**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| Module Title(s): | Programming for DA  Statistics for Data Analytics  Machine Learning for Data Analysis  Data Preparation & Visualisation |
| Assessment Title: | MSC\_DA\_CA2 |
| Lecturer(s) Names: | Sam Weiss  John O’Sullivan  Muhammad Iqbal  David McQuaid |
| Student Full Name: | Clelia Caetano |
| Student Number: | 2023060 |
| Assessment Due Date: | 26/05/2023 |
| Date of Submission: | 29/05/2023 |
| Word Count: | 3198 |

**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 another source. I declare it to be my work and that all material from third parties has been appropriately referenced. I further confirm that this work has yet to be submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Author: Clelia Caetano

E-mail: [2023060@student.cct.ie](mailto:2023060@student.cct.ie)

Student ID:2023060

GitHub: [CleliaCaetano/CA2\_Construction\_local (github.com)](https://github.com/CleliaCaetano/CA2_Construction_local)

Constructions:

Workplace Accidents

CCT College Dublin

May (2023)

Table of Contents

[List of Tables: 4](#_Toc136293808)

[Abstract 6](#_Toc136293809)

[1. Introduction 6](#_Toc136293810)

[2. Statistics for Data Analytics 7](#_Toc136293811)

[2.1 Collect Initial Data 7](#_Toc136293812)

[2.2 Describe Data 7](#_Toc136293813)

[2.2.1 Distribution (Kurtosis Test) 10](#_Toc136293814)

[2.2.2 Log Transformation 11](#_Toc136293815)

[2.2.3 Shapiro-Wilk test 12](#_Toc136293816)

[2.2.4 Probability density function (PDF) of the normal distribution 13](#_Toc136293817)

[2.3 The Kolmogorov-Smirnov (non-parametric) 14](#_Toc136293818)

[2.4 Kruskal-Wallis (non-parametric) test 15](#_Toc136293819)

[2.5 Mann-Whitney U (non-parametric) test 16](#_Toc136293820)

[3. Data Preparation & Visualization 19](#_Toc136293821)

[3.1 Dataset Description 19](#_Toc136293822)

[3.2 Select Data 19](#_Toc136293823)

[3.3 Clean Data 20](#_Toc136293824)

[3.4 Building Data 20](#_Toc136293825)

[3.5 Format Data 23](#_Toc136293826)

[4. Modeling 27](#_Toc136293827)

[4.1 Classification Models 27](#_Toc136293828)

[4.1.1 Decision Tree Model 27](#_Toc136293829)

[4.1.2 Random Forest 28](#_Toc136293830)

[4.1.3 K-Nearest Neighbor Classifier 29](#_Toc136293831)

[4.2.4 K-fold cross-validation 30](#_Toc136293832)

[5. Evaluation 31](#_Toc136293833)

[5.1 Evaluate Results 31](#_Toc136293834)

[5.1.1 Confusion matrix 31](#_Toc136293835)

[5.2 Determine Next Steps: 32](#_Toc136293836)

[5.2.1 Applying Machine Learning to Classify the Risk Level Range in the construction sector in the Union European Countries. 32](#_Toc136293837)

[6. Conclusion 35](#_Toc136293838)

[References 36](#_Toc136293839)

[Appendix 39](#_Toc136293840)

[1. Evidencing statistical calculations 39](#_Toc136293841)

# List of Tables:

[Figure 1: DataFrame head. 8](#_Toc136013186)

[Figure 2: Outliers 9](#_Toc136013187)

[Figure 3:Winsorized outliers. 9](#_Toc136013188)

[Figure 4: Statistical metrics. 10](#_Toc136013189)

[Figure 5: DataFrame Description. 11](#_Toc136013190)

[Figure 6: Distribution (Kurtosis Test). 12](#_Toc136013191)

[Figure 7: Log Transformation Applied. 13](#_Toc136013192)

[Figure 8: Shapiro-Wilk test results table. 13](#_Toc136013193)

[Figure 9.: Probability density function result (table). 14](#_Toc136013194)

[Figure 10: Probability density function visual result. 15](#_Toc136013195)

[Figure 11:The Kolmogorov-Smirnov (non-parametric) test result. 16](#_Toc136013196)

[Figure 12: Kruskal-Wallis (non-parametric) test plotting result. 17](#_Toc136013197)

[Figure 13: Determine the condition for the null hypothesis. 18](#_Toc136013198)

[Figure 14:Mann-Whitney U (non-parametric) test: Comparison with Ireland and other 29 European Union countries. 19](#_Toc136013199)

[Figure 15: Original DataFrame of Employeement size. 20](#_Toc136013200)

[Figure 16: Original DataFrame of Number non-fatal accidents. 20](#_Toc136013201)

[Figure 17: Original DataFrame of Number fatal accidents. 21](#_Toc136013202)

[Figure 18: DataFrame info( ). 21](#_Toc136013203)

[Figure 19: Missing values. 22](#_Toc136013204)

[Figure 20: TIME\_PERIOD convert as datetime64[ns]. 22](#_Toc136013205)

[Figure 21: New merged DataFrame. 23](#_Toc136013206)

[Figure 22: Multivariate analysis (Correlation between Features). 24](#_Toc136013207)

[Figure 23: Pairplot correlation variables. 25](#_Toc136013208)

[Figure 24: Choropleth 'Total accidents per country over the years'. 26](#_Toc136013209)

[Figure 25: Comparison of non-fatal accidents between the countries. 26](#_Toc136013210)

[Figure 26: Comparison of fatal accidents between the countries. 27](#_Toc136013211)

[Figure 27: DataFrame (new features). 27](#_Toc136013212)

[Figure 28: Last version DataFrame before Machine Learning application. 28](#_Toc136013213)

# Abstract

*This report is intended to summarize the results of the analyses carried out on the number of workers in the construction sector in the European Union countries and to account for the number of accidents that occurred between the years 2011 and 2020. The entire dataset was cleaned, processed and analysed statistically. After, classification between the risk levels in these countries was made using Machine Learning models and shown in a dashboard.*

*Finally, an analysis of the feelings of workers in this area was added.*

***Keywords: construction workplace accidents, risk level, European Union countries, classification.***

A picture containing text, line, screenshot, font

Description automatically generated

# 1. Introduction

The construction sector is a complex business as it encompasses several sales & purchase relationships. Contracts are signed in all the processes, whether with suppliers, buyers or hiring labour. All these relationships have a direct impact on progress and compliance with financial schedules and budgets.

After some research, it was found that there are not many studies on the risks of accidents to which workers in this sector are exposed. It is worth mentioning that this workforce has a high connection with the costs and schedules of work. Because if an employee suffers an accident, he will have to be absent from his duties, leading to delays, medical expenses, demotivation and even an overload of other workers.

According to HAS, “ In 2020, the NACE economic sector with the highest rate of work-related injuries leading to four or more days of absence from work was Construction (15.5 per 1,000 workers)”. (Authority, 2022).

Thus, this project merges all the information data collected and shows the results raised.

This case study aims to statistically analyse data from the dataset to compare numerical variables across countries and apply Machine Learning models to classify 30 European Union countries into a risk range.

A picture containing text, line, screenshot, font

Description automatically generated

# 2. Statistics for Data Analytics

## 2.1 Collect Initial Data

Initially, three different datasets from the European Union (Eurostat) website were collected. These are the Number of employees (SBS\_SC\_CON\_R2), the Number of non-fatal accidents (HSW\_N2\_01) and the Number of fatal accidents (HSW\_N2\_02). (Union, 2023).

These files have information of 30 countries of the European Union (AT: Austria, BE: Belgium, BG: Bulgaria, CY: Cyprus, CZ: Czech Republic, DE: Germany, DK: Denmark, EE: Estonia, EL: Greece, ES: Spain, FI: Finland, FR: France, HR: Croatia, HU: Hungary, IE: Ireland, IS: Iceland, IT: Italy, LT: Lithuania, LU: Luxembourg, LV: Latvia, MT: Malta, NL: Netherlands, NO: Norway, PL: Poland, PT: Portugal, RO: Romania, SE: Sweden, SI: Slovenia, SK: Slovakia and UK: United Kingdom) from the years 2011 to 2020.

## 2.2 Describe Data

After carrying out the pre-preparation and cleaning process, the three files were merged into a single dataset.

A picture containing text, screenshot, font, number

Description automatically generated

Figure : DataFrame head.

Which means:

* there is in a clear DataFrame a sample with 294 observations & 5 variables ('geo' is a categorical variable, and the other columns are numerical variables).
* the variables are considered multivariate analysis.
* the dataset has a continuous numerical probability, which the outcome can take any value within a certain interval.

With the Data Frame defined, it was time to observe the outliers and check their extreme values ​​and how they can interfere with the results of the analyses. It was verified that there were extremely high numbers above the maximum value in the three measured variables.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure : Outliers

It was also defined that these outliers were essential data for the project, so they were winsorized to reduce noise. But it was determined that this technique would be applied to reduce only 5% of the anomaly. Where the first variable reduced from 4M to 3M, the second variable from 1.5M to 1.2m and the third variable from 300 to 200. As it is displayed in the figure below.

A picture containing text, screenshot, diagram, parallel

Description automatically generated

Figure :Winsorized outliers.

Subsequently reducing anomalies and possible errors, the next point is to check the statistical metrics.A picture containing text, screenshot, font, number

Description automatically generated

Figure : Statistical metrics.

It is observed:

* the average (mean) indicates that the number of employees in the construction area of the European Union countries covered in this study is 705,826.96 and that the number of non-fatal accidents is 203,102.63 and fatal accidents 53.86.
* the standard deviation shows that 940,941.73 is the value of the number of employees, 387,533.18 is the number of non-fatal accidents and 68.05 is the value of the number of fatal accidents.
* Other measures as quartiles (25%, 50% and 75%) and maximum & minimum values are also indicating in the analyse and can be observed in the table.

All this information can also be observed in an interactive graph in Jupyter notebook version.

A picture containing text, diagram, plot, line

Description automatically generated

Figure : DataFrame Description.

## 2.2.1 Distribution (Kurtosis Test)

To detect whether the variables are independent or random, it is essential to test the normal distribution of the data set to check the statistical probability in the distribution.

Ideally, when returning from this test, the distribution should appear symmetrical to the mean value. Unlike the ideal, the bell of the curve in this data is positively sloped to the right, meaning that there is no normal distribution and that the variables are skewed to the right.

In the kurtosis test, p-values ​​and statistics were also printed, which returned as zero, reassuring that the data does not follow a normal distribution.

A picture containing text, diagram, screenshot, plot

Description automatically generated

Figure : Distribution (Kurtosis Test).

## 2.2.2 Log Transformation

Employing the log transformation to modify the scale to reduce the skewness and review the normal distribution.

For this, some rules were settled, where the value '+1' was applied to keep the '0' instead of turning it into a negative value in the variables Number of fatal accidents and Number of fatal accidents.

A picture containing diagram, text

Description automatically generated

Figure : Log Transformation Applied.

Visibly, the data still does not follow the normal distribution because where it should be a bell curve, it presents a wave shape.

## 2.2.3 Shapiro-Wilk test

The Shapiro-Wilk test was established to check the normality of the sample dataset and evaluate the log transformation, replying in two main statistical metrics: test statistics and p-value.



Figure : Shapiro-Wilk test results table.

Overall, the test statistic values returned very close to 1, indicating a good fit of the normal distribution of the variable tested. And the p-value is less than 0.05, suggesting that the log-transformed variables still do not follow a normal distribution. We can reject the null hypothesis.

## 2.2.4 Probability density function (PDF) of the normal distribution

While the Shapiro-Wilk test provides statistical metrics for the assumption of the distribution from a sample, the probability density function (PDF) shows the likely possibility distribution of continuous random variables. In this study, the PDF was applied to calculate the mean (mu\_hat) and standard deviation (sigma\_hat) of the log-transformed data for each column in log\_columns and it was used to compare how the fit value applied from the log transformation differ from the original data.



Figure .: Probability density function result (table).

The PDF test result demonstrates the mean and the standard deviation for each variable (Number of employees, Number of non-fatal accidents, Number of fatal accidents, Total accidents, percentage\_fatal, risk\_level). Then, it provides the statement ‘Reject’, based on the Shapiro-Wilk test, the log-transformed does not follow the normal distribution, because the curve is still asymmetric, indicating it is skewed. Also, the p-value is less than 0.05, which is evidence enough to reject the null hypothesis.

Finally, the Fit value indicates how many standard deviations is the mean from the minimum value of each variable. In other words, the Fit shows the normalization of the mean (Mu), by calculating: Fit Value = (Mu - Minimum Value) / (Maximum Value - Minimum Value) .

And a graph to illustrate the results.

A picture containing text, diagram, screenshot, plot

Description automatically generated

Figure : Probability density function visual result.

## 2.3 The Kolmogorov-Smirnov (non-parametric)

Confirming the statement that the data does not follow a normal distribution, another non-parametric test was alternatively used to deepen the analysis without the presumption of results due to equality of variances and normality.

The Kolmogorov-Smirnov (non-parametric) technique was tested to compare the normality of the tested variables in their original form and after applying the logarithmic transformation.



Figure :The Kolmogorov-Smirnov (non-parametric) test result.

It can be noted that some variables had a small improvement in their results after being transformed, where the static value was lower than 1, showing better fit. However, these changes were not significantly consistent to the point of transforming the p-value and consequently deviating the data format, remaining as non-normal distribution.

## 2.4 Kruskal-Wallis (non-parametric) test

Kruskal-Wallis is a non-parametric alternative of ANOVA test and has the objective to investigating the medians to find the differences of distribution shapes or central tendencies.

In regarding to this analysis, Kruskal-Wallis was utilised to compare the medians correlation between the risk level and percentages of types of accidents variables.

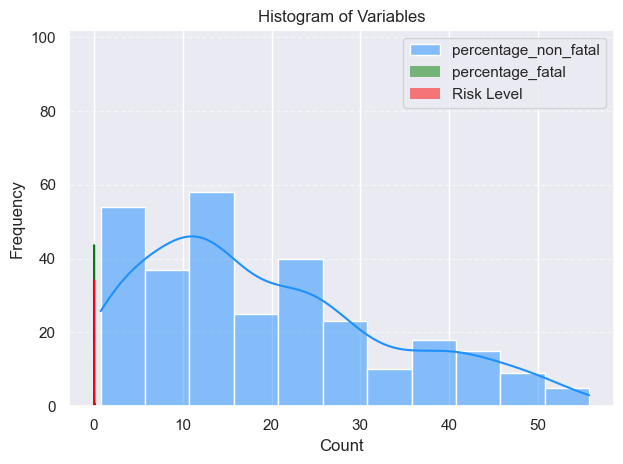


Figure : Kruskal-Wallis (non-parametric) test plotting result.

This graph presents that there is no similarity between the medians of the variables risk level and percentage types of accidents. Also, it was observed that the p-value is equal 0.00. That means, they are different in its statistical values.

## 2.5 Mann-Whitney U (non-parametric) test

To create a trend and logic in these analyses, another non-parametric test was also implemented that, alternatively to the t-test (parametric), investigates the relationship between two variables. The technique Mann-Whitney U tested the comparison between Ireland and the other 29 countries on the list to verify the distribution across these countries. For this, a single command was given, and a table was generated with the p-values ​​of the label’s Total accidents, Percentage Non-Fatal and Percentage Fatal.

It should be mentioned that a condition of p-values was implemented for the result to be rejected or not.

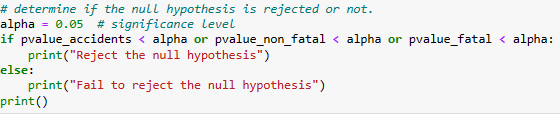


Figure : Determine the condition for the null hypothesis.

Although in the table below it can be seen that Ireland has some similarities with some countries in a few variables like Italy and Iceland for example, the test was rejected because the condition was that all tested p-values ​​that returned less than 0.05 'Reject the null hypothesis' would be printed.



Figure :Mann-Whitney U (non-parametric) test: Comparison with Ireland and other 29 European Union countries.

A picture containing text, line, screenshot, font

Description automatically generated

# 3. Data Preparation & Visualization

## 3.1 Dataset Description

At first, the idea was to find data from few countries to compare the metrics between them, but due to the variety of different languages ​​and also the lack of data with similar subjects, the objective became to look for files that integrated more than one country. The data required to construct this project was found after searching the Eurostat website. Both files come from the same authority and they are all open source. The copyright license can be verified in the link provided in the references. (Union, 1995).

## 3.2 Select Data

Three files, Number of employees, Number of non-fatal accidents and Number of fatal accidents, were selected to compose the analysis about the risk level of working in construction.

These row data are listed in next.

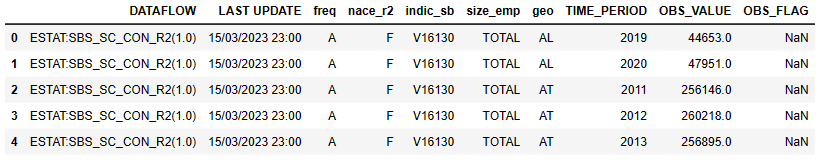


Figure : Original DataFrame of Employeement size.

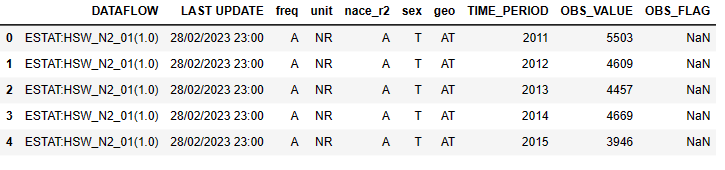


Figure : Original DataFrame of Number non-fatal accidents.

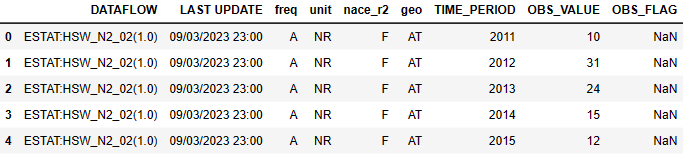


Figure : Original DataFrame of Number fatal accidents.

## 3.3 Clean Data

Datasets needed to be cleaned up before being merged. Features were renamed with the variable name so that they could be identified and all irrelevant and/or duplicate columns, such as searching code and year of last update were dropped. The rows that divided the information between construction sub-sectors were grouped. Some data that differed from one file to another, such as countries, were also dropped. This made the data frames have the same shape.

## 3.4 Building Data

Using the functions intersections to delimit the variable, grouped so that the values ​​are integrated and merged, the datasets were allocated in a data collection.

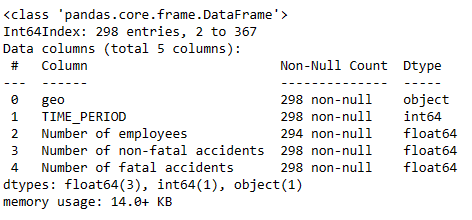


Figure : DataFrame info( ).

This dataframe underwent other cleaning techniques, so it was necessary to look for missing values ​​and address a possible solution for them.

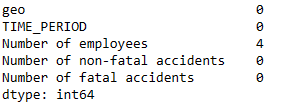


Figure : Missing values.

Due to the small number of missing values, they were dropped, as they would not cause major impacts on the analyses.

The column TIME\_PERIOD was converted to datetime64[ns] format to validate as an index.

A picture containing text, screenshot, font, number

Description automatically generated

Figure : TIME\_PERIOD convert as datetime64[ns].

Along with the application of all these techniques mentioned, a new dataframe was built.

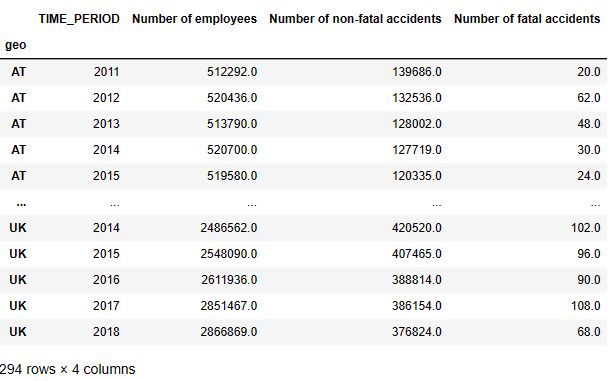


Figure : New merged DataFrame.

To continue the analyses, it is necessary to check the correlation between all the variables gathered within a data structure. Applying the correlation matrix implies understanding the dimension of independence or dependence among the variables and the existence of trends or patterns. A heatmap completes the analysis as a visual way to better recognize the relationships.

A picture containing text, screenshot, number, diagram

Description automatically generated

Figure : Multivariate analysis (Correlation between Features).

This figure shows that the TIME\_PERIOD resource has no correlation with any other resource, as it is an independent variable. However, the Number of employees, Number of non-fatal accidents & Number of fatal accidents have a correlation above 70%, this means that they are dependent variables.

Another visual format (pair plot) was designed to illustrate the correlation of these variables.

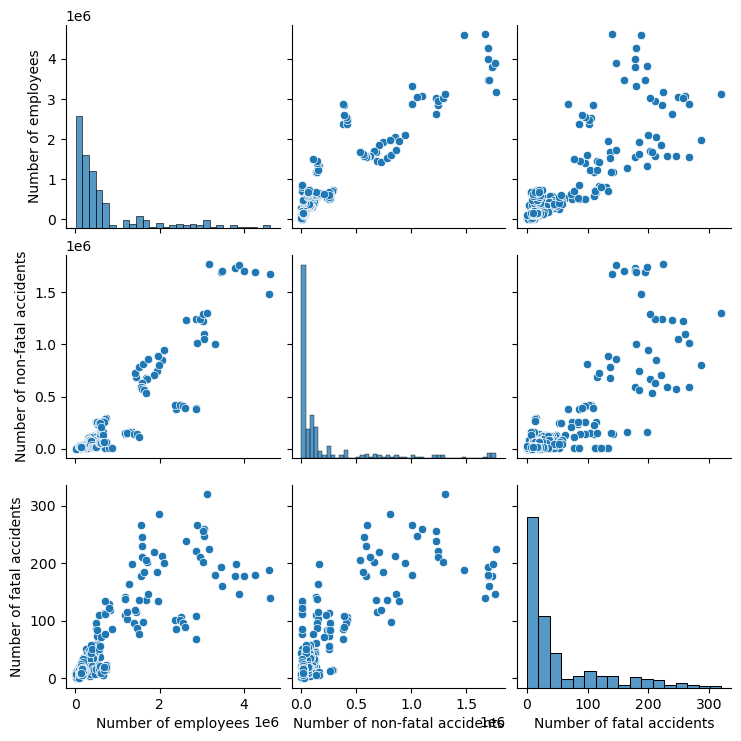


Figure : Pairplot correlation variables.

## 3.5 Format Data

Sequentially, the first analysis applications were carried out directly in the countries on the list and a comparison between them was raised. With a simple mathematical calculation (Total accidents = Number of non-fatal accidents + Number of fatal accidents) the Total number of Accidents was found and put into practice. It was verified through a choropleth of how this measure behaved over the years.

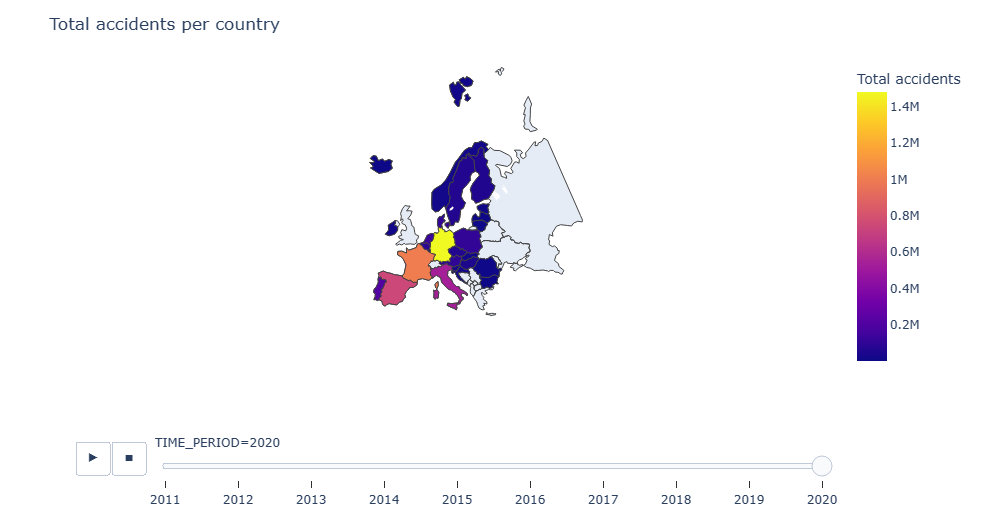


Figure : Choropleth 'Total accidents per country over the years'.

And then, an interactive graph comparing non-fatal and fatal accidents between countries was plotted.

A picture containing text, screenshot, plot, diagram

Description automatically generated

Figure : Comparison of non-fatal accidents between the countries.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure : Comparison of fatal accidents between the countries.

Note that Germany, Spain, France & Italy have the highest number of accidents on the list and Ireland is among the medium-value countries.

It should be mentioned that two new features were created to allocate accident percentages by type. And once again simple calculation was applied to find these values.

* percentage\_non\_fatal = Number of non-fatal accidents / Number of employees \* 100
* percentage\_fatal = Number of fatal accidents / Number of employees \* 100

A picture containing text, font, line, number

Description automatically generated

Figure : DataFrame (new features).

Later, it was necessary to use other functions to adjust the measures, so the function scaler = MinMaxScaler() was applied to scale between 0 and 1 the selected columns 'Total accidents', 'percentage\_non\_fatal', 'percentage\_fatal'. Another function was implemented to encode the label ‘geo’ into a numerical variable.

Other features were created to allocate risk level and risk range. Then, the mapping function was used to transform the risk range from categorical ('high', 'moderate', 'low') to a numerical variable.

This is the DataFrame format before implementing the Machine Learning classification models.

A picture containing text, menu, screenshot

Description automatically generated

Figure : Last version DataFrame before Machine Learning application.

A picture containing text, line, screenshot, font

Description automatically generated

# 4. Modeling

To answer the question of what is the risk range of accidents that each country of the European Union is inserted in the construction area, it was determined that for the application of the supervised models of Classification of Machine Learning in this study, they would be Decision Tree, Forest Random and K-Nearest Neighbor.

## 4.1 Classification Models

To build the classification learning algorithm model, first, the data was split between the training (70%) and test (30%) sets. This percentage division gave a balanced proportion between the sets and it can learn the patterns to classify the rows.

Then, the tests were applied within each model. As these are classification models, the alpha was not determined here, but the number of estimators, in this study '10' was established, to be used to standardize the fit value and thus making the prediction within this parameter. In this process, GridSearchCV was applied to perform on the grid and give the best hyperparameter to train the models in order to make them more accurate and result in a parameter for cross-validation.

As previously mentioned, the objective of this study is to classify countries between risk ranges of accidents and, for this purpose, supervised classification models were chosen because they try to teach the machine and train it for the known and possible returns.

## 4.1.1 Decision Tree Model

To carry out the decision tree test, the entropy criterion was determined as the best hyperparameter to measure the impurity and indicate the best trait to divide the nodes during the tree construction process. The maximum leaf depth also was settled at 4.

After indicating the parameter and applying it to the analysis, the test precision return value was 0.9775, which is an excellent score.

The decision tree graph shows the values chosen as the good sample and the measure of accuracy in the test set.

A picture containing text, font, receipt, screenshot

Description automatically generated

Figure : Decision Tree illustration test result.

## 4.1.2 Random Forest

As a hyperparameter of the Random Forest model, it was indicated that the best number of estimators was 7. Thus, the best precision was 0.9658, but the calculation of the precise score in the set test performed even better 0.9888.

Other analyses were also carried out to measure the importance of each feature within the set, and only percent\_non\_fatal had a relevant number of importance 0.9856, while percent\_fatal and Total accidents scored 0.0144 and 0.0000 respectively.

A picture containing text, screenshot, plot, diagram

Description automatically generated

Figure : Importance Features.

## 4.1.3 K-Nearest Neighbor Classifier

The K-Nearest Neighbor Classifier model uses the number of neighbors as a parameter, which in this study was to indicate that the best number would be 9. This resulted in a score of 0.9561, which indicates a very good percentage of test accuracy.

In order to measure the proportion of variance in dependent variables which can be explained by independent variables, the R-squared (2R) was also calculated and the accuracy of the test set was 0.977, which is a very high percentage.



Figure : K-Nearst Neighbor test & training accuracy.

## 4.2.4 K-fold cross-validation

The technique of mean score in K-fold 5 cross-validation folders was applied to cross-validate the results between the chosen models:

* Decision Tree Accuracy: 0.9829924021040327
* Random Forest Accuracy: 0.9558153126826416
* K-Nearest Neighbor Accuracy: 0.9319111630625365

According to the results, Decision Tree indicates better accuracy than other models and K-Nearest Neighbor was the least accurate among them.

A picture containing text, line, screenshot, font

Description automatically generated

# 5. Evaluation

## 5.1 Evaluate Results

## 5.1.1 Confusion matrix

A confusion matrix was applied to evaluate the accuracy and scores of the chosen tree classification model. As a result, it can be observed that the models had a small difference between their split sets.

Random Forest performing the best score with the accuracy on the training-set as 1.0 and on the test-set has 0.9888. While K-Nearest Neighbor performed slightly below the other models, with a training-set score of 0.9561, and on the test-set was 0.9775.

Random Forest also had the best accuracy record on the test set (0.9881), Recall on the test-set (0.9896) and F1 score on the test-set (0.9886). And, K-Nearest Neighbor consequently scored 0.977, 0.9792, 0.9773 and ranked third in performance. See the entire values in table and graph below:



Figure : Classification models evaluation (table).

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure : Classification models evaluation (graph).

## 5.2 Determine Next Steps:

## 5.2.1 Applying Machine Learning to Classify the Risk Level Range in the construction sector in the Union European Countries.

After training and testing the models and finding accurate results, the next step was to apply them to the data under study.

For this, each model was tested to classify countries within an accident risk range between low=1, moderate=2 and high=3, using the results of each prediction model grouped. Subsequently, a final classification was generated, which can be seen in the table below.



Figure : Classification of the countries (table).

Note that Ireland is ranging as moderate risk level of accidents in the construction sector. It can be verified in the table below a summary of the final classification result.

A picture containing text, screenshot, line, font

Description automatically generated

Figure : Final Classification result counted.

As a form to synthesize the results found and display them in a more visually attractive and understandable way, a bar graph panel was created to show the values ​​of each country that can be explored by Jupyter Notebook or through this link that is provided in there.

A picture containing text, screenshot, plot, diagram

Description automatically generated

Figure : Dashboard: Risk Classification of Countries.

In addition, a sentimental analysis was performed to validate how employees feel about working in the construction sector, and the clustering model was applied to evidence the result.

A picture containing screenshot, text, diagram, plot

Description automatically generated

Figure : Clustering the Sentiment analysis.

In other words, after categorizing the texts into feelings: Positive, Neutral and Negative, they were grouped according to their similarities and the groups or clusters to which they belong.

A picture containing text, line, screenshot, font

Description automatically generated

# 6. Conclusion

Although the number of accidents by itself does not have a determined role to indicate whether a country is dangerous to work in the construction area or not, this case study aimed to analyse the numbers and values ​​of each country in the European Union and placed in a classification so that it could have a ranking at the end.

Of course, in order to deepen the analysis and understand how the numbers impact each country differently, many issues need to be raised, such as work safety policies, as well as the personal protective equipment (PPE) used in each place. But, again, that was not the purpose of this study.

It is worth mentioning that some challenges already mentioned above were encountered, such as looking for data for each country separately and also posts and discussion forums about the coverage region to carry out the sentiment analysis and this part was covered with data from another region.

Finally, it can be concluded that the objective was reached, and the result was delivered in the form of analysis, graphs, dashboard and tables to illustrate it.

# References

Aandahl, E. M., 2022. *Parametric versus nonparametric tests..* [Online]   
Available at: https://ledidi.com/academy/parametric-versus-nonparametric-tests  
[Accessed 19 May 2023].

Authority, H. a. S., 2022. *hsa.ie.* [Online]   
Available at: https://www.hsa.ie/eng/topics/statistics/annual\_review\_of\_workplace\_injuries\_illnesses\_and\_fatalities\_20202021.pdf  
[Accessed 2 May 2023].

Foundation, P. S., 2021. *Python 3.10.1 documentation.* [Online]   
Available at: https://docs.python.org/3/library/index.html  
[Accessed 12 April 2023].

Frost, J., 2017. *Nonparametric Tests vs. Parametric Tests.* [Online]   
Available at: https://statisticsbyjim.com/hypothesis-testing/nonparametric-parametric-tests/  
[Accessed 19 May 2023].

Frost, J., 2022. *Skewed Distribution: Definition & Examples..* [Online]   
Available at: https://statisticsbyjim.com/basica/skewed-distribution/  
[Accessed 20 May 2023].

Grolemund, G. &. W. H., 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* 1st ed. Sebastopol, CA, USA.: O'Reilly Media, Inc. .

Guido, A. C. M. &. S., 2016. *Introduction to Machine Learning with Python.* 1st ed. Sebastopol, CA, USA.: O’Reilly Media, Inc..

Hartman, J., 2023. *Guru 99.* [Online]   
Available at: https://www.guru99.com/java-vs-python.html  
[Accessed 14 May 2023].

Lawrence S. Meyers, G. C. G. A. J. G., 2012. *Applied Multivariate Research: Design and Interpretation.* 2nd ed. California, USA: Sage Publications Inc..

Pedamkar, P., 2023. *EDUCBA.* [Online]   
Available at: https://www.educba.com/python-vs-javascript/  
[Accessed 14 May 2023].

Raschka, S. &. M. V., 2019. *Python Machine Learning.* 3rd ed. Birmingham, UK.: Packt Publishing.

Sullivan, M., 2018. *Fundamentals of statistics.* 5th ed. s.l.:Pearson Education, Inc..

Team, G. L., 2022. *Great Learning.* [Online]   
Available at: https://www.mygreatlearning.com/blog/open-source-python-libraries/  
[Accessed 12 May 2023].

Trevisan, V., 2022. *Towards Data Science - Comparing sample distributions with the Kolmogorov-Smirnov (KS) test..* [Online]   
Available at: https://towardsdatascience.com/comparing-sample-distributions-with-the-kolmogorov-smirnov-ks-test-a2292ad6fee5  
[Accessed 20 May 2023].

Union, E., 1995. *Eurostat.* [Online]   
Available at: https://ec.europa.eu/eurostat/web/main/about-us/policies/copyright  
[Accessed 3 May 2023].

Union, E., 2023. *Eurostat.* [Online]   
Available at: https://ec.europa.eu/eurostat/databrowser/view/SBS\_SC\_CON\_R2\_\_custom\_6155496/default/table?lang=en  
[Accessed 28 April 2023].

Union, E., 2023. *Eurostat.* [Online]   
Available at: https://ec.europa.eu/eurostat/databrowser/view/HSW\_N2\_01\_\_custom\_6154887/default/table?lang=en  
[Accessed 28 April 2023].

Union, E., 2023. *Eurostat.* [Online]   
Available at: https://ec.europa.eu/eurostat/databrowser/view/HSW\_N2\_02\_\_custom\_6154914/default/table?lang=en  
[Accessed 28 April 2023].

# Appendix

## 1. Evidencing statistical calculations





