**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: | 26/05/2023 |

**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 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

[Abstract 3](#_Toc136013074)

[List of Tables 6](#_Toc136013075)

[1. Introduction 7](#_Toc136013076)

[2. Statistics for Data Analytics 8](#_Toc136013077)

[2.1 Collect Initial Data 8](#_Toc136013078)

[2.2 Describe Data 8](#_Toc136013079)

[2.2.1 Distribution (Kurtosis Test) 11](#_Toc136013080)

[2.2.2 Log Transformation 12](#_Toc136013081)

[2.2.3 Shapiro-Wilk test 13](#_Toc136013082)

[2.2.4 Probability density function (PDF) of the normal distribution 14](#_Toc136013083)

[2.3 The Kolmogorov-Smirnov (non-parametric) 15](#_Toc136013084)

[2.4 Kruskal-Wallis (non-parametric) test 16](#_Toc136013085)

[2.5 Mann-Whitney U (non-parametric) test 17](#_Toc136013086)

[3. Data Preparation & Visualization 20](#_Toc136013087)

[3.1 Dataset Description 20](#_Toc136013088)

[3.2 Select Data 20](#_Toc136013089)

[3.3 Clean Data 21](#_Toc136013090)

[3.4 Building Data 21](#_Toc136013091)

[3.5 Format Data 25](#_Toc136013092)

[4. Modeling 29](#_Toc136013093)

[4.1 Select Modeling Technique 29](#_Toc136013094)

[4.2 Generate Test Design 29](#_Toc136013095)

[4.3 Build Model 29](#_Toc136013096)

[4.4 Assess Model 29](#_Toc136013097)

[Decision Tree Model 29](#_Toc136013098)

[Random Forest 29](#_Toc136013099)

[K-Nearest Neighbor Classifier 29](#_Toc136013100)

[K-fold cross-validation: 29](#_Toc136013101)

[5. Evaluation 29](#_Toc136013102)

[5.1 Evaluate Results 29](#_Toc136013103)

[5.2 Review Process 31](#_Toc136013104)

[5.3 Determine Next Steps 31](#_Toc136013105)

[6. Conclusion 32](#_Toc136013106)

[6.1 Plan Deployment 32](#_Toc136013107)

[6.2 Plan Monitoring and Maintenance 32](#_Toc136013108)

[6.3 Product Final Report 32](#_Toc136013109)

[References 32](#_Toc136013110)

[Appendix 34](#_Toc136013111)

[1. Evidencing statistical calculations 34](#_Toc136013112)

# 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

***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 1: 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 2: 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 3: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 4: 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 5: 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 6: 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 7: 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 8: 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 9.: 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 10: 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 11: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.

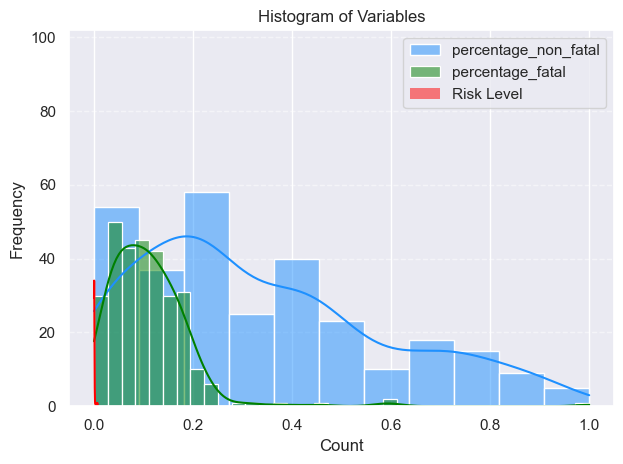


Figure 12: Kruskal-Wallis (non-parametric) test plotting result.

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.

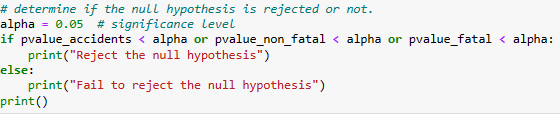


Figure 13: Determine the condition for the null hypothesis.

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 14: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.

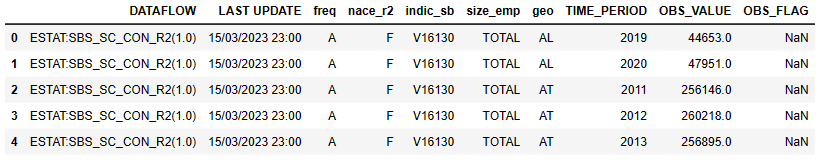


Figure 15: Original DataFrame of Employeement size.

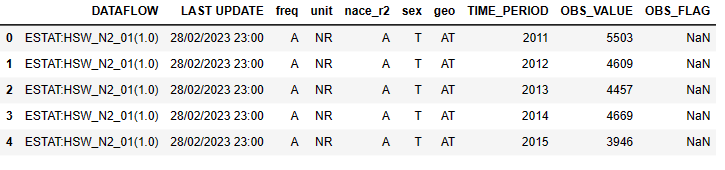


Figure 16: Original DataFrame of Number non-fatal accidents.

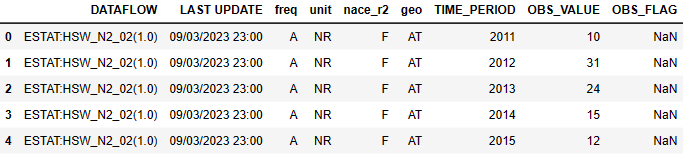


Figure 17: Original DataFrame of Number fatal accidents.

## 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.

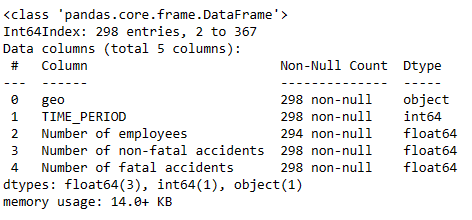


Figure 18: DataFrame info( ).

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

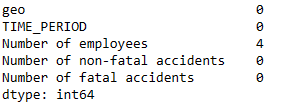


Figure 19: Missing values.

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

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

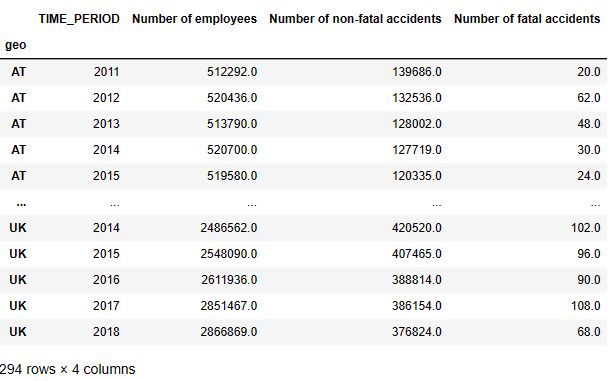


Figure 21: New merged DataFrame.

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 22: 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.

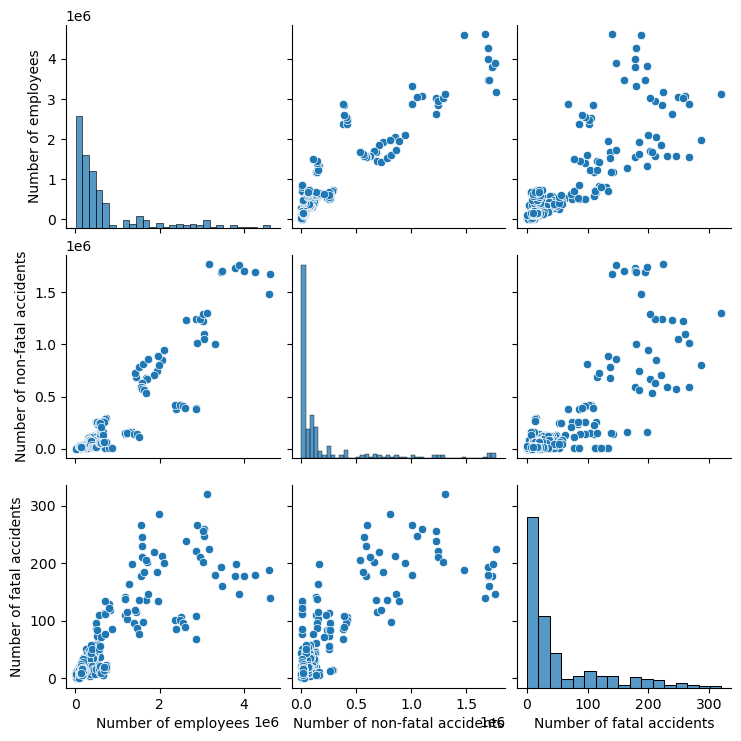


Figure 23: Pairplot correlation variables.

## 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.

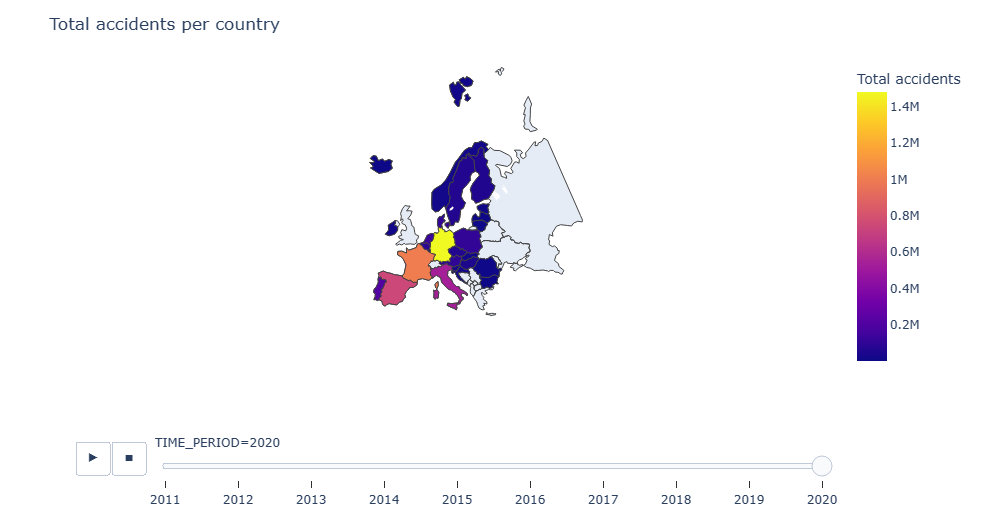


Figure 24: Choropleth 'Total accidents per country over the years'.

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

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 26: 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 27: 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 28: 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 Select Modeling Technique

## 4.2 Generate Test Design

## 4.3 Build Model

## 4.4 Assess Model

## Decision Tree Model

## Random Forest

## K-Nearest Neighbor Classifier

GridSearchCV

### K-fold cross-validation:

# 5. Evaluation

A picture containing text, line, screenshot, font

Description automatically generated

## 5.1 Evaluate Results

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.



A picture containing text, screenshot, diagram, plot

Description automatically generated

## 5.2 Review Process

## 5.3 Determine Next Steps



A picture containing text, screenshot, plot, diagram

Description automatically generated

A picture containing text, line, screenshot, font

Description automatically generated

# 6. Conclusion

## 6.1 Plan Deployment

## 6.2 Plan Monitoring and Maintenance

## 6.3 Product Final Report

# 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].

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].

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

# 



