***Analysis and predictions on land sales in Ireland***

Abstract.

Determining land prices is very challenging task now a days for investors and customers both as land prices are dynamically changes.

The focus of this project is to apply data science and machine learning techniques that can help investors by predicting what prices they can maintain.

**Introduction**

*Project scenario*

In Ireland an investor wants to know about land prices. But he get confused when he comes to know different prices rates.

To overcome above mentioned problems this this investor decides to make a robust prediction system for its company.

**Data Description:**

The dataset which I am going to use here is dataset of land sales in Ireland. It is openly available on Kaggle.

Dataset source: <https://data.gov.ie/dataset/ara02-agricultural-land-sales?package_type=dataset>

The important features are:

* **Year**
* **Type of Land Use**
* **Region**
* **UNIT**
* **VALUE**

**Reason for Analysis:**

The reasons of doing this Analysis are following:

* To extract information, facts and figures from raw data
* To understand pattern of data.
* To know about trends.
* To make visualizations to represents data in the form of graphs and charts. As a picture is worth than 1000 words.
* To show what relationships different feature are having among them.
* To make an accurate Machine Learning Model which can be used to land sales on the basis of input data?

**Technical requirements:**

There are two main tools which are used to do analysis which are following:

* **Python:**

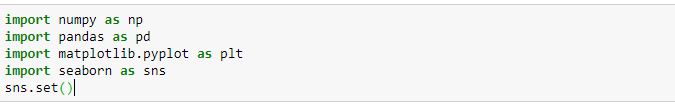
Python is programming language which is used here .I have to perform this data analysis using python

* **Jupiter notebook:**

Jupiter notebook is tool on have to I have to write python script

Importing libraries and dataset

The first step is to import all required libraries.so here I am importing four Major Libraries which are here:



1)**pandas**:

pandas is used for data crunching.

2)**numpy** :

It is used for scientific computing it provides some mathematics and statistics like functions.

3)**Matplotlib:**

It is used for data Visualization For example to make Charts and graphs.

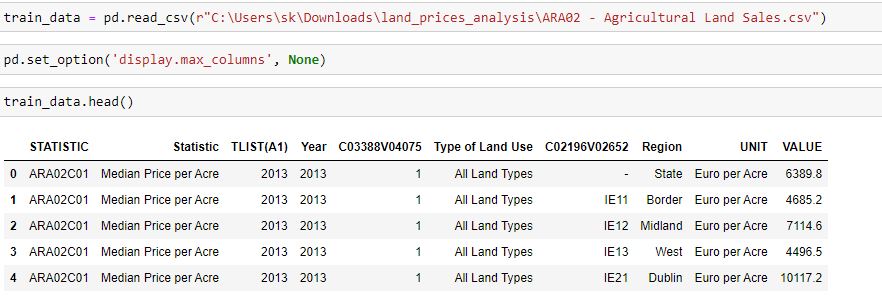
4)**Seaborn**:

It is also used for data Visualization For example to make Charts and graphs.

so these are four major python libraries which I have imported to use.

Import csv file and create a data frame

Now here I am loading data. As data is in excel file so I am importing Excel File with .csv file format.

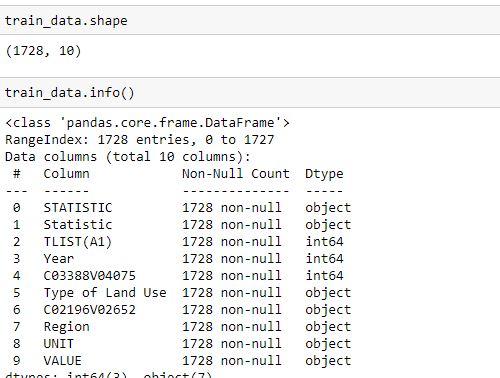


As here excel data is loaded into a data frame 'train data'.

Data Frame is also like a table.

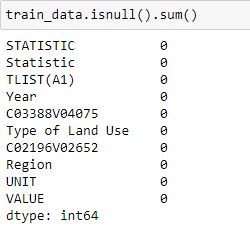
At the end first 5 records are displayed.

Size/shape of data and general information about data



There are 10 features or columns and 1728 rows or records in dataset.

Checking for any missing values

****

As shown above there are no null or missing values in data.

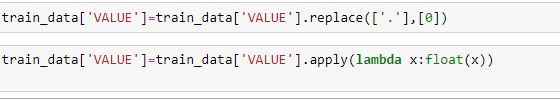
# Feature Engineering

***Removing ‘LIST(A1)’ feature***

******

As ‘’LIST(A1)’’ is similar to year feature so it I removed.

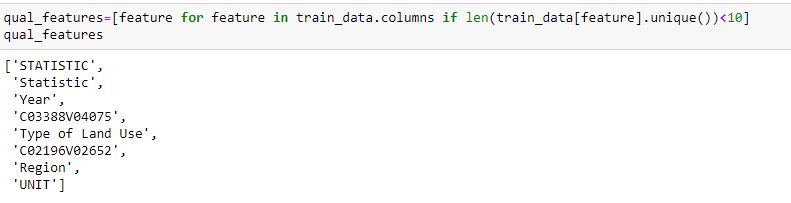
***Converting value feature into float type***



As value feature has data type of object so it is converted to integer but problem is that there is dot character is exists so at first dot character is converted to zero than type of Value feature is changed by using lambda function

# Exploratory data analysis

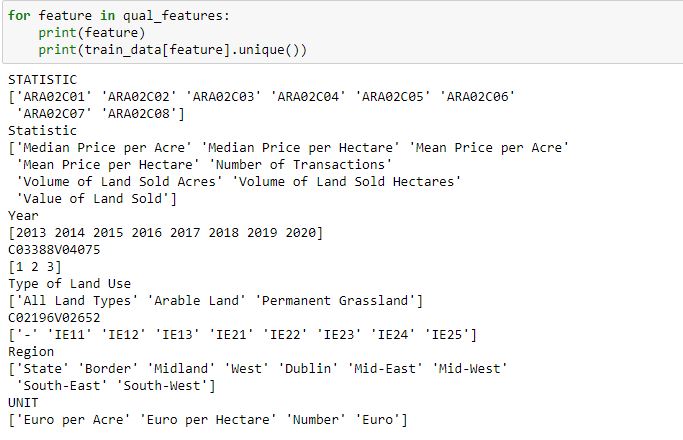
Extracting qualitative features



Here above qualitative features are those features which are non-numerical features having less than 10 categories.

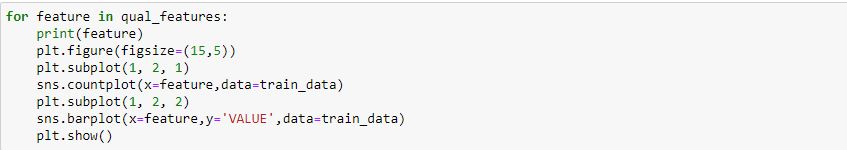
These features are also called categorical features.

Unique values of categorical features

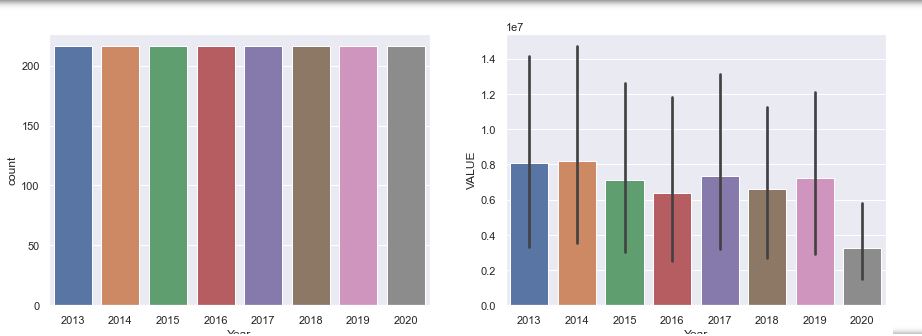


Here above are unique values of all categorical features .

Exploring categorical features using bar charts



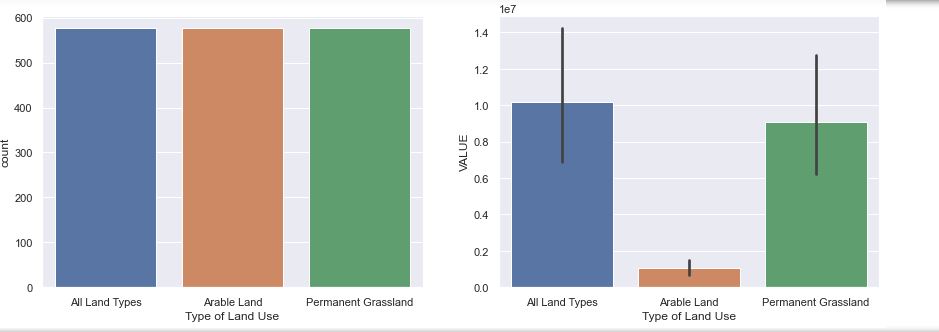
Relationship of years with sales



Left bar chart shows all years has same frequency.

Right bar charts shows that 2013 and 2014 are years with high sales

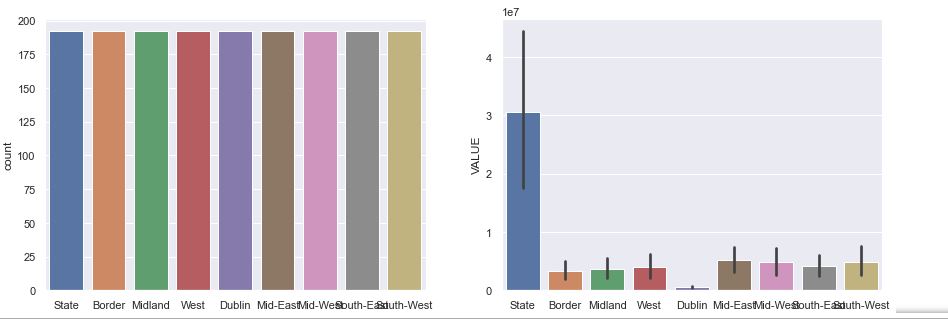
Relationship of land use with sales



Left bar chart shows that all types of lands have same frequency.

right chart shows all types of land and grassland are lands with high sales

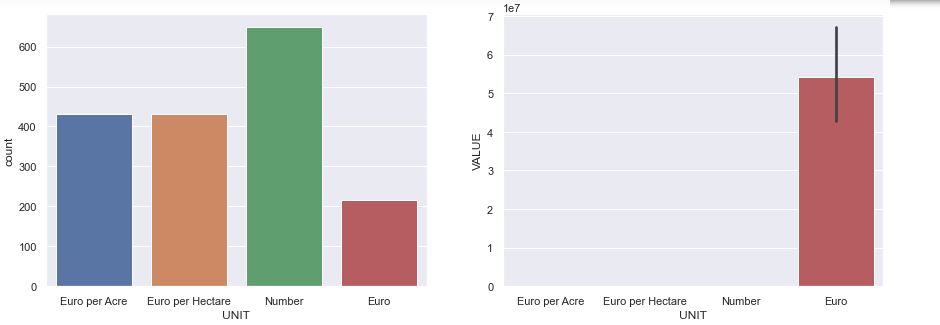
Relationship of region with sales



Left bar chart shows that all region have same frequency

Right bar chart shows that state is region with high sales of land

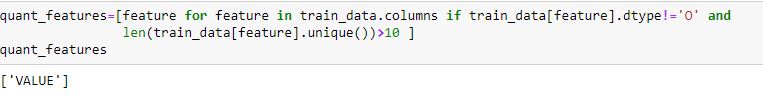
Relationship of unit with sales



Left bar chart shows that Number has high frequency than other

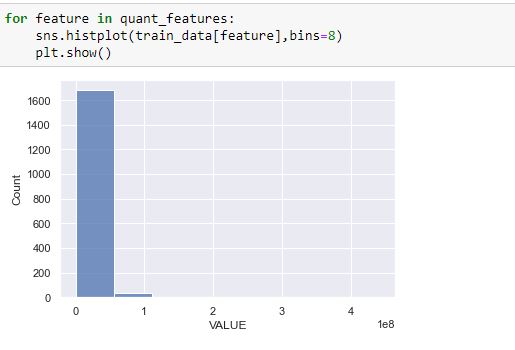
Right bar chart shows that Euro is currency with high sales.

Quantitative features:



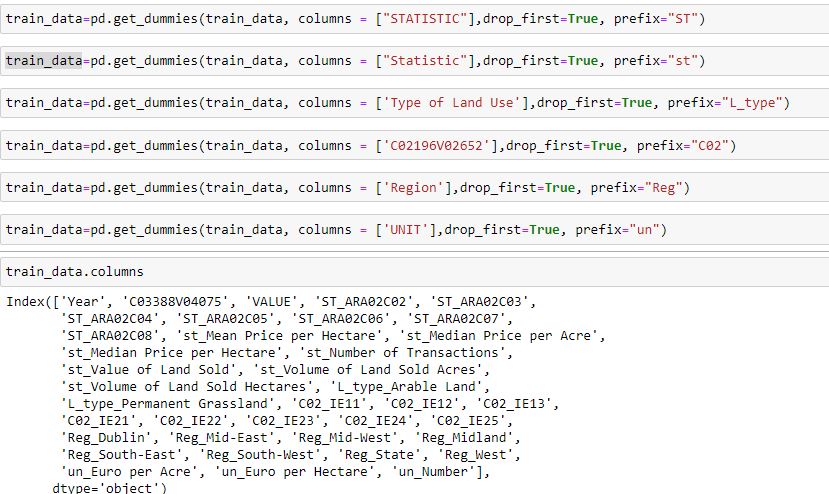
There is only one feature which is numerical and have more than 10 unique values.

Showing distribution of values in quantitative features using histogram



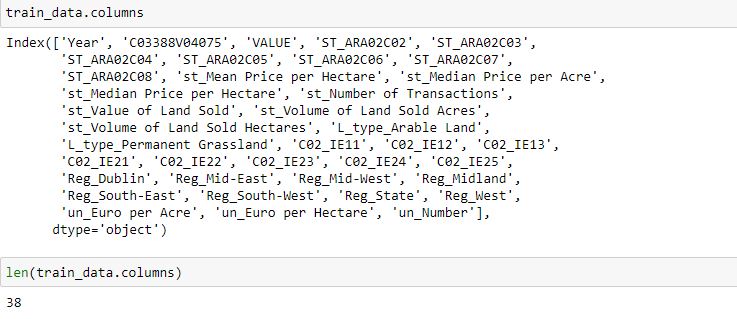
So this is distribution of VALUE so as by looking on histogram it is clear that 0-50% has highest sales VALUE

Handling categorical features using One hot encoding

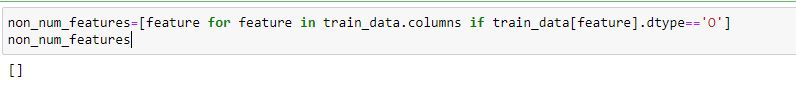


Here six categorical features are converted to numerical features.

So that can be used as inputs in training.



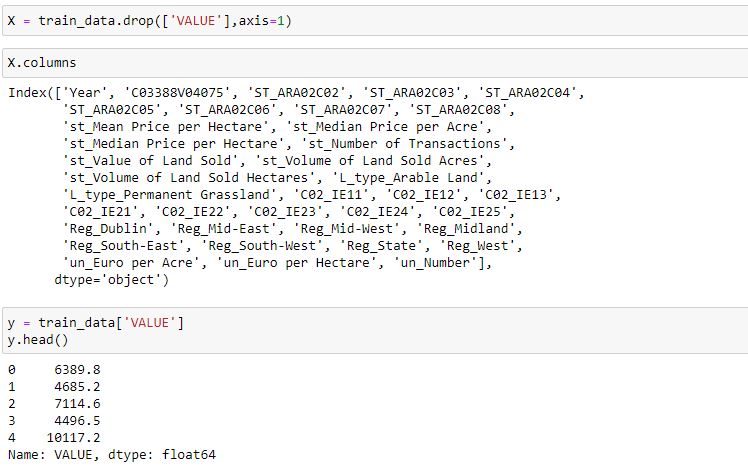
So there are total 38 features after performing one hot encoding so at beginning there are 10 features. After one hot encoding it becomes 38 features.



As after doing one hot encoding there is no categorical features

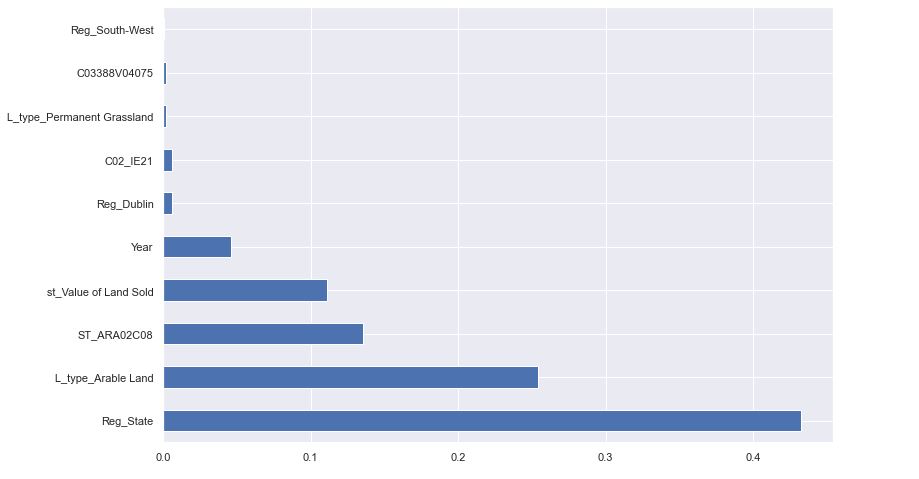
# Preparing dataset for regression

Splitting of data into dependant and independent



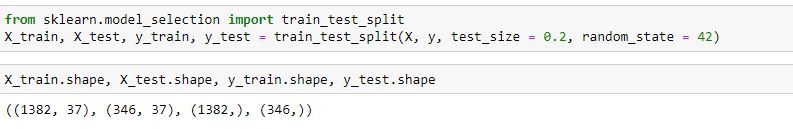
As value is targeted feature so it cannot be used as inputs so that is why it is removed from independent or inputs feature. Here inputs are represented by X

Visualizing feature importance



This bar chart is showing importance of each feature with respect to target feature sales. As by looking here it is shown that region of state and Arable land has high feature importance.

Splitting of data into train and test



Here dataset is divided into train and test part at the ratio of 20% .test part is used for evaluation of model.

Machine learning Model Building

 Here following three regression algorithms are used:

* Linear Regression
* Decision Tree Regression
* Random Forest Regression

Data training phase is now done here

# Predictions

Exploratory Data Analysis or EDA is used to take insights from the data. Data Scientists and Analysts try to find different patterns, relations, and anomalies in the data using some statistical graphs and other visualization techniques

1.Get maximum insights from a data set <br>

2.Uncover underlying structure <br>

3.Extract important variables from the dataset <br>

4.Detect outliers and anomalies(if any) <br>

5.Test underlying assumptions <br>

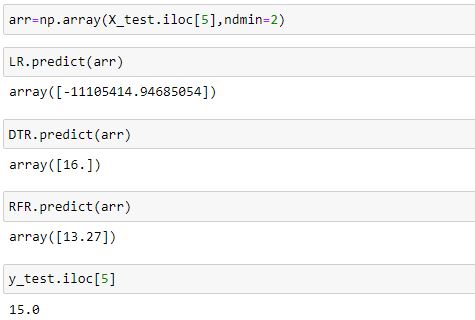
6.Determine the optimal factor settings <br>

Building predictions

Here below I am going to perform predictions so 5th record of test data is selected as inputs.

As row data is one dimensional data .So that is why it is converted to two dimensional data

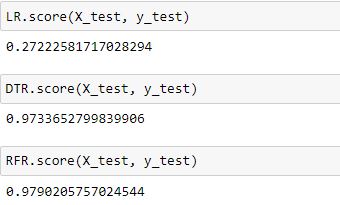
As predict function requires inputs to be two dimensional.



Hare it the difference between predictions and actual value can be compared.

Decision Tree Regression and Random Forest regression have predictions near than actual value.

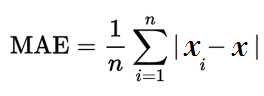
Accuracy scores of machine learning models

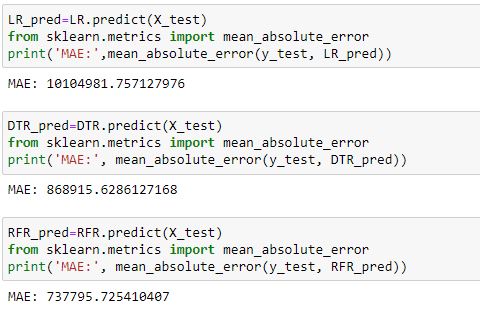


Here are accuracy scores of three different regression models.

Random Forest and Decision Tree has high accuracy scores but linear regression has very less accuracy score

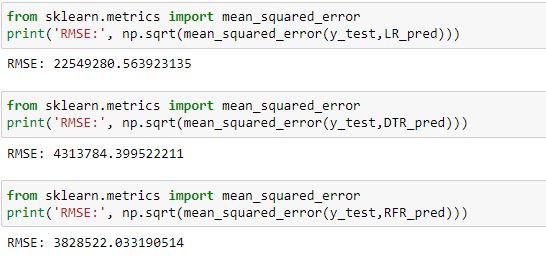
Mean Absolute Error

The Mean Absolute Error(MAE) is the [average](https://www.calculushowto.com/average-value-of-a-function/#def) of all absolute errors. The formula is:  
[](https://www.statisticshowto.com/wp-content/uploads/2016/10/MAE.png)



Here is Mean Absolute Error of three models .It shows that Random Forest Regression has very less MAE error which shows that it is very accurate.

Root Mean Square Error

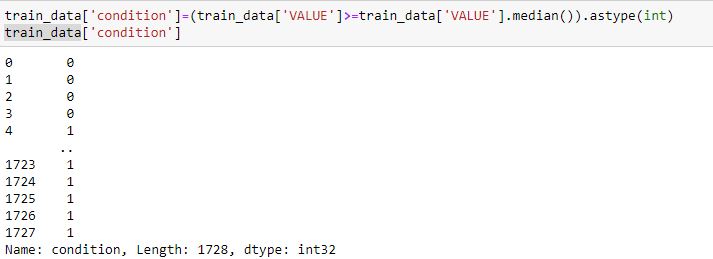


Here are root mean square errors of three models.

As by looking above it is clear that Random Forest regression is machine learning model very less Root mean Square Error root mean square error.

# Preparing dataset for classification

Making a new categorical feature for performing classification



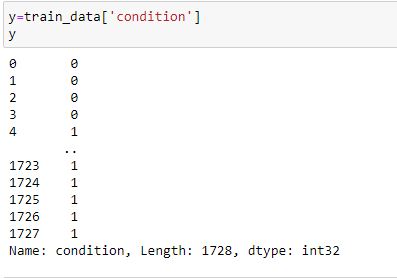
Here a new feature is developed by taking median as threshold of VALUE so above median is considered as expensive and below median is considered as cheaper than above median is converted to 1 and below median is converted to 0.

Preparing independent data by removing two features



As condition is targeted feature so it is removed with its parent feature VALUE so here X is data frame which have inputs features

Preparing dependant feature



Here condition is that feature which is to be predicted so it is targeted feature or dependant feature.

It has values of zeroes and ones.

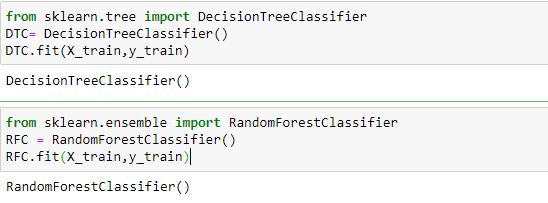
Splitting data into train and test



Here X\_train and y\_train is 80 % of data which is giving as inputs to model

Here X\_test and y\_test is covering 20% of data which is used to evaluate models.

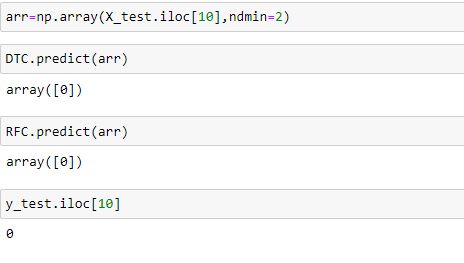
Building classification models.



Here two classification models are used which are following:

* Decision Tree Classification
* Random Forest Classification

Doing predictions

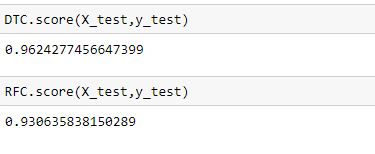


Here are predictions of two classification models on the basis of row 10 of test data.so all are giving same result.

# Evaluating classification models

Accuracy score

dividing the number of correct predictions (the corresponding diagonal in the matrix) by the total number of samples is called accuracy



Here Decision Tree classification has high accuracy than Random forest regression

F1\_score

The F1 score is defined as the harmonic mean of precision and recall.



As by examining F1 score it is clear that Decision Tree Classification has high F1\_score

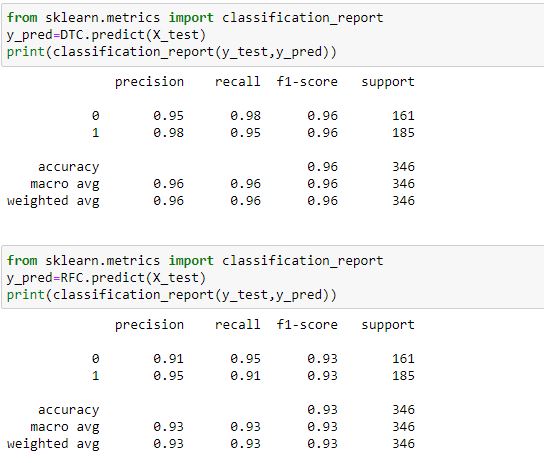
Precision ( ) is defined as the number of true positives ( ) over the number of true positives plus the number of false positives ( ). Recall ( ) is defined as the number of true positives ( ) over the number of true positives plus the number of false negatives ( ).

So it has high performance.

Classification report

What is a classification report?

A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of your trained classification model.



Here this classification report shows precision recall and F1\_score of each category

So in both models category of one and zero has same F1\_score.

# Conclusions:

On the basis of analysis and visualization these are conclusions.

Missing values

There are no missing values in dataset so it is not needed to clean dataset.

Feature Engineering:

Here one feature is removed as this is similar to year

Here a feature ‘’VALUE” IS Converted to 1 and zero by replacement

Best model:

As Random Forest regression has high accuracy in regression but decision tree classifier is best model with high accuracy rate in classification.

# Reference:

Dataset at <https://data.gov.ie/>:

<https://data.gov.ie/dataset/ara02-agricultural-land-sales?package_type=dataset>

**Article on this topic ‘’How To Perform Exploratory Data Analysis -A Guide for Beginners’’**

**Machine learning tasks:**

<https://machinelearningmastery.com/machine-learning-in-python-step-by-step/>

here is step by step method of building machine learning model

**Exploratory data analysis:**

[**https://medium.com/@hammad173216/my-first-exploratory-data-analysis-project-in-python-on-sales-analysis-2e48880d4094**](https://medium.com/@hammad173216/my-first-exploratory-data-analysis-project-in-python-on-sales-analysis-2e48880d4094)