifier .

In task 3 was created and commented on with statistical summaries of the dataset, and it was conducted an appropriate hypothesis test to see if there is evidence of a difference between the three means using “crosstab”, “chi2\_contingency”, “ANOVA”, “OLS Regression Results” and more, also checked if there is evidence of a difference between the three means using “pairwise\_tukeyhsd”, created also an exploration of the data with Linear Regression, Decision Tree, Random Forest and Logist Regression Machine Learning algorithms.

Finally, in task 4, it was per performed a correlation analysis between all numerical variables with comments and hypothesis tests using “pearsonr”, also performed a regression analysis to decide which of the three possible models you would recommend using, created also, an exploration of the data with Linear Regression Machine Learning to Girth and Volume features, Height and Volume features, Girth and Height features, where "Girth" and "Volume" have the best performance as a linear Regression, once the Data Points and Linear Regression Line are most of the time near to each other, and also the values got from ShapiroResult test and Linear Regression Machine Learning algorithms and tested also with KNeighborsRegressor.

# **References**

GeeksforGeeks (2021). Using pandas crosstab to create a bar plot. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/using-pandas-crosstab-to-create-a-bar-plot/ [Accessed 16 May 2024].

Grammarly (2009). Grammarly. [online] Grammarly.com. Available at: https://app.grammarly.com/.

guest\_blog (2020). Introduction to ANOVA for Statistics and Data Science. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/06/introduction-anova-statistics-data-science-covid-python/ [Accessed 17 May 2024].

Hugo Bowne-Anderson (2022). if…elif…else in Python Tutorial. [online] www.datacamp.com. Available at: https://www.datacamp.com/tutorial/elif-statements-python [Accessed 14 May 2024].

iamaziz (2015). PyDataset/examples/basic-usage.ipynb at master · iamaziz/PyDataset. [online] GitHub. Available at: https://github.com/iamaziz/PyDataset/blob/master/examples/basic-usage.ipynb [Accessed 1 May 2024].

Ibezim, C. (2023). Using Contingency Tables to Find Associations Between Categorical Variables in Python. [online] Medium. Available at: https://medium.com/@ibezimchike/using-contingency-tables-to-find-associations-between-categorical-variables-in-python-2a2a4f14082d [Accessed 16 May 2024].

JulienAkpalu (2024). PlantGrowth-DataAnalysis/PlantGrowth Dataset Analysis Summary.md at main · JulienAkpalu/PlantGrowth-DataAnalysis. [online] GitHub. Available at: https://github.com/JulienAkpalu/PlantGrowth-DataAnalysis/blob/main/PlantGrowth%20Dataset%20Analysis%20Summary.md [Accessed 17 May 2024].

Macedo, C. (2024). CristhianMacedo2024104/Statistical-Techniques-CA2. [online] GitHub. Available at: https://github.com/CristhianMacedo2024104/Statistical-Techniques-CA2 [Accessed 2024].

Mahadevan, M. (2022). Step-by-Step Exploratory Data Analysis (EDA) using Python -. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/ [Accessed 14 May 2024].

mlexplained (2023). How To Calculate Correlation Among Categorical Variables? [online] ML EXPLAINED. Available at: https://mlexplained.blog/2023/01/23/how-to-calculate-correlation-among-categorical-variables/ [Accessed 16 May 2024].

Navlani, A. (2019). Python Logistic Regression Tutorial with Sklearn & Scikit. [online] www.datacamp.com. Available at: https://www.datacamp.com/tutorial/understanding-logistic-regression-python [Accessed 18 May 2024].

Polat, E.I. (2022). Hypothesis Testing with Python: Step by step hands-on tutorial with practical examples. [online] Medium. Available at: https://towardsdatascience.com/hypothesis-testing-with-python-step-by-step-hands-on-tutorial-with-practical-examples-e805975ea96e [Accessed 17 May 2024].

Scavetta, R. (n.d.). Chapter 2 Getting Started: Plant Growth Case Study | Data Analysis with Python. [online] www.ylz.ncx.mybluehost.me. Available at: https://www.ylz.ncx.mybluehost.me/scavetta.academy/misk/7\_Misk\_DSI\_Python/getting-started-plant-growth-case-study.html [Accessed 17 May 2024].

stackoverflow (2018). Correlation among multiple categorical variables. [online] Stack Overflow. Available at: https://stackoverflow.com/questions/48035381/correlation-among-multiple-categorical-variables [Accessed 16 May 2024].

Stephen Roddewig (2022). Using If-Else Statements in Pandas: A Practical Guide [+ Examples]. [online] blog.hubspot.com. Available at: https://blog.hubspot.com/website/pandas-if-else [Accessed 14 May 2024].

The Pandas Development Team (2024). pandas.core.groupby.DataFrameGroupBy.aggregate — pandas 2.2.2 documentation. [online] pandas.pydata.org. Available at: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.core.groupby.DataFrameGroupBy.aggregate.html [Accessed 17 May 2024].

The SciPy community (2014). scipy.stats.f\_oneway — SciPy v1.3.1 Reference Guide. [online] Scipy.org. Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f\_oneway.html [Accessed 16 May 2024].

The SciPy community (2019). scipy.stats.pearsonr — SciPy v1.3.2 Reference Guide. [online] Scipy.org. Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html [Accessed 16 May 2024].

The SciPy community (2024a). pandas.factorize — pandas 2.2.2 documentation. [online] pandas.pydata.org. Available at: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.factorize.html [Accessed 16 May 2024].

The SciPy community (2024b). scipy.stats.chi2\_contingency — SciPy v1.8.0 Manual. [online] docs.scipy.org. Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2\_contingency.html [Accessed 16 May 2024].

The SciPy community (2024c). scipy.stats.contingency.crosstab — SciPy v1.12.0 Manual. [online] docs.scipy.org. Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.contingency.crosstab.html [Accessed 16 May 2024].

The SciPy community (2024d). scipy.stats.ttest\_1samp — SciPy v1.7.1 Manual. [online] docs.scipy.org. Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest\_1samp.html [Accessed 14 May 2024].

Vidhi Chugh (2023). An Introduction to Python T-Tests. [online] www.datacamp.com. Available at: https://www.datacamp.com/tutorial/an-introduction-to-python-t-tests [Accessed 17 May 2024].

Zach (2020). How to Calculate Confidence Intervals in Python. [online] Statology. Available at: https://www.statology.org/confidence-intervals-python/ [Accessed 17 May 2024].

Zach (2022). How to Rotate Axis Labels in Seaborn Plots. [online] Statology. Available at: https://www.statology.org/seaborn-rotate-axis-labels/ [Accessed 14 May 2024].