**CCT College Dublin**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**Table of Content**

Abstract……………………………………………………………………… 4

Introduction………………………………………………………………….. 5

Python

Modules, Libraries and packages in python

Data analysis and visualization

Machine Learning

Types of machine learning

Machine learning applications

Module: Statistics for DA………………………………………………….. 7

Dataset Summary

Descriptive statistics

Inferential Statistics

Outcomes for deepen Research and challenges faced in the process

Module: Module: Data preparation and Visualization………………….. 12

Discussion on Acquiring Raw Data

EDA

Data enriching (Feature Engineering)

Interactive Dashboard

Module: Machine Learning for Data Analysis

How machine learning models work

Why to choose Supervised and unsupervised approaches?

Linear Regression

Random Forest Regressor

Logistic Regression

Grid searchCV (Cross Validation) on different algorithms

Sentimental Analysis

Tabular comparison of machine learning models

Module: Module: Programming for DA…………………………………… 24

Programming

Data from diverse sources.

Data manipulation.

Data structure.

Testing.

Optimization.

**ABSTRACT**

*This report is representing a comprehensive analysis of agricultural sector of Ireland using python programming language. By applying different techniques for data analysis and machine learning models, we have a detailed understanding of progress and pattern of agriculture in Ireland. This report contain multiple stages starting for collecting the dataset and then applying different statistical techniques and machine learning models and visualizing our data. The outcomes from this report provides the valuable insights for a farmers and government.*

**Introduction**

* **Python**

Python is high-level and dynamically typed programming language known for its simplicity and flexibility. It was created by Guido van roussam and got released in 1991. It is the most popular and widely used programming language has applications in web development, data science, artificial intelligence and more. Its extensive libraries are its strength which allows developer to do complex tasks with minimal code. Python support multiple paradigms of programming including object-oriented, imperative, functional and more which offers developer any approach which suits best for their needs.

Python’s interpreted nature facilitates in debugging and testing and dynamically typing enhance productivity. Having a vast community of developers contributing, python continuously evolving and upgrading. Even any beginner can start it easily.

* **Modules, Libraries and packages in python**

Python supports a large number of modules and packages which are facilitates the code organization and its usability. Here are the following modules and packages we are going to use:

* **Numpy**

Numpy stands for numerical python is a python library use for numerical computing and data analysis. It provides the multi-dimensional arrays which are used to store numerical data. It also provides the functions which can used to perform different mathematical, statistical and algebraic operations on that arrays. These functions are optimized and can be applied directly on arrays. Numpy can be integrated with different python modules and libraries such as pandas, scipy, matplotlib and more to extend its capabilities for data analysis.

* **Pandas**

Pandas is a most popular and widely used python library for data manipulation and analysis. Pandas contains the data frame object which resemble with the SQL tables includes the rows and columns. You can perform different operations on that data frames like sorting, grouping, merging and more same as SQL. Pandas also store the series data. A series is just a single dimensional array which includes data and is also a building block of a data frame. One of the ability of pandas is that it can handle the missing data. You can fill this missing data with the neighbouring values or remove them. This is a technique of data pre-processing. Pandas can also integrate with other libraries to for advance analysis.

* **Sklearn**

Sklearn also known as scikit-learn is a popular machine learning library use for different types of tasks including classification, regression, clustering, model evaluation, pre-processing and more. It is an easy to use library includes many machine learning algorithms. It has also use in data analysis for the pre-processing task.

* **Matplotlib**

Matplotlib is a powerful library use for visualization of your data. It includes the wide range of plotting functionalities which can be used by anyone to visualize their data. In matplotlib there is a module name as pyplot, is a essential tool for the visualization of data. It helps the data analysts to plot your data like scatter plot, line plot, bar chart, histogram and many more. Matplotlib can also integrate with different python libraries to create an interactive visualization.

* **Seaborn**

Seaborn is also a powerful library for data visualization and it based on matplotlib. One of the use of seaborn is that it can give the useful visualization with a minimal code. Some of the functions of seaborn are lineplot, boxplot and heatmap which provides relational information within the data. Another feature of seaborn is that it has the built-in styling and theming. It contains the variety of colour themes which allows the user to change the interference of their plots. In short, whether you are exploring dataset or visualizing trends, seaborn provides the tools which enhance the visualization of your data.

**Module: Statistics for Data Analytics**

* **About Dataset:**

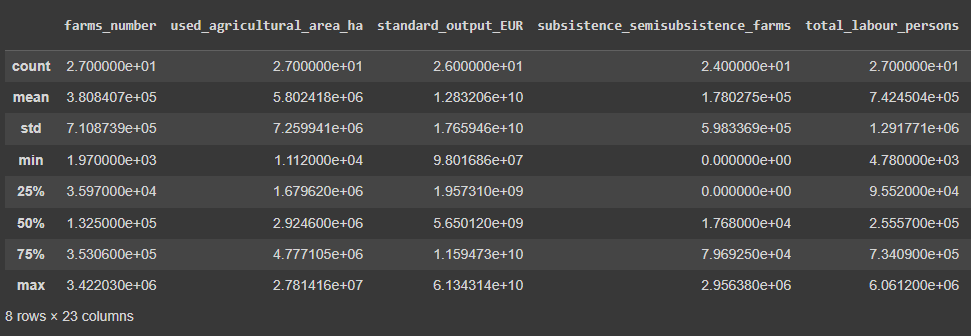
This dataset contains the data of agricultural sector of European Union countries (UK was removed later from the dataset. This data covers the year 2016. The dataset was downloaded from different resources. Further analysis and techniques were then applied to it.

* **Descriptive statistics:**

Descriptive statistics is a statistical technique use for the summarization and understanding of our data. This achievement provides a simple and clear overview of the data.

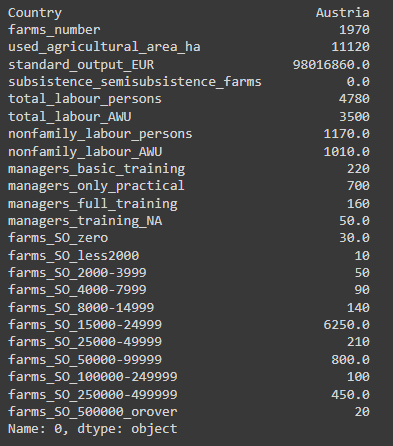
* Some important statistical results from a dataset:

This includes the count, mean, standard deviation, minimum, maximum, 25%, 75%, 50% values in the dataset.



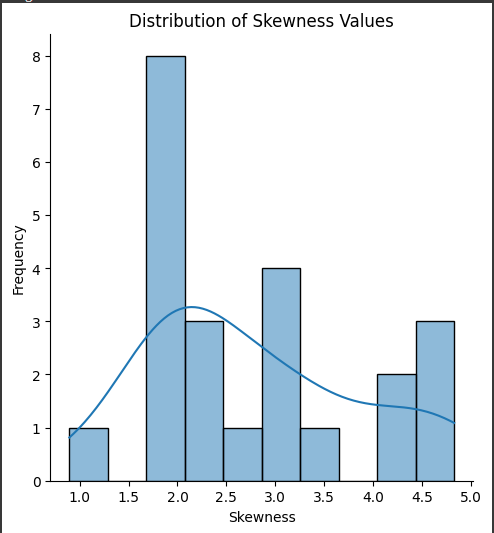
* Mode for each country:

Here it is telling one from each column but in the dataset every value is unique.



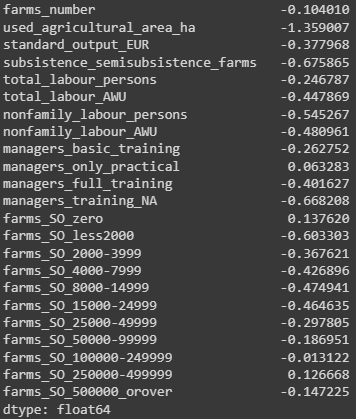
* Checking for skewness:

Skewness tells us that in which side the distribution of our data is. In our dataset’s case, the data is positive skewed which can be seen below.



* Log transformation:

Log transformation is a statistical technique used to stabilize the variance of our data. This technique involves of applying the algorithmic function.



* **Inferential Statistics:**

Inferential statistics is a statistical technique that focuses on making predictions and generalizations about a population based on sample data. We conduct several tests in it.

* Confidence Interval:

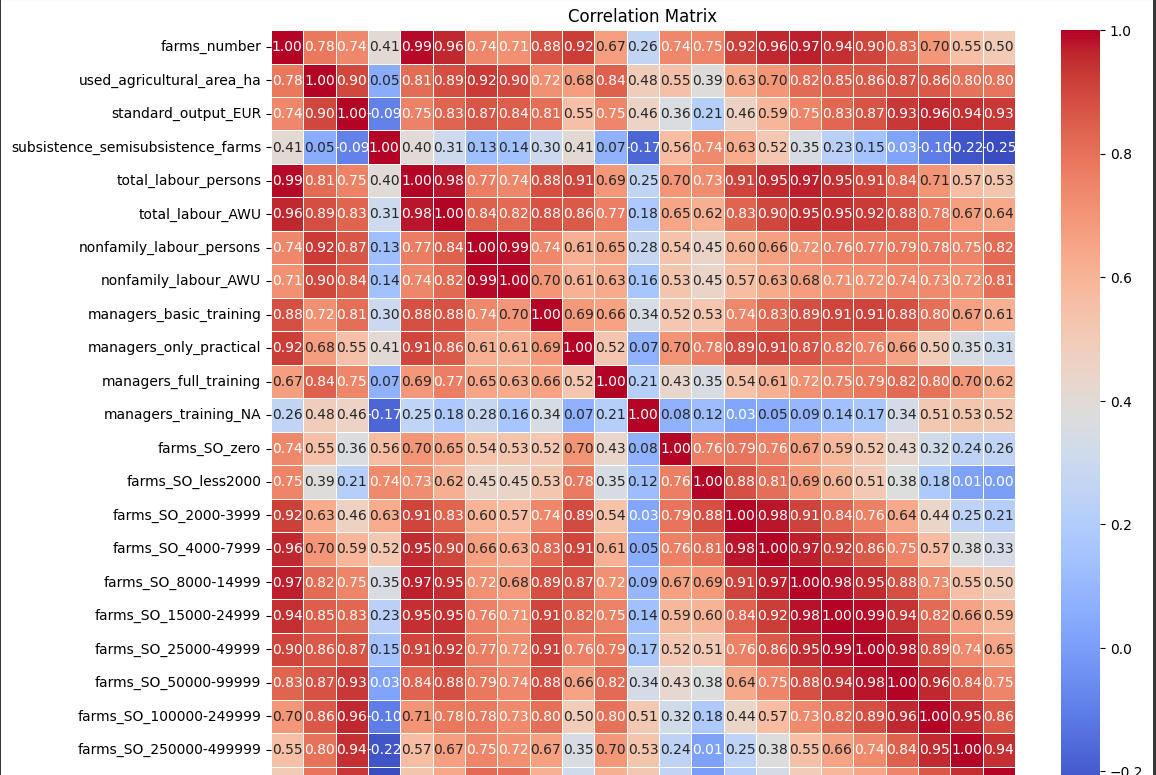
This statistical technique provides a comprehensive summary of parameters for the population.

The results from two columns are given below:



* Correlational analysis:

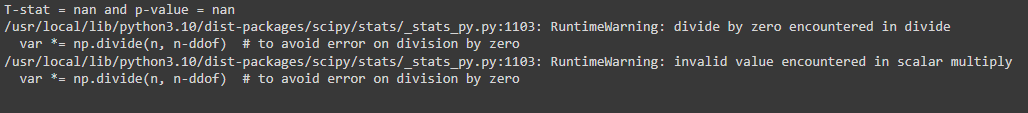
Correlational analysis is a statistical technique used to find and evaluate the direction and strength between two or more variables.



* T-test (Parametric):

A t-test is a measure of significant difference between two groups which can be related in certain features. As it is a parametric test means that it makes assumptions about parameters.

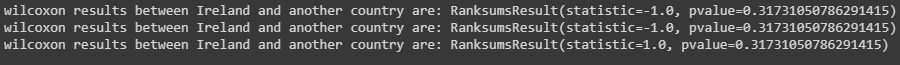
The results from t-test parametric are given below:



* Wilcoxon Test (Non-parametric):

The Wilcoxon test is the non-parametric test used for related samples or paired just to observe whether their population mean differ or not.

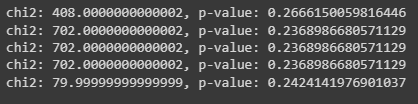
Results from Wilcoxon test are given below:



* Chi-Squared Test (Non-Parametric):

This statistical test is also non-parametric used to determine any significant association between categorical variables.

The results from this test are given below:



* Kruskal-wallis test:

This statistical test is used to determine any significant difference between three more groups.

The results from this test are given below:



* **Outcomes for deepen Research and challenges faced in the process:**

The results from this research provided insights that can leverage to deepen research. Here are the some ways we can further go in it:

* Interpreting results to formulate new hypothesis.
* Post-hoc analysis.
* Refining theoretical models.
* Extend research scope.
* Addressing limitations and biases.
* Validate results with additional data.

The following challenges we faced:

* Data quality and availability.

Inconsistent and missing data can impact the analysis’s validity.

* Statistical assumptions.

Making sure data meets the assumption of statistical tests we used.

* Computational complexity.

Large dataset and complex analysis can be computationally expensive.

* Ethical and practical constraints.

Results that are non-significant are difficult to interpret.

**Module: Data preparation and Visualization**

* **Discussion on Acquiring Raw Data:**
* Data Source:

The data was sourced by Eurostat. It is the statistical office of European Union. Eurostat collaborate with NSIs of each EU state for that data. The method for collecting the data includes survey and reporting from farms. The data undergoes through various checks to ensure consistency. The uniformity was ensured in the data. Eurostat published the data on their website that is accessible for public. Data can be downloaded in various formats i.e. csv.

* Positive aspect of Data acquisition:

The open data policy of Eurostat ensure the free access of data.

Data is available in various formats.

They ensured high quality and reliability.

The dataset provides the holistic view of agriculture sector.

In depth analysis is possible due to comprehensive coverage.

* Negative aspect of Data acquisition:

In the process of data collection, there can be lag.

The harmonizing process of data from multiple sources can be complex.

There is a data collection variations in different countries.

Open access can lead to misuse of the data.

* Relevance and Implications of Licensing/Permissions:

Allows the copy, share and distribution of data.

Data can be adapted by users.

Users must give the appropriate credit to Eurostat.

Ensure that data remains open.

* Implications for Research and Policy-Making:

Policymakers can leverage the data to make informed decisions.

Start-ups and businesses can use the data for market analysis and trends.

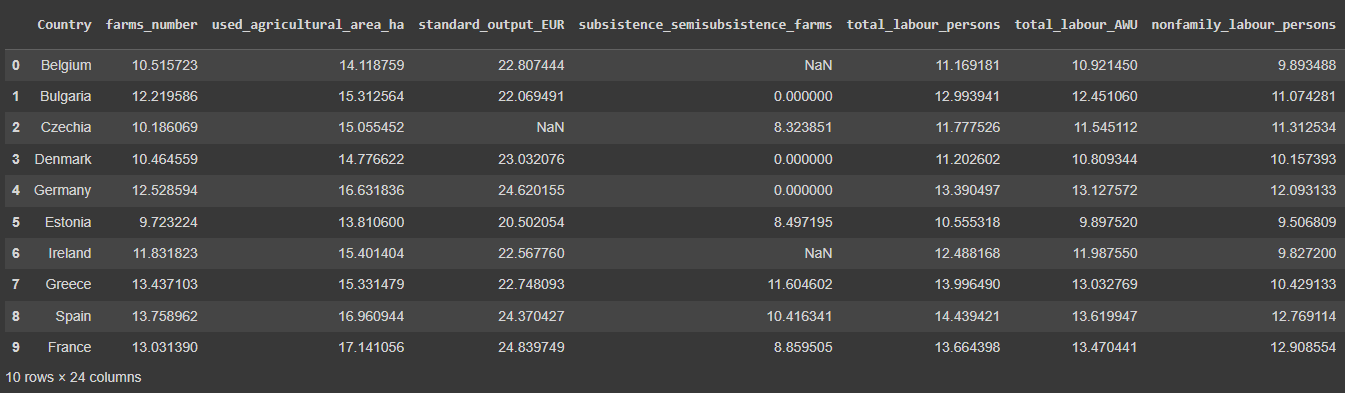
Open licensing supports transparency.

* **EDA:**
* Shape of dataset:

First step is always to check for the shape of your data.

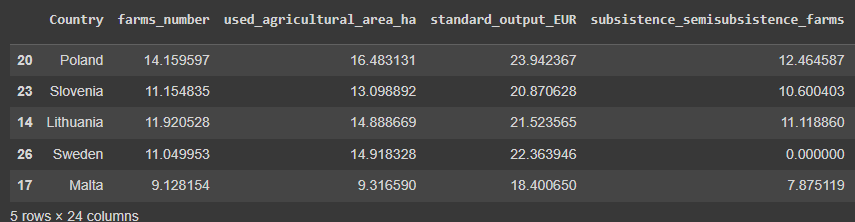
The shape of our data is (27, 24).

* An overview of dataset:



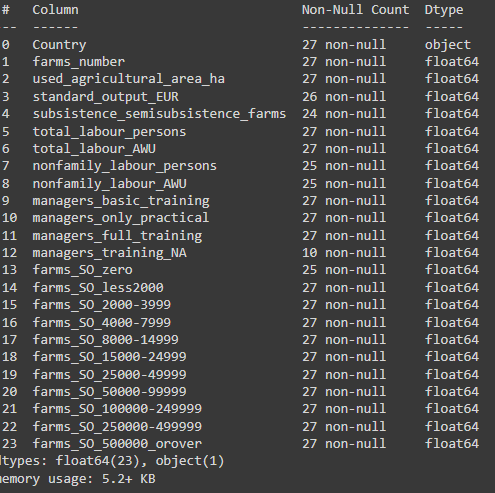
* Checking for random rows in dataset:

It is a good practice to always see the random rows from dataset because not every time the initial rows tells us about our dataset.



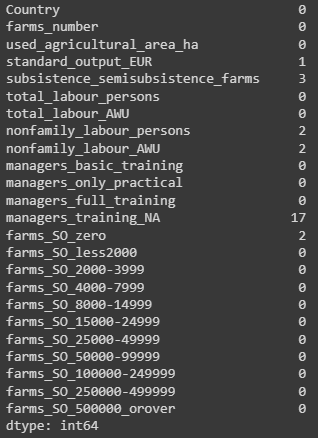
* Information about dataset ( data types of columns and non-null values ):

The df.info() function tells about the basic information about our dataset.



* Checking for total null values in each column:

Null values effects the process of pre-processing and throw errors in the process of training. For that purpose we have to see the number of null values in each column so that we can deal with them.

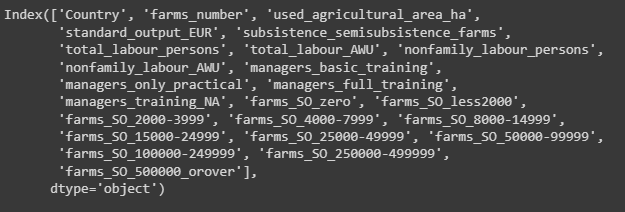


* Checking for duplicates:

There are zero duplicate values in the dataset.

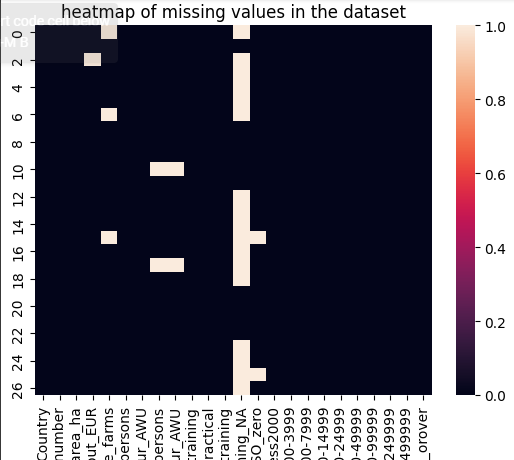
* To check the all columns by name:

Then it comes to see the names of columns in our dataset. Because sometimes having a large number of columns, we are unable to see every column and also it takes more time and effort for going through each column. This approach put all columns at one place.



* A heat-map of missing values:

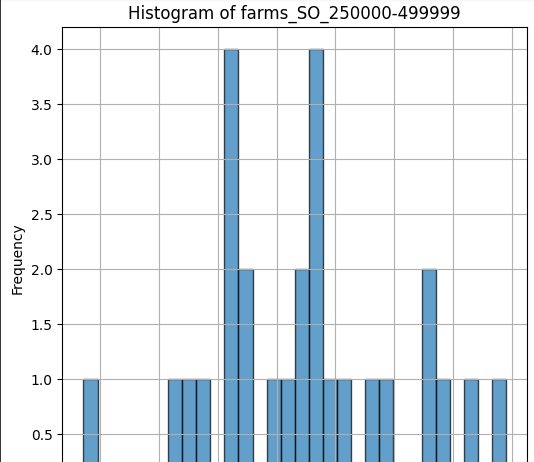
This tells us about the pattern of missing values in the dataset.



* A histogram:

This divide the numerical values in bins and tells us about the frequency of different values. The histogram of our dataset is given below:

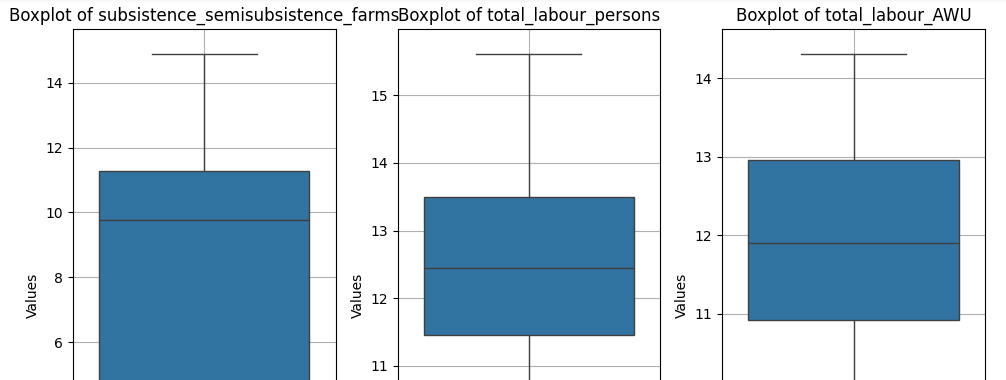
There are multiple histograms of each column but here we will represent just one.



* Anomalies detection:

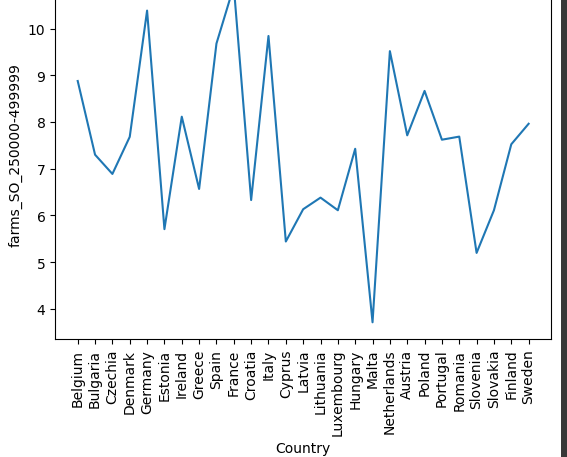
The box plot has a best use case here to check the anomalies in our dataset:

We checked for anomalies in each column but here are the representation of three:



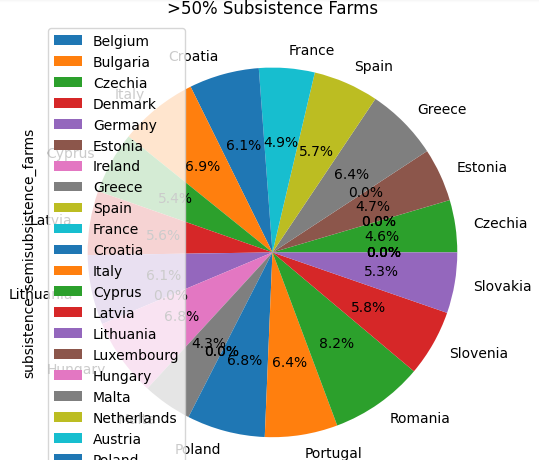
* The Line plot:

The line plot tells us a graphical representation between two variables. We draw multiple line plots between columns for different countries but here are one of them:

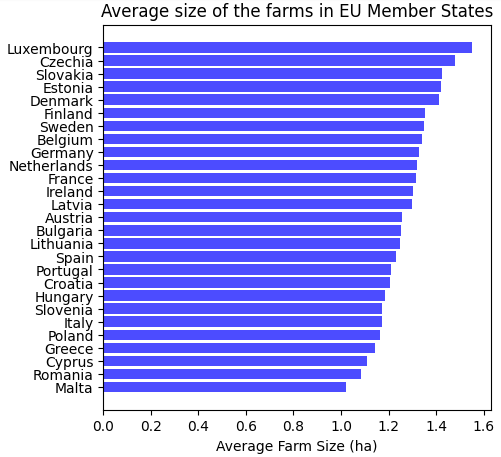


* Checking for >50% Subsistence Farms:

The pie chart is in best use case here. It representing has Romania has more >50% Subsistence Farms than any other country.



* Average farm size in European Union:



* **Data enriching (Feature Engineering):**

To improve the performance of our machine learning algorithms, we use a technique called as feature engineering. It starts with filling missing values, feature scaling, feature extraction and goes up to feature construction. We will see the techniques we used.

* Drop columns:

The more number of null values a column have, the more useless it going to be for a machine leaning model to train. So here we drop a column which has more number of missing values in that.

* Filling missing values:

In our dataset, we can’t drop the rows as it is representing any county. So we have to fill the missing values. For that purpose, we filled the missing values with the mean value of each column.

* Removing Outliers:

We checked for outliers in our dataset. We calculated the z-score values for each column. Then we set a threshold. Any value exceeding that threshold was an outlier. We didn’t removed any because we just one outlier which is not going to effect our model.

* Standardization:

Standardization is a concept of placing all values in a specific range. We standardized our columns so any feature can’t dominate any other in the process of training.

* One-Hot-Encoding:

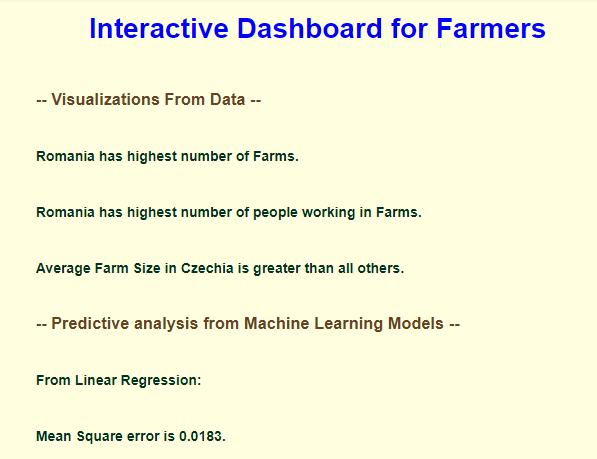
Machine learning models can’t train on a categorical column, so we one hot encoded our features so that categorical column got their own identity for training.

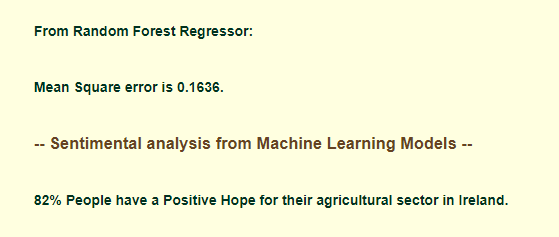
Till that point, the process of feature enriching has been completed.

* **Interactive Dashboard:**

As modern farmers relies on technologies for information communication. For that purpose we have developed a dashboard which represents all the important analysis and finding of our dataset and machine learning models.

Here it is:





**Module: Machine Learning for Data Analysis**

As it is a Regression problem, we have to implement the Regression algorithms of machine learning. All our data is pre-processed now and ready to apply different algorithms.

* **How a machine learning algorithm works?**

First of all we select our input and output columns as x and y. These inputs are the columns of the data frame. Then we split our input and output in training and testing. After that, we train our algorithm on training data. When training is done, predict the output on test data or unseen data. At the end, we calculate the accuracy of our model means how well it predicted on unseen data.

* **Why to choose Supervised and unsupervised approaches?**

Our problem was a regression task and our data was labelled. For that purpose we used the supervised learning technique. Because supervised learning techniques work on labelled data.

* **Linear Regression:**

Linear regression is one of the fundamental and commonly used algorithm in machine learning. It do a relationship between a target variable and one or more independent variable i.e. features.

Rationale and justification:

Linear regression is simple to implement and evaluate. The coefficient provides an insight on dependent and independent variables. The model coefficients can easily interpreted. Linear regression is computationally efficient which allow the quick training. It is a benchmark for complex models. When relation between targets and features is almost linear, then this model is best in use. Linear regression provides the high accuracy. Several techniques in linear regression can help in handling multi-collinearity.

Linear Regression MSE: 0.018286724960515543

* **Random Forest Regressor:**

It is an ensemble learning method that work by constructing multiple decision trees during training. Then it output the mean prediction.

Rationale and justification:

Random forest regressor can capture the non linearity relationship between input and target variables. Due to their ensemble nature, the model do not overfit. It can estimate the feature importance. It can be used for both regression and classification tasks. Random forest provide the stable and consistent prediction. Rather than a single decision tree, random forest provides the better accuracy. I can also perform better on noisy data. It can even perform better if features and not scaled.

Best parameters: {'max\_depth': 20, 'min\_samples\_split': 2, 'n\_estimators': 100}

Mean Squared Error: 0.15927887151392736

* **Logistic Regression**

A supervised machine learning algorithm used to predict the output of a categorical data. Goal is to find the probability that an instance belongs to a particular class. We use logistic regression and got our results. The accuracy was more than 86 percent. We use that model for sentimental analysis.

Rationale and justification:

Using logistic regression is one of the efficient technique for solving classification related problems. It is really fast, easy to train and perform well if the data is linearly separable. For modelling the dependent variables, it uses the logistic function. Logistic regression has three different types. First is binomial. This technique is used when the target variable has 2 types. Second one is multinomial. This one is used when the target variable has 3 or more types. And the last one is ordinal. This is used when the target variable has ordered categories. The sigmoid function is used to map the predicted values.

* **Grid searchCV (Cross Validation) on different algorithms:**

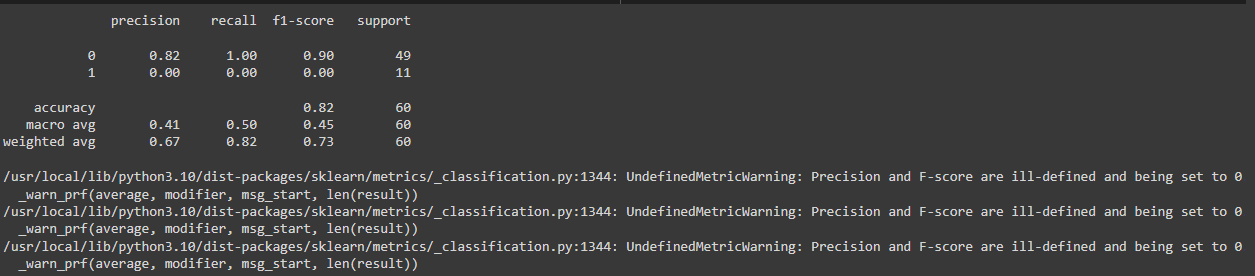
Grid search CV is a cross-validation technique. We basically use it for hyper parameters tuning of the model. It returns the best set of hyper parameters which returns the best accuracy. We applied it on random forest regressor. It gives us the best set of parameters on which model accuracy was high.

* **Sentimental Analysis:**

Sentimental analysis is also known as opinion mining is the process of identifying and categorizing opinion based on text. The goal is to categorize then positive or negative. It has a lot of applications in social media, business and many other places.

The approach we are going to do here is a machine learning approach. We will create a random data by our self. The data will include the positive and negative sentences about Ireland agricultural sector. Then we will create the random two to three features and assign them some values. After that we will gave the sentiments the random values 0 or 1 for machine learning model. Then we will divide our data in train test and split and then feed it to machine learning model. We will make an accuracy score at the end which will tell us how better our model performed on unseen data.

Here is accuracy report of our model:



Our model performed well on unseen data. 86% people has a positive point about the agricultural sector of Ireland. These people includes the consumer and producers.

* **Tabular comparison of machine learning models:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Used for** | **MSE or Accuracy** | **Performance** | **Similarities** |
| Linear regression | Predictive analysis | 0.018286724 - MSE | - | Use for regression problems |
| Random Forest | Predictive analysis | 0.159278871 - MSE | Good | Use for regression problems |
| Logistic Regression | Sentimental analysis | 0.82 - Accuracy | Good | Use for binary classification problems |

Above table represents the comparison between the different machine learning algorithms used for predictive and sentimental analysis.

**Module: Programming for DA**

* **Programming:**

The project has been explored programmatically and all the code has been implemented in jupyter notebook. The documentation (report) of the dataset tells the clear justification and explanation of code choices. Code quality standards has been applied.

* **Data from diverse sources:**

The important libraries like pandas, numpy and sklearn has been used to processed data from diverse sources. These libraries with their working has been introduced above in introduction’s section.

* **Data manipulation:**

We downloaded our data from two different sources. Then we manipulate our data and combine them to make one dataset.

* **Data structure:**

The data has been downloaded from multiple sources in both csv and json formats.

* **Testing:**

Testing strategy plays an important role in any data analysis task that ensure the accuracy. The primary goals of testing strategy are: ensuring data accuracy, validating the correctness of data accuracy and verify and fix bugs in early development of the code.

Some testing strategies are:

* Unit testing:

Testing individual components such as functions and classes.

Coverage in project: Data cleaning functions, data transformation scripts and statistical and machine learning model functions.

* Integrating testing:

This ensure that modules are working good together.

Coverage in project: data pipeline from input to final output, interaction between data processing, visualization and analysis modules.

* Performance testing:

To ensure that code can handle large dataset effieciently.

Coverage in project: execution time for key functions, memory usage during data pre-processing.

Trades off:

* Priority was given to critical components while less critical components received lighter testing.
* Automatic testing provides consistency while manual testing ensure accuracy.

Challenges:

* Standardized tests were difficult as agricultural data is highly variable due to seasonal and regional differences.
* Performing testing was difficult with low computational resources.
* **Optimization:**

Optimization is necessary for ensuring that data analysis is efficient and scalable often dealing with large dataset and low computational resources.

Primary goal of optimization strategy is: reduce execution time, minimize memory usage, ensuring scalability for large dataset.

Optimization approaches:

* Efficient data processing: replaced loops with vectorized operations using libraries such as numpy and pandas.
* Parallel and asynchronous processing: the purpose was to reduce the cpu cores and improve execution.
* Memory management: to optimize and manage memory.
* Algorithm optimization: this strategy was used to improve the algorithm’s efficiency.

Trade-offs:

* Resource constraints.
* Complexity of agricultural data.