



SAPIENZA
UNIVERSITÀ DI ROMA

TENDER MARKET ANALYSIS USING ELASTIC SEARCH AND KIBANA

Facolta di Ingegneria dell'Informazione, Informatica e
Statistica Corso di Laurea Magistrale in Data Science

Candidate

Chittamuru Greeshma Spandana

Thesis Advisor

Prof. Ioannis Chatzigiannakis

Co - Advisor

Juan Carlos Martinez Perez

Academic Year - 2021/2022

TENDER MARKET ANALYSIS USING ELASTIC SEARCH AND KIBANA

Master's thesis. Sapienza – University of Rome

© 2022 Chittamuru Greeshma Spandana.

All rights reserved Author's email:

spandana061297@gmail.com

Abstract

This thesis proposes an approach of Tender Market analysis with data (From Telemat) which comes from a firm working in the Data-Driven Transformation Company specialized in the design and delivery of solutions. The proposed model used in the thesis uses an Elastic search and Kibana to provide users with the ability to analyze and visualize the data. Daily data from 2021 to 2022 are obtained from the website Telemat has been providing information on public tenders daily, obtained through constant monitoring of the official sources and the websites of the Bodies, including the Central Purchasing Centers and the platforms of the Electronic Market. Telemat is a dynamic and lively reality able to offer customers new services and constant and personalized support for their business growth approach contributes to studies aimed that providing more accurate and reliable visualization, analysis and prerequisites for efficient platform management.

Table of Contents

ABSTRACT	3
1 INTRODUCTION.....	3
1.1 ABOUT ICONSULTING	3
1.2 ABOUT TELEMAT.....	3
WHO THEY ARE: SPECIALISTS FOR PUBLIC TENDERS.	3
2 DATA DESCRIPTION.....	4
2.1 DATASET DESCRIPTION.....	4
3 DATA PREPROCESSING.....	6
3.1 WHY DO WE NEED DATA PRE-PROCESSING?	6
3.1.1 <i>Get the Dataset</i>	6
3.1.2 <i>Importing Libraries</i>	7
3.1.3 <i>Importing the Datasets</i>	7
4 DATA ANALYTICS	8
4.1 ELASTIC SEARCH	8
4.2 KIBANA	8
4.3 VISUALIZATIONS USING KIBANA.....	9
4.3.1 <i>Adding Data to the Dashboard</i>	9
4.3.2 <i>Creating the First Dashboard</i>	9
4.3.3 <i>Open the visualization editor and get familiar with the data.</i>	10
4.3.4 <i>Create your first visualization</i>	11
4.3.5 <i>View a metric over time</i>	12
4.3.6 <i>View the top values of a field</i>	13
4.3.7 <i>Comparing the documents with others</i>	14
4.3.8 <i>Set the Time Range</i>	15
4.3.9 <i>Importo Values</i>	17
4.3.10 <i>Importo Values</i>	17
4.3.11 <i>Tenders based on Month</i>	18
4.3.12 <i>Importo Value Over Time</i>	19
4.3.13 <i>Average Importo Value</i>	20
4.3.14 <i>Performance Graph</i>	21
4.3.15 <i>Searching for Text in the Data</i>	22
4.3.16 <i>Dashboard and Visualization</i>	23
5 DATA CLEANING.....	25
5.1 ZONE:	25
5.2 PROCEDURA DI GARA:	25
5.3 ENTE APPALTANTE:.....	26
5.4 DATA INSERIMENTO AND SCADENZA:	27
5.5 FONTI:	27
5.6 CATEGORIE:	28
5.7 OGGETTO:	28
5.8 REMOVING NS VALUES	29
5.9 TOKENIZATION PROCESS	30
6 FEATURE ENGINEERING	32
6.1 WHAT IS FEATURE ENGINEERING	32
6.2 STEPS FOR FEATURE ENGINEERING.....	32
6.3 TECHNIQUES FOR FEATURE ENGINEERING	32
6.4 READING THE DATASET.	33
7 PREDICTING THE IMPORTO VALUES:	33
7.1 TEXT CLASSIFICATION:	34

7.2	TEXT SUMMARIZATION:	34
7.3	NAMED ENTITY RECOGNITION:	35
7.4	LABEL ENCODING.....	35
8	TEXT FEATURE EXTRACTION:.....	36
8.1	TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF):	36
8.2	AVERAGE WORD2VEC:.....	36
9	DATA MODELING	37
10	EVALUATION METRICS:.....	38
11	METHODOLOGIES	39
11.1	DECISION TREE.	39
	HOW IS SPLITTING DECIDED FOR DECISION TREES?	40
11.2	RANDOM FOREST REGRESSION.	40
11.3	KNN (K-NEAREST NEIGHBOUR).....	42
11.4	REGRESSION	44
11.5	DEEP LEARNING TECHNIQUES.....	45
11.5.1	<i>Artificial Neural Networks</i>	45
12	RESULTS FOR IMPORTO	47
12.1.	TF-IDF:	47
12.2	WORD2VEC:	51
13	COMPARISON OF DIFFERENT METHODS:	55
14	RESULTS OF TEXT CLASSIFICATION:.....	56
14.1	LOGISTIC REGRESSION:	56
14.2	DECISION TREE CLASSIFIER:	57
14.3	RANDOM FOREST:	59
14.4	KNN:.....	60
15	COMPARISON BASED ON TRAIN AND TEST ACCURACY:	63
16	RESULTS OF TEXT SUMMARIZATION:	64
17	RESULTS OF NAMED ENTITY RECOGNITION:	65
18	CONCLUSION	66
19	ACKNOWLEDGEMENT:	67
20	BIBLIOGRAPHY.....	ERROR! BOOKMARK NOT DEFINED.

Table of Figures

Figure 1 Adding Dataset to Kibana.....	9
Figure 2 Creating Dashboard.....	10
Figure 3 Visualization Editor.....	10
Figure 4 Visualization Types in Kibana.....	11
Figure 5 Count of Timestamp.....	12
Figure 6 Chart for Timestamp.....	13
Figure 7 Top Values Zones.....	14
Figure 8 Ranges of documents.....	15
Figure 9 Time Range	16
Figure 10 Table of Importo values.....	17
Figure 11 Maximum and Average value.....	18
Figure 12 Percentage of Tenders.....	19
Figure 13 Maximum Importo Value.....	20
Figure 14 Average Importo Value.....	21
Figure 15 Performance Graph.....	22
Figure 16 Mapping text fields.....	23
Figure 17 Text search.....	23
Figure 18 Dashboard visualization.....	24
Figure 19 Dashboard visualization.....	24
Figure 20 Unique values of Procedure di Gara	26
Figure 21 Unique values of Ente Appaltante.....	26
Figure 22 Values of Data Inserimento & Scadenza	27
Figure 23 One Hot Representation.....	28
Figure 24 Description of Oggetto	29
Figure 25 Importo with NS values.....	29
Figure 26 Importo without NS values.....	30
Figure 27 Feature engineering.....	32
Figure 28 Columns in Dataset.....	33
Figure 29 Process for Text summarization.....	35
Figure 30 Structure of Decision trees	39
Figure 31 Structure of Random Forest regression.....	41
Figure 32 KNN performance.....	43
Figure 33 Simple Regression.....	44
Figure 34 Deep hidden layers.....	45
Figure 35 Structure of ANN.....	45
Figure 36 Bar plot representation for Train MAE.....	48
Figure 37 Bar plot representation for Validation MAE.....	48
Figure 38 Bar plot representation for Train RMSE.....	49
Figure 39 Bar plot representation for Test RMSE.....	49
Figure 40 Bar plot representation for Train MAE.....	52
Figure 41 Bar plot representation for Validation MAE.....	53
Figure 42 Bar plot representation for Train RMSE.....	53
Figure 43 Bar plot representation for Train RMSE and Validation RMSE.....	55
Figure 44 Confusion matrix representation for Train and Test.....	56
Figure 45 ROC curve representation for Train and Test.....	57

Figure 46 Confusion matrix representation for Train and Test.....	58
Figure 47 ROC curve representation for Train and Test.	59
Figure 48 Confusion matrix representation for Train and Test.....	59
Figure 49 ROC curve representation for Train and Test.	60
Figure 50 Confusion matrix representation for Train and Test.....	61
Figure 51 ROC curve representation for Train and Test.	62

1 INTRODUCTION

1.1 About Iconsulting

Today, data is a key part of the process of making decisions. The new challenge is to make the data we use more intelligent and relevant. We have a method that uses data value and human potential to make the best decision in any situation.

This firm specialize in designing and delivering solutions, methodologies, algorithms, and technologies that can transform and empower client companies with the power of data, the leading and most abundant asset in the marketplace today. They are a driven change company.

Reality was born from a group of university researchers and today, they are a strategic partner to over 150 client companies and all the most important international technology vendors. They have over 1000 successful projects completed over 20 years of constant and continuous growth. With offices in Bologna, Rome, Milan and London, they boast a staff of 300 highly specialized professionals. Through consulting services, big data platform, integrated platforms, blockchain, business intelligence, location analytics, artificial intelligence and machine learning, product management and client data platforms, this firm supports all levels of client companies, giving shape and concreteness to their vision.

1.2 About Telemat

Who they are: Specialists for public tenders.

Telemat is a division of DB Information that has been supporting companies in the procurement world for 35 years. Since 1987, Telemat has been providing information on public tenders daily, obtained through continuous monitoring of official sources and websites of bodies, including central purchasing centers and electronic market platforms. Telemat is a dynamic and dynamic reality capable of providing new services to customers and continuous and personalized support for the growth of their business.

2 DATA DESCRIPTION

With the advent of the internet, public, semi-public all around the Italy have been publishing procurement documents on public platforms or websites. This resulted in a large body of valuable economic information. Despite the challenges of getting the right information at the right time, companies face many challenges in finding and using strategic information. Much of this information is hidden in a vast amount of unstructured or semi-structured documents.

2.1 Dataset description

The below table indicates the information about the dataset, which contains many useful (related) columns that are useful for the analysis and prediction. Every column has a specific description that are related to the tenders.

Rif. Bando	Data Inserimento	Scadenza	Importo (€)	Oneri (€)	Onorario (€)	Fonti	Ente Appaltante
13149385	14/04/2021	30/04/2021	905.000,00	NS	NS	ALBO	SALERNO ENERGIA HOLDING SPA, VIA STEFANO PASSARO N. 1, 84134 SALERNO (SA)
13116746	19/03/2021	26/04/2021	396.000,00	NS	NS	ALBO, CEE,	SORESA SOCIETA' REGIONALE PER LA SANITA' SPA DI NAPOLI, CENTRO DIREZIONALE ISOLA F9, 80143 NAPOLI
12857671	16/11/2020	17/12/2020	575.500,00	NS	NS	ALBO, CEE,	UNIVERSITA' DEGLI STUDI DI BARI ALDO MORO, PIAZZA UMBERTO I N. 1, 70121 BARI (BA)
13144973	14/04/2021	28/04/2021	250.000,00	NS	NS	ALBO, CEE,	UNIACQUE SPA, VIA DELLE CANOVINE N. 21, 24126 BERGAMO (BG)
13149583	14/04/2021	28/04/2021	24.000,00	NS	NS	ALBO	COMUNE DI ANCONA, PIAZZA XXIV MAGGIO N. 1, 60100 ANCONA (AN)
13143183	08/04/2021	23/04/2021	16.440.000,00	NS	NS	ALBO, CEE,	ARIA AZIENDA REGIONALE PER L'INNOVAZIONE E GLI ACQUISTI SPA, VIA TORQUATO TARAMELLI N. 26, 20124 MILANO (MI)
13149634	14/04/2021	23/04/2021	64.550,00	NS	NS	ALBO	PARCAM SRL - AZIENDA DELLA CAMERA DI COMMERCIO DI MILANO, VIA MERAVIGLI N.7, 20123 MILANO (MI)
13149763	14/04/2021	30/04/2021	174.590,16	NS	NS	ALBO	CNR CONSIGLIO NAZIONALE DELLE RICERCHE DI MESSINA - ITAE ISTITUTO DI TECNOLOGIE AVANZATE PER L'ENERGIA NICOLA GIORDANO
13149892	14/04/2021	27/04/2021	35.000,00	NS	NS	ALBO	ARTER ATTRATTIVITA' RICERCA TERRITORIO, VIA GOBETTI N. 101, 40129 BOLOGNA (BO)

3 DATA PREPROCESSING.

Data pre-processing is a process of preparing the raw data so that it can be used in a machine learning model. The first step in creating a machine learning model is selecting the right data. When working with machine learning, it's not always easy to find clean and well-formed data. When working with data, it is essential to clean it and put it into a formatted format. So, we use data pre-processing tasks to do this.

3.1 Why do we need Data Pre-processing?

Real-world data often contains noises, missing values, and an unusable format that cannot be directly used for machine learning models. Data pre-processing is necessary tasks that clean the data and make it ready for a machine learning model, which also improves the accuracy and efficiency of the machine learning model.

1. Getting the dataset
2. Importing libraries
3. Importing datasets
4. Finding Missing Data
5. Encoding Categorical Data
6. Splitting dataset into training and test set
7. Feature scaling

3.1.1 Get the Dataset

To create a machine learning model, we need a dataset to work with. The dataset on a particular problem is a collection of data arranged in a specific way. The data set may be in a different format for different purposes, such as if we are trying to create a machine learning model for business purposes, each dataset is unique. To use the dataset in our code, we usually save it to a CSV file. Sometimes, we may need to use an HTML or Excel file. CSV stands for "Comma Separated Values" files; it is a file format that allows us to save tabular data such as spreadsheets. The software is useful for large datasets and can be used to process these datasets in programs

3.1.2 Importing Libraries

To use data pre-processing features in Python, we need to import some pre-defined Python libraries. These libraries are specifically designed for performing specific jobs. I have used libraries such as Pandas, Matplotlib, and NumPy for data pre-processing.

3.1.3 Importing the Datasets

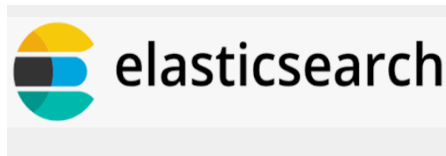
The datasets that we have gathered for our machine learning research must now be imported. However, we must first establish the current directory as the working directory before importing a dataset.

The dataset has been loaded. The information below describes how the dataset was pre-processed. The datasets that we have gathered for our machine learning research must now be imported.

This dataset has a total of 6630 rows * 18 columns of entries, according to the result.

4 DATA ANALYTICS

ELASTIC SEARCH AND KIBANA



4.1 ELASTIC SEARCH

Elasticsearch is a distributed, free and open search and analytics engine for all types of data, including textual, numerical, geospatial, structured, and unstructured. Elasticsearch is built on Apache Lucene and was first released in 2010 by Elasticsearch N.V. (now known as Elastic). Known for its simple REST APIs, distributed nature, speed, and scalability, Elasticsearch is the central component of the Elastic Stack, a set of free and open tools for data ingestion, enrichment, storage, analysis, and visualization. Commonly referred to as the ELK Stack (after Elasticsearch, Logstash, and Kibana), the Elastic Stack now includes a rich collection of lightweight shipping agents known as Beats for sending data to Elasticsearch. (1)

The best features of Elastic Search are

- Scalability and Resiliency
- Clustering and High Availability
- Automatic Node Recovery
- Automatic Node Rebalancing

4.2 KIBANA

Kibana is a free and open frontend application that sits on top of the Elastic Stack, providing search and data visualization capabilities for data indexed in Elasticsearch. Commonly known as the charting tool for the Elastic Stack (previously referred to as the ELK Stack after Elasticsearch, Logstash, and Kibana), Kibana also acts as the user interface for monitoring, managing, and securing an Elastic Stack cluster — as well as the centralized hub for built-in solutions developed on the Elastic Stack. Developed in 2013 from within the Elasticsearch community, Kibana has grown to become the window into the Elastic Stack itself, offering a portal for users and companies. (2)

In Kibana, users can:

- Grant all access to Dashboard at an individual space
- Grant all access to one space and read access to another
- Grant read access to all spaces and write access to an individual space

4.3 VISUALIZATIONS USING KIBANA

As, our data set contains 18 columns, all the columns have unique features/values. In order to understand the dataset well, we use Elastic Search and Kibana for good visualizations. As, our main goal is to predict and analyze the data, we compare different columns together and understand the importance of them.

Data visualization is the most effective technique to comprehend it. With dashboards, you can organize your data into a series of panels that provide clarity, tell a strong story about your data, and let you concentrate just on the information that matters to you.

Panels provide your data in charts, tables, maps, and other formats that enable side-by-side comparison of your data to spot trends and connections. To display your data, dashboards enable a variety of panel kinds and panel creation options.

4.3.1 Adding Data to the Dashboard

Add the sample web logs data and create and set up the dashboard.

- Go to the Home page, then click Try sample data.
- On the Sample web logs card, click Add data.

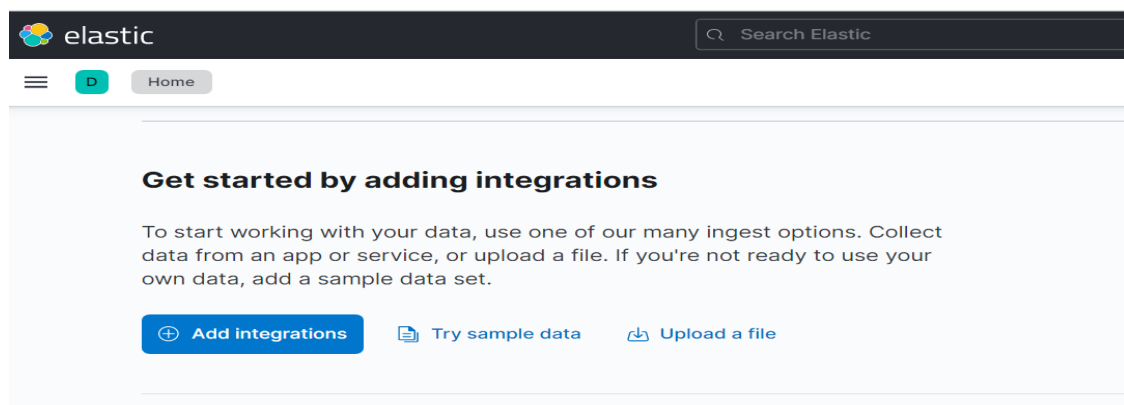


Figure 1 Adding Dataset to Kibana.

4.3.2 Creating the First Dashboard

Create the dashboard where you'll display the visualization panels.

- Open the main menu, then click Dashboard.

- Click Create dashboard.
- Set the time filter to Last 90 days.

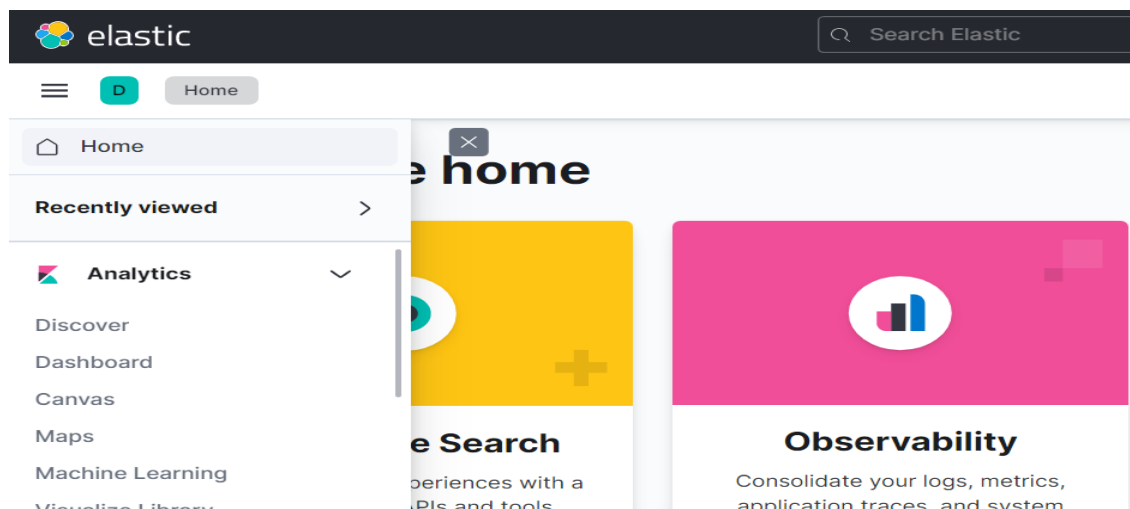


Figure 2 Creating Dashboard.

4.3.3 Open the visualization editor and get familiar with the data.

Open the visualization editor, then make sure the correct fields appear.

- On the dashboard, click Create visualization.
- Make sure the kibana_sample_data_logs data view appears.

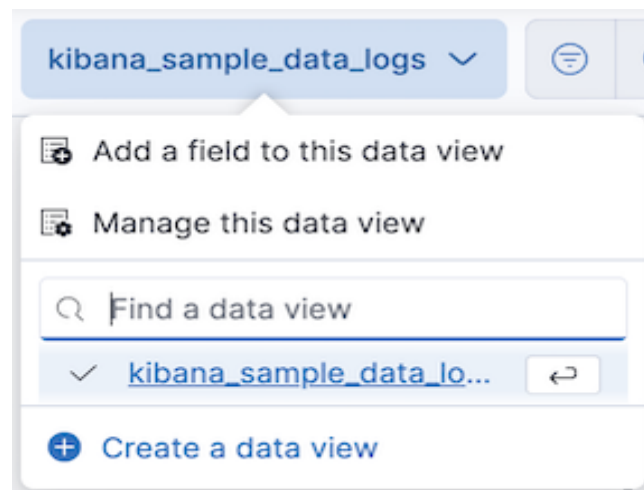


Figure 3 Visualization Editor.

To create the visualizations, you will be using the following fields:

- **Timestamp**
- **Articoli**
- **Data Inserimento**

- Durata
- Ente Appaltante
- Fonti
- Importo
- Oggetto
- Oneri
- Onorario
- Procedura di Gara
- Rif_Bando
- Scadenza
- Zone

To see the most frequent values in a field, hover over the field name, then click **i**.

4.3.4 Create your first visualization

Select a field for analysis, like timestamp. Use the Metric visualization to display the information as a number to analyze solely the Timestamp field. Only the number function known as Unique count, also known as cardinality, which roughly approximates the number of unique values, can be used with Timestamp.

- Open the Visualization type dropdown, then select Metric.

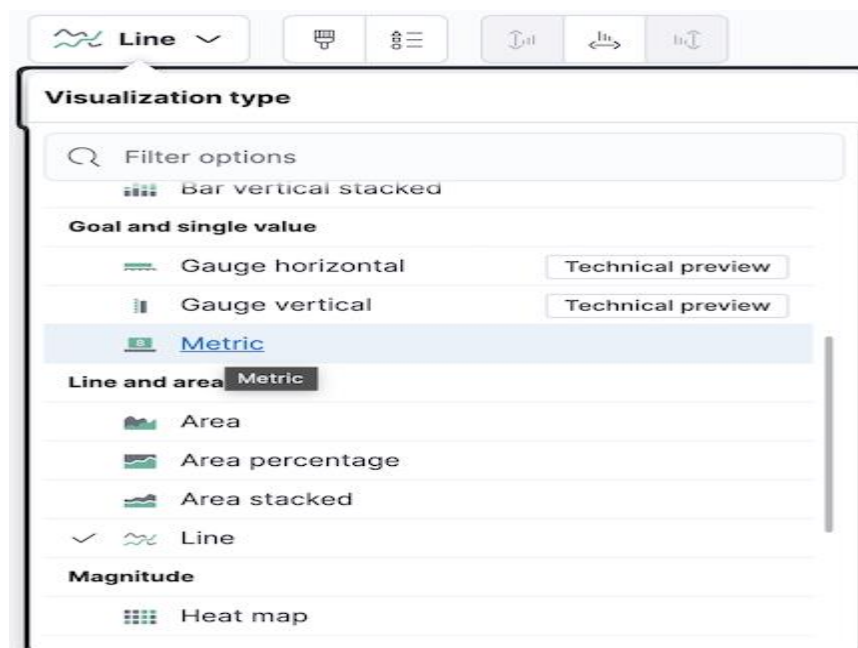


Figure 4 Visualization Types in Kibana.

- From the Available fields list, drag Timestamp to the workspace or layer pane.

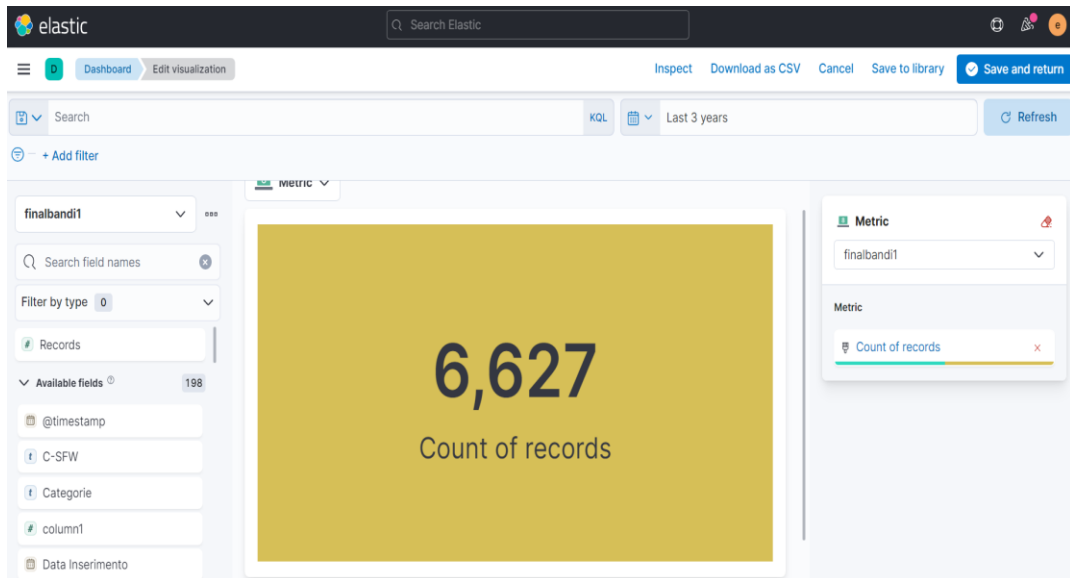


Figure 5 Count of Timestamp.

Because the editor automatically applies the Unique count function to the Timestamp field, Unique count of Timestamp displays in the layer pane. The only mathematical operation that functions with IP addresses is unique count.

- In the layer pane, click Unique count of Timestamp.
- In the Name field, enter Count of Records.
- Click Close.
- Click Save and return.

4.3.5 View a metric over time

There are two shortcuts you can use to view metrics over time. When you drag a numeric field to the workspace, the visualization editor adds the default time field from the data view. When you use the Date histogram function, you can replace the time field by dragging the field to the workspace.

To visualize the bytes field over time:

- On the dashboard, click Create visualization.
- From the Available fields list, drag bytes to the workspace.

The visualization editor creates a bar chart with the timestamp and Median of bytes fields.

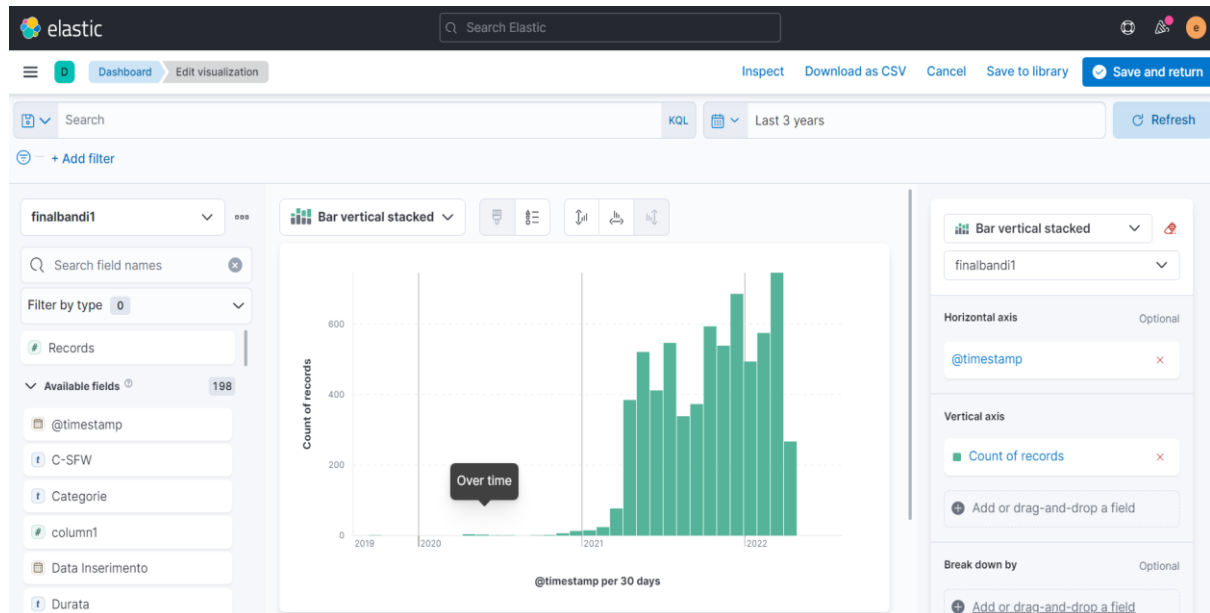


Figure 6 Chart for Timestamp.

4.3.6 View the top values of a field

Create a visualization that displays the most frequent values. To create the visualization, use Top values of zone ranked by unique count of records.

The Top values function ranks the unique values of a field by another function. The values are the most frequent when ranked by a Count function, and the largest when ranked by the Sum function.

- On the dashboard, click create visualization.
- From the Available fields list, drag zone to the vertical axis field in the layer pane.

The visualization editor automatically applies the Unique count function. If you drag zone to the workspace, the editor adds the field to the axis.

- Drag zone field to the workspace. (3)

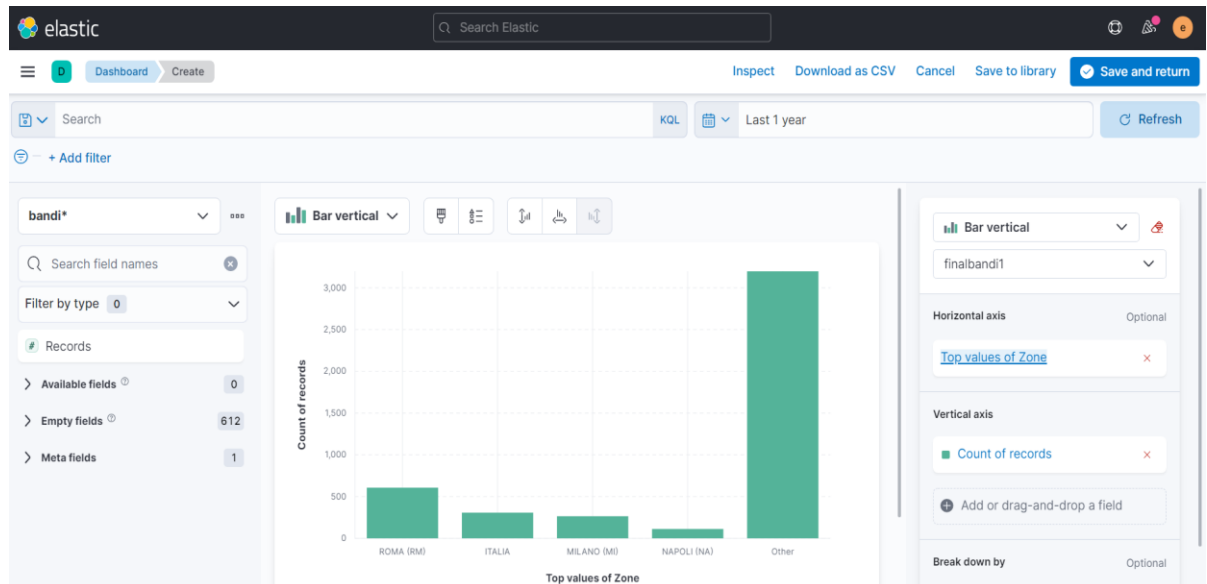


Figure 7 Top Values Zones.

4.3.7 Comparing the documents with others

- Click Create visualization on the dashboard.
- Drag bytes to the Vertical axis field in the layer pane from the list of available fields.
- Click Median of Bytes in the layer pane.
- After selecting the Sum option, click Close.
- Drag bytes to the Break down by field in the layer pane from the Available fields list.

Use the Intervals function to choose documents based on a field's number range. You might use the Filters function if the query has many clauses, or the ranges are not numeric. Specify the file size ranges:

- Click bytes under the layer pane.
- Select Create custom ranges; then, in the Ranges field, type the following information and hit Return:
 Ranges – 0 → 10240
 Label – Below 10KB
- Click Add range, then press Return.

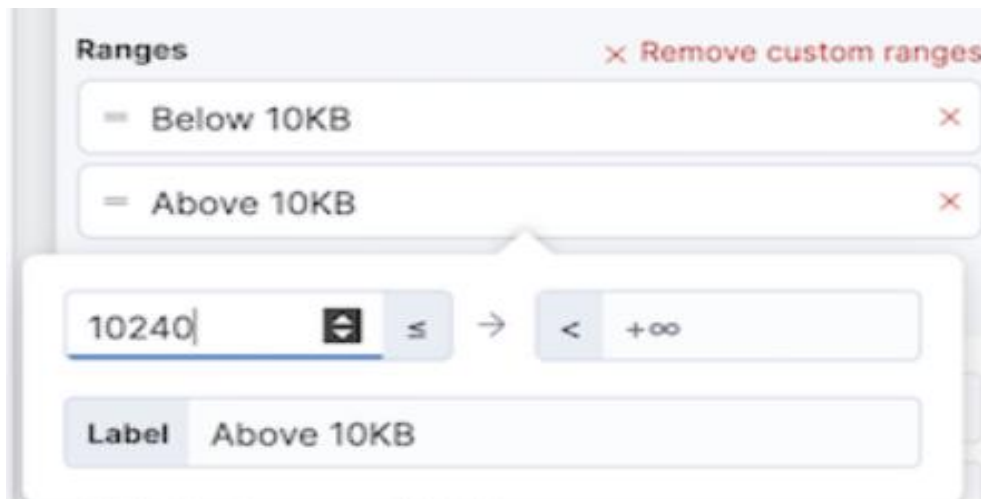


Figure 8 Ranges of documents.

4.3.8 Set the Time Range

When your index contains time-based events and a time-field is set up for the chosen data view, display data for the specified period. In Advanced Settings, you can change the default time range of 15 minutes.

- Press the calendar icon.

Select from the following:

- Decide on a time based on the previous or upcoming number of seconds, minutes, hours, or another time unit.
- Choose a period from the Last 15 minutes, Today, and Week to date selections, dates that have recently been utilized. Use a set of data that you've already chosen.
- Specify an automatic refresh rate.

Last 3 years

Quick select

Last

3

years

Apply

Commonly used

Today

This week

Last 15 minutes

Last 30 minutes

Last 1 hour

Last 24 hours

Last 7 days

Last 30 days

Last 90 days

Last 1 year

Recently used date ranges

Last 3 years

Last 1 year

Aug 29, 2020 @ 17:20:51.544 to Dec 23, 2020 @ 12:23:27.083

Mar 21, 2017 @ 00:00:00.000 to Apr 14, 2022 @ 00:00:00.000

☐ Refresh every

0

seconds

- Click the bar next to the time filter to set the start and end times. Select Absolute, Relative, or Now in the popup, then enter the necessary information.

~ 4 days ago

→

~ in 10 hours

5 m

Refresh

Absolute

Relative

Now

4

Days ago

Start date

Jun 3, 2022 @ 00:00:00.000

☒

Round to the day

Figure 9 Time Range

4.3.9 Importo Values

- We can generate visualizations by selecting the Visualize option from the Kibana Home screen.
- Next, we choose the Data Table option from the range of possible visualization options.
- In this figure we are displaying the top values of import based on the column procedura di gara.
- This table displays the maximum import value of the tender based on the time that we have selected as per the requirements.



Top values of Procedura di Gara	Maximum of Importo
AVVISO DI INFORMAZIONE PRELIMINARE	3,300,000,000
PROCEDURA RISTRETTA	3,000,000,000
PROCEDURA APERTA	723,300,000
CONSULTAZIONE PRELIMINARE DI MERCATO	162,000,000
INDAGINE DI MERCATO	52,000,000
AVVISO DI MANIFESTAZIONE D'INTERESSE	35,000,000
PROCEDURA NEGOZIATA	15,000,000

Figure 10 Table of Importo values.

4.3.10 Importo Values

- Select a field for analysis, like Importo. Use the Metric visualization to display the information as a number to analyze solely the Importo field.
- To visualize the Maximum and Average Values of Importo which displays the tender values, can be used with Importo.

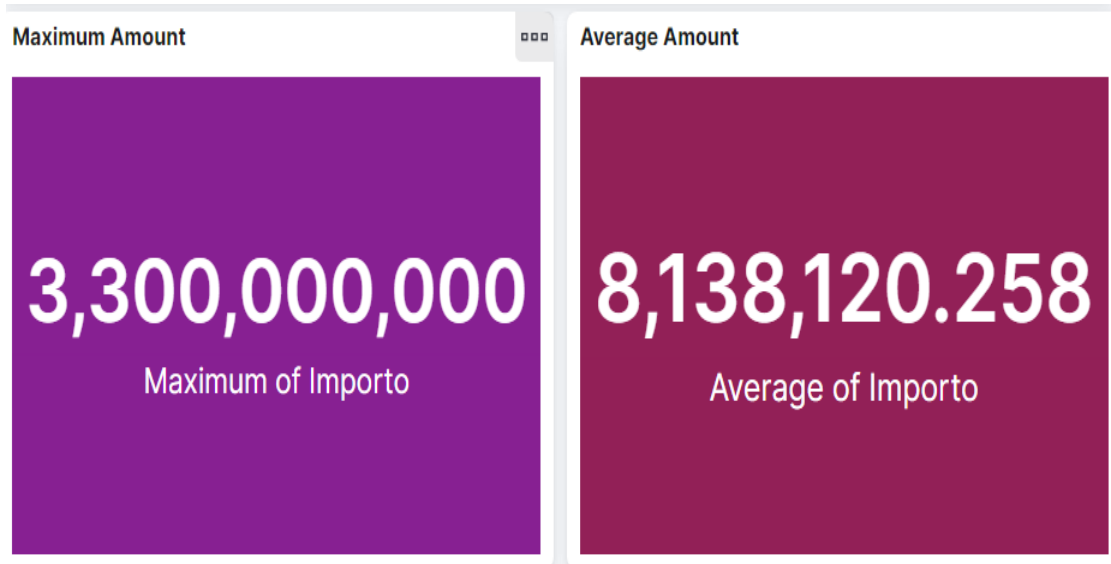


Figure 11 Maximum and Average value

4.3.11 Tenders based on Month

To display the values as a percentage of the sum of all values, use the **Pie** chart.

- Open the Visualization Type dropdown.
- Click the Create a Visualization button.
- Select the Pie chart.
- By default, a pie chart with just one bucket will be produced.
- In the pie chart editor, configure the Metrics.
- Open Options tab and then follow the steps.
- Check the Show Label box from the Label settings.
- Click the right arrow on the top of the tool bar to update the changes.
- Click Save and return.

In the below pie chart, we are displaying the results of tenders based on the zone. It displays the percentage of each tender with respect to zones or regions thar are present in the dataset. (3)

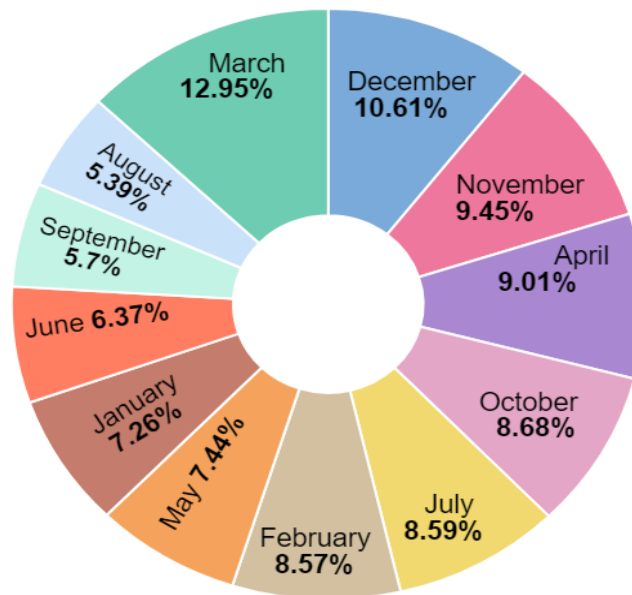


Figure 12 Percentage of Tenders.

4.3.12 Importo Value Over Time

Create a line chart that shows the maximum amount of tender value, then add a line chart layer that shows the number of tenders to analyze several visualization kinds.

- On the dashboard, click Create visualization.
- From the Available fields list, drag Importo to the workspace.
- In the layer pane, click Maximum of Importo.
- Click the Maximum function.
- In the Name field, enter desired name, then click Close.
- Open the Visualization type dropdown, then select Area.



- The visualization editor by default shows time series data with stacked charts that illustrate how the various document sets alter over time.

- This line graph is used to display the maximum number of tender values.

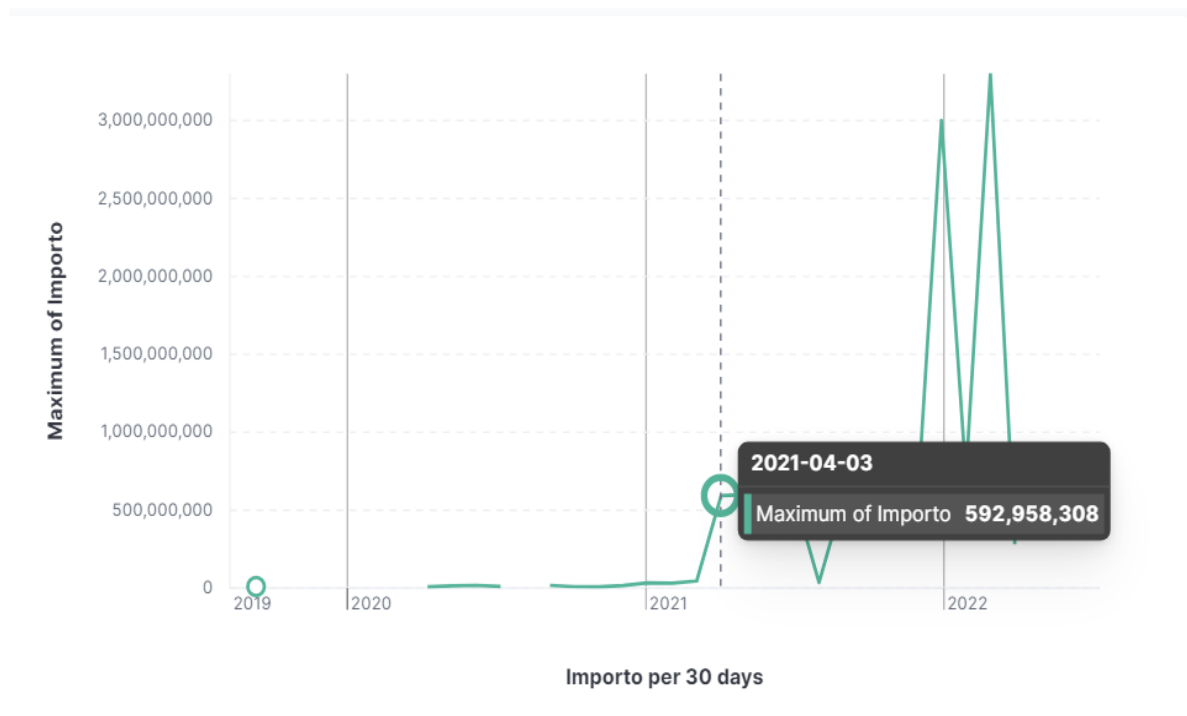


Figure 13 Maximum Importo Value.

4.3.13 Average Importo Value

We select the option for Vertical Bar under New Visualization on the Kibana Home screen to create the vertical bar chart.

- Select the Visualize tab from the left menu bar
- Click the Create a Visualization button
- Select the Vertical Bar chart



- The default configuration will produce a bar chart with just one bucket.

- This graph displays the average amount of Importo value based on the time.

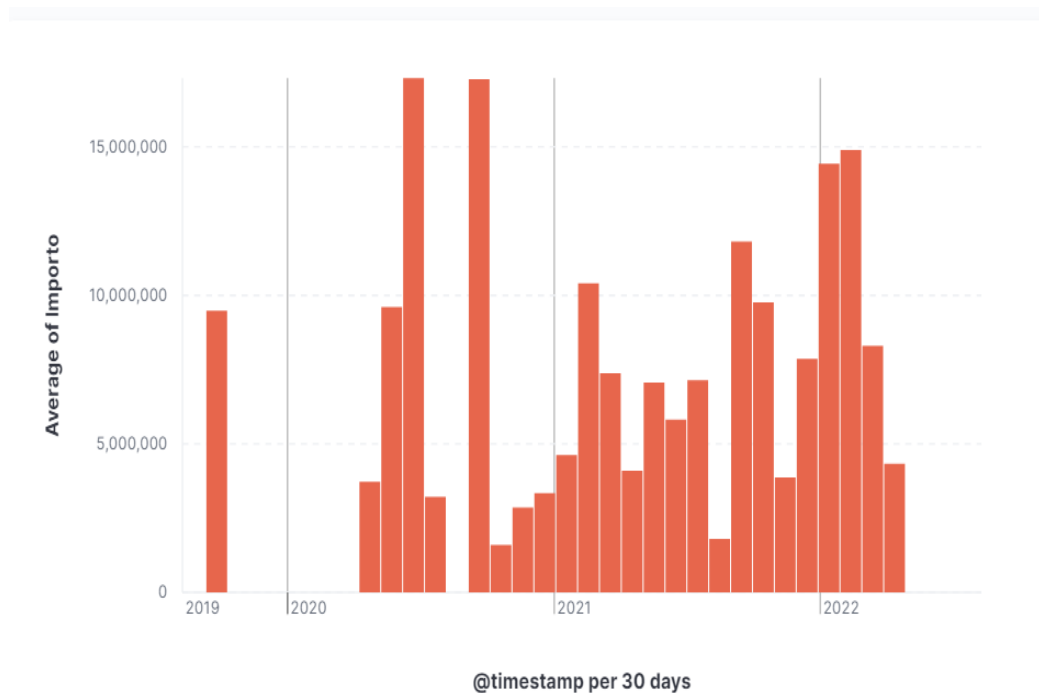


Figure 14 Average Importo Value.

4.3.14 Performance Graph

Create a line chart that shows the maximum amount of tender value which displays the performance, then add a line chart layer that shows the number of tenders based on each month.

- On the dashboard, click Create visualization.
- From the Available fields list, drag Importo to the workspace.
- In the layer pane, click Maximum of Importo.
- Click the Maximum function.
- Open the Visualization type dropdown, then select Area
- To zoom in on the data, click and drag your cursor across the bars. (3)



Figure 15 Performance Graph.

4.3.15 Searching for Text in the Data

The text family includes the following field types:

- `Text`, the traditional field type for full-text content such as the body of an email or the description of a product.
- `match_only_text`, a space-optimized variant of `text` that disables scoring and performs slower on queries that need positions. It is best suited for indexing log messages.
- A field to index full-text values, such as the body of an email or the description of a product.
- These fields are analyzed, that is they are passed through an analyzer to convert the string into a list of individual terms before being indexed.
- The analysis process allows Elasticsearch to search for individual words within each full text field.
- Text fields are not used for sorting and used for aggregations.
- Sometimes it is useful to have both a full text (`text`) and a keyword (`keyword`) version of the same field: one for full text search and the other for aggregations and sorting.
- This can be achieved with multi-fields mapping. (4)

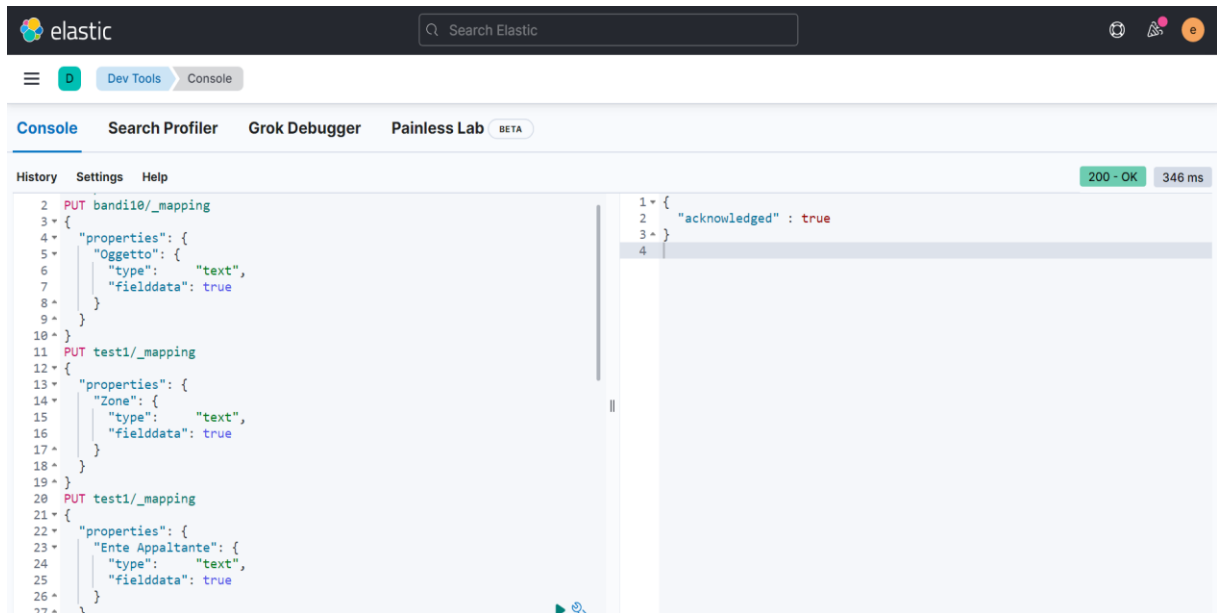


Figure 16 Mapping text fields.

- In Kibana, we cannot be able to show the text data using any visualization charts, so we are using text fields to search for keywords or text. The below image shows in which document the search word appears.

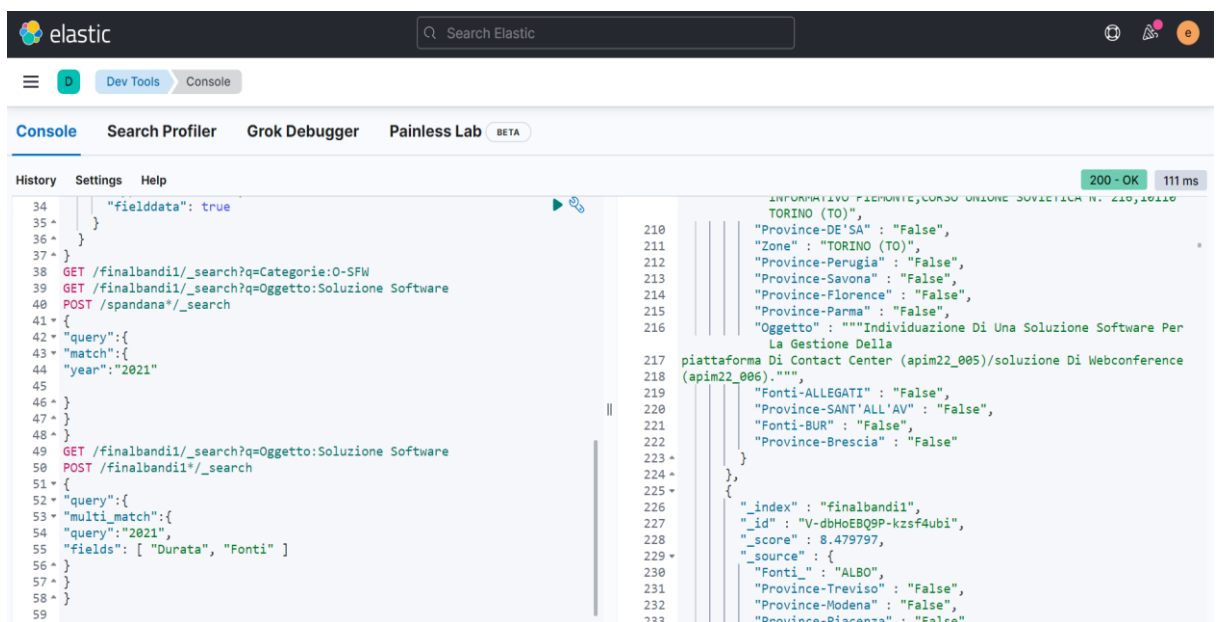


Figure 17 Text search.

4.3.16 Dashboard and Visualization

The best way to understand your data is to visualize it. With dashboards, you can turn your data from one or more data views into a collection of panels that bring clarity to your data, tell a story about your data, and allow you to focus on only the data that's important to you.

Panels display your data in charts, tables, maps, and more, which allow you to compare your data side-by-side to identify patterns and connections. Dashboards support several types of panels to display your data, and several options to create panels. (5)

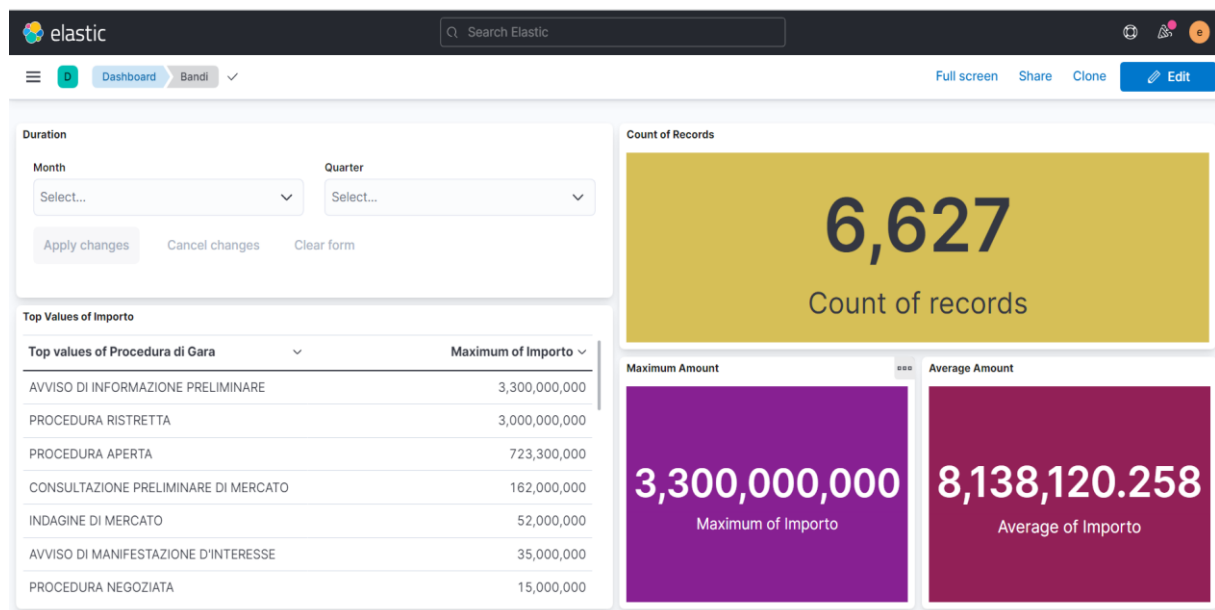


Figure 18 Dashboard visualization.

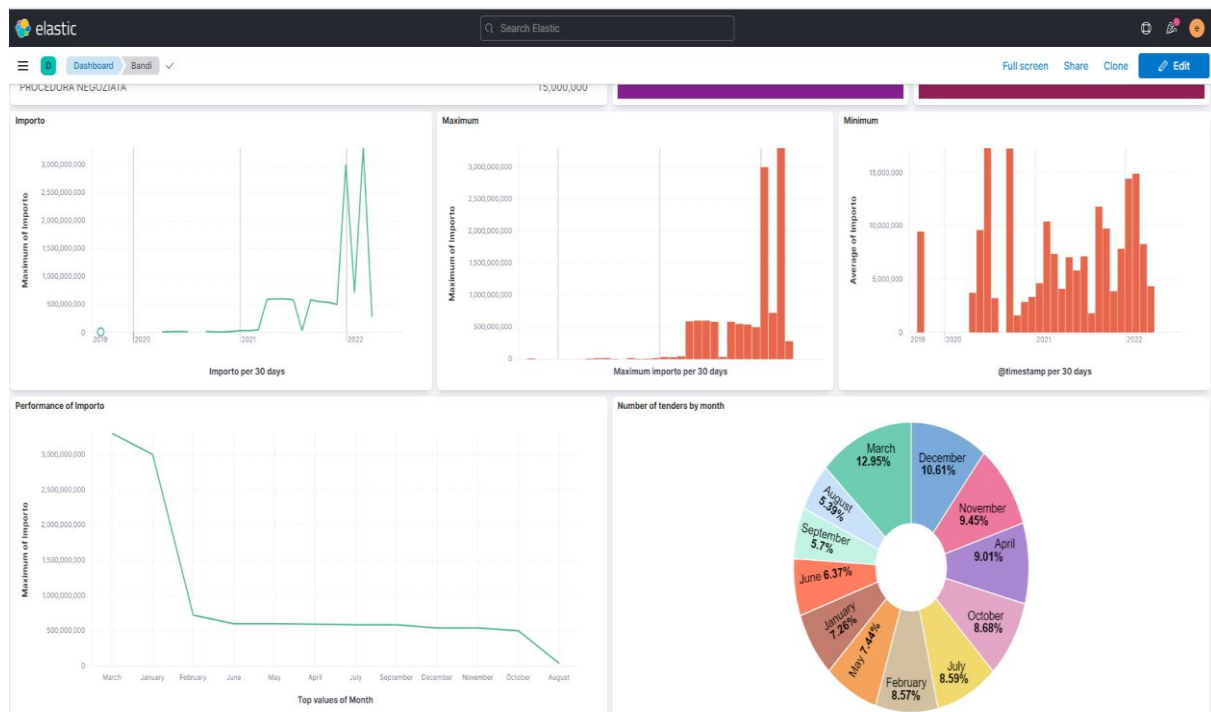


Figure 19 Dashboard visualization.

5 DATA CLEANING

5.1 Zone:

In the dataset we are having so many zones where we cannot be able to use them to visualize. So, we are mapping all the zones to regions to familiar with them. Conversion of zone to regions and total 21 regions identified including (ITALIA) and 21 values are represented in one hot representation.

```
['Lombardy',  
'Tuscany',  
'Emilia-Romagna',  
'Sicily',  
'Piedmont',  
'Sardinia',  
'Veneto',  
'Apulia',  
'Calabria',  
'Campania',  
'Marche',  
'Lazio',  
'Abruzzo',  
'Friuli-Venezia Giulia',  
'Liguria',  
'Umbria',  
'Basilicata',  
'Trentino-South Tyrol',  
'Molise',  
'Aosta Valley',  
'ITALIA']
```

5.2 Procedura di Gara:

- This column represents the tender procedure.
- This column is represented as a label encoder to convert each category into numerical representation because machine learning algorithms can understand only numbers.
- Here we are displaying all the unique values that are present in the Procedura di Gara column.

PROCEDURA APERTA	2956
AVVISO DI MANIFESTAZIONE D'INTERESSE	1364
INDAGINE DI MERCATO	982
CONSULTAZIONE PRELIMINARE DI MERCATO	423
NS	188
ALBO FORNITORI	170
PROCEDURA NEGOZIATA	123
PROCEDURA TELEMATICA	79
AVVISO DI INFORMAZIONE PRELIMINARE	74
PROCEDURA COMPARATIVA	70
AFFIDAMENTO DIRETTO	68
PROCEDURA RISTRETTA	52
PROCEDURA SELETTIVA	17
DIALOGO COMPETITIVO	12
SISTEMA DI QUALIFICAZIONE	12
CONCORSO DI IDEE	11
BANDO DI FINANZIAMENTO	8
AVVISO INFORMATIVO PERIODICO	5
AVVISO DI TRASPARENZA EX ANTE	5
ASTA PUBBLICA	4
GARA AD EVIDENZA PUBBLICA	2
PROCEDURA COMPETITIVA CON NEGOZIAZIONE	2
PARTENARIATO PER L'INNOVAZIONE	1
DIALOGO TECNICO	1
BANDO DI GARA INDICATIVO	1

Figure 20 Unique values of Procedure di Gara

5.3 Ente Appaltante:

- This column represents the authority of the contract.
- It can only interpret numbers; a label encoder is used to this column to turn each category into a numerical representation.
-

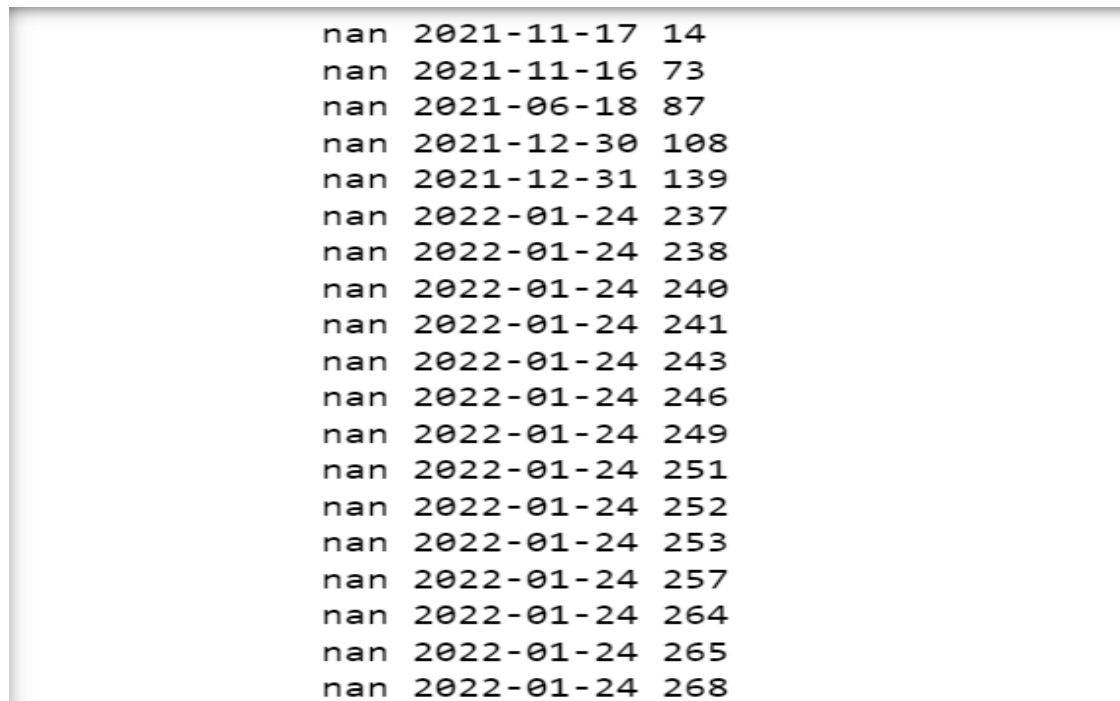
CONSIP SPA, VIA ISONZO N. 19/E, 00198 ROMA (RM)	159
RAI RADIOTELEVISIONE ITALIANA SPA - DIREZIONE ACQUISTI E SERVIZI, VIALE MAZZINI N. 14, 00195 ROMA (RM)	97
ARIA AZIENDA REGIONALE PER L'INNOVAZIONE E GLI ACQUISTI SPA, VIA TORQUATO TARAMELLI N. 26, 20124 MILANO (MI)	85
TERNA SPA DI ROMA, VIALE EGIDIO GALBANI N. 70, 00156 ROMA (RM)	77
ASST AZIENDA SOCIO SANITARIA TERRITORIALE DI MANTOVA, STRADA LARGO PAIOLO N. 10, 46100 MANTOVA (MN)	63
...	
ENTE CAPITULO CATTEDRALE DI SAN PIETRO, PIAZZA DUOMO N. 1, 76011 BISCEGLIE (BT)	1
COMUNE DI MASCALUCIA, PIAZZA LEONARDO DA VINCI , 95030 MASCALUCIA (CT)	1
SCAPIGLIATO SRL DI ROSIGNANO MARITTIMO, LOC. SCAPIGLIATO - SR 206 KM 16,5, 57016 ROSIGNANO MARITTIMO (LI)	1
REGIONE PIEMONTE DI TORINO, PIAZZA CASTELLO N. 165, 10100 TORINO (TO)	1
COMUNE DI SAINT CHRISTOPHE, LOCALITA' CHEF LIEU N. 11, 11020 SAINT CHRISTOPHE (AO)	1

Name: Ente Appaltante, Length: 2171, dtype: int64

Figure 21 Unique values of Ente Appaltante.

5.4 Data Inserimento and Scadenza:

- From the start date and end date columns timeline can be extracted using datetime module in python and one feature is extracted i.e., timeline from 2 raw date features.



nan	2021-11-17	14
nan	2021-11-16	73
nan	2021-06-18	87
nan	2021-12-30	108
nan	2021-12-31	139
nan	2022-01-24	237
nan	2022-01-24	238
nan	2022-01-24	240
nan	2022-01-24	241
nan	2022-01-24	243
nan	2022-01-24	246
nan	2022-01-24	249
nan	2022-01-24	251
nan	2022-01-24	252
nan	2022-01-24	253
nan	2022-01-24	257
nan	2022-01-24	264
nan	2022-01-24	265
nan	2022-01-24	268

Figure 22 Values of Data Inserimento & Scadenza.

5.5 Fonti:

Extracted unique Fonti present in the data and represented the column in one hot because of a smaller number of categories

```
['QUOT',  
'RETTIFICA',  
'BUR',  
'ALBO',  
'ALLEGATI',  
'GURI',  
'CEE',  
'ESITO',  
'GURS',  
'RET_ALBO']
```


['ALBO']	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0],
['ALBO']	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0],
['ALBO', 'CEE', 'GURI', 'RET_ALBO']	[0, 0, 0, 1, 0, 1, 1, 0, 0, 1],
['ALBO', 'CEE', 'GURI']	[0, 0, 0, 1, 0, 1, 1, 0, 0, 0],
['ALBO']	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0],
['ALBO']	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0],
['ALBO']	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0],
['ALBO', 'CEE']	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0],
['ALBO', 'CEE', 'GURI']	[0, 0, 0, 1, 0, 0, 1, 0, 0, 0],
['ALBO']	[0, 0, 0, 1, 0, 1, 1, 0, 0, 0],
['ALBO']	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0],
['ALBO', 'CEE', 'GURI']	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0],
['ALBO', 'CEE']	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0],
['ALBO']	[0, 0, 0, 1, 0, 1, 1, 0, 0, 0],
['ALBO', 'CEE', 'GURI', 'RET_ALBO']	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0],
['CEE']	[0, 0, 0, 1, 0, 1, 1, 0, 0, 1],
['ALBO', 'RET_ALBO']	[0, 0, 0, 0, 0, 0, 0, 1, 0, 0],
['CEE']	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1],
	[0, 0, 0, 0, 0, 0, 0, 1, 0, 0],

Figure 23 One Hot Representation.

5.6 Categorie:

In the categorie column we are having so many unique values, so we are considering only top 5 important categories as per the requirements.

Top five Categories:

['O-SFW',
'O-MSW',
'O-GEST',
'O-SIC',
'C-SFW']

These categories are represented in one hot representation

5.7 Oggetto:

- Oggetto is a text field where it describes about each tender.
- This text field is transformed into tokens using text-to-vector conversion techniques such as word2vec and TFIDF.
- These tokens are then represented as numerical vectors that machine learning and deep learning algorithms can understand and learn patterns.
- The feature extraction process from the oggetto column is explained in detail.

```

0      Fornitura Di Software Gestionale Per Mmg Assoc...
1      S21023 - Procedura Aperta Per La Conclusione D...
2      Id Sintel 146749727. Gara Aria_2021_21 - Siste...
3      Progetto Per Una Cro Da Utilizzare Per Servizi...
4      Fornitura, Posa In Opera E Montaggio Di Attrez...

...

6625   Fornitura In Saas Di Una Soluzione Informatica...
6626   Id Sintel:152235562 Indagine Di Mercato Per L'...
6627   Id Sintel:152196130 - Gara 8488395 - Fornitura...
6628   Indagine Di Mercato/avviso Esplorativo Per La ...
6629   Avviso Pubblico Esplorativo Di Manifestazione ...
Name: Oggetto, Length: 6630, dtype: object

```

Figure 24 Description of Oggetto

5.8 Removing NS Values

- Many non- information (NS) values are present in the Importo field. Summing up all the NS's would give me the resultant output.

Unnamed: 0	Rif. Bando	Data Inserimento	Scadenza	Importo (€)	Oneri (€)	Onorario (€)	Fonti	Ente Appaltante	Categorie	Zone	Oggetto
0	0	13350922	2021-11-23 00:00:00	12000	NS	NS	ALBO	ASL AZIENDA SANITARIA LOCALE 02 LANCIANO VASTO...	T-SANITA, O-SFW	CHIETI (CH)	Fornitura Di Software Gestionale Per Mmg Assoc...
1	1	13351176	2021-12-17 00:00:00	1498598	3648	NS	ALBO	COMUNE DI BARI - RIPARTIZIONE STAZIONE UNICA A...	O-SFW, O-MSW, O-GEST	BARI (BA)	S21023 - Procedura Aperta Per La Conclusione D...
2	2	13325837	2021-11-29 00:00:00	430000	NS	NS	ALBO, CEE, GURI, RET_ALBO	ARIA AZIENDA REGIONALE PER L'INNOVAZIONE E GLI...	O-SIC, O-SFW	MILANO (MI)	Id Sintel 146749727. Gara Aria_2021_21 - Siste...
3	3	13341901	2021-12-06 00:00:00	1000000	NS	1000000	ALBO, CEE, GURI	IRCCS ISTITUTO NAZIONALE PER LO STUDIO E LA CU...	T-SANITA, O-SFW, O-MSW, O-GEST, F-SEG, F-ASSO,...	NAPOLI (NA)	Progetto Per Una Cro Da Utilizzare Per Servizi...
4	4	13349952	2021-12-01 00:00:00	138520	300	NS	ALBO	UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II,...	O-SFW, O-ELAB, O-APPA	NAPOLI (NA)	Fornitura, Posa In Opera E Montaggio Di Attrez...

Figure 25 Importo with NS values.

- All the (NS) values are replaced with NaN.

Unnamed: 0	Rif. Bando	Data Inserimento	Scadenza	Importo (€)	Oneri (€)	Onorario (€)	Fonti	Ente Appaltante	Categorie	...	Abruzzo	Friuli-Venezia Giulia	Liguria	Umbria	Basilic
0	0	13350922	2021-11-17	2021-11-23 00:00:00	12000	NaN	NaN	ALBO	153	T-SANITA, O-SFW	1	0	0	0	
1	1	13351176	2021-11-17	2021-12-17 00:00:00	1498598	3648	NaN	ALBO	638	O-SFW, O-MSW, O-GEST	0	0	0	0	
2	2	13325837	2021-10-22	2021-11-29 00:00:00	430000	NaN	NaN	ALBO, CEE, GURI, RET_ALBO	126	O-SIC, O-SFW	0	0	0	0	
3	3	13341901	2021-11-17	2021-12-06 00:00:00	1000000	NaN	1000000	ALBO, CEE, GURI	1512	T-SANITA, O-SFW, O-MSW, O-GEST, F-SEG, F-ASSO,...	0	0	0	0	
4	4	13349952	2021-11-17	2021-12-01 00:00:00	138520	300	NaN	ALBO	2075	O-SFW, O-ELAB, O-APPA	0	0	0	0	

Figure 26 Importo without NS values.

5.9 Tokenization process

- **Removing HTML tags**

Removing the html tags from the statement is the first stage in the tokenization process because html tags can be appended to text when site scraping, which is how this information was obtained.

- **Removing Links - (https:/)**

In this step html links don't receive any text-related information, the words will be eliminated.

- **Removing Words having numbers**

Remove words like usernames (John1) and email addresses if they contain text with digits because they don't provide any information.

- **Removing Special characters (@, \$)**

Special characters like #, @, and \$ should be removed because they don't convey any semantic information.

- **Removing Stop words:**

When stop words are eliminated from sentences, the vocabulary size will decrease, and it will be easier to spot sentence patterns because stop words will be used less frequently and have less informational value. Some words are only used to connect words they contain no information about the context of the sentence.

- **Italian Stop Words in NLTK library:**

“ ad, al, allo, ai, agli, all, agl, alla, alle, con, col, coi, da, dal, dallo, dai, dagli, dall, dagl, dalla, dalle, di, del, dello, dei, degli, dell, degl, della, delle, in, nel, nello, nei, negli, nell, negl, nella, nelle, su, sul, sullo, sui, sugli, sull, sugl, sulla, sulle, per, tra, contro, io, tu, lui, lei, noi, voi, loro, mio, mia, miei, mie, tuo, tua, tuoi, tue, suo, sua, suoi, sue, nostro, nostra, nostri, nostre, vostro, vostra, vostri, vostre, mi, ti, ci, vi, lo, la, li, le, gli, ne, il, un, uno, una, ma, ed, se, perché, anche, come, dov, dove, che, chi, cui, non, più, quale, quanto, quanti, quanta, quante, quello, quelli, quella, quelle, questo, questi, questa, queste, sì, tutto, tutti, a, c, e, i, l, o, ho, hai, ha, abbiamo, avete, hanno, abbia, abbiate, abbiano, avrò, avrai, avrà, avremo, avrete, avranno, avrei, avresti, avrebbe, avremmo, avreste, avrebbero, avevo, avevi, aveva, avevamo, avevate, avevano, ebbi, avesti, ebbe, avemmo, aveste, ebbero, avessi, avesse, avessimo, avessero, avendo, avuto, avuta, avuti, avute, sono, sei, è, siamo, siete, sia, siate, siano, sarò, sarai, sarà, saremo, sarete, saranno, sarei, saresti, sarebbe, saremmo, sareste, sarebbero, ero, eri, era, eravamo, eravate, erano, fui, fosti, fu, fummo, foste, furono, fossi, fosse, fossimo, fossero, essendo, faccio, fai, facciamo, fanno, faccia, facciate, facciano, farò, farai, farà, faremo, farete, faranno, farei, faresti, farebbe, faremmo, fareste, farebbero, facevo, facevi, faceva, facevamo, facevate, facevano, feci, facesti, fece, facemmo, faceste, fecero, facessi, facesse, facessimo, facessero, facendo, sto, stai, sta, stiamo, stanno, stia, stiate, stiano, starò, starai, starà, staremo, starete, staranno, starei, staresti, starebbe, staremmo, stareste, starebbero, stavo, stavi, stava, stavamo, stavate, stavano, stetti, stesti, stette, stemmo, steste, stettero, stessi, stesse, stessimo, stessero, stando”

6 Feature Engineering

All machine learning algorithms typically use input data to produce output, which is known as a feature. The input data continues to be presented in a tabular format with rows denoting instances or observations and columns denoting variables or attributes; these attributes are frequently referred to as features issue. (6)

6.1 What is Feature Engineering

Feature engineering is a machine learning technique that leverages data to create new variables that aren't in the training set. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy.

Feature engineering is required when working with machine learning models. Regardless of the data or architecture, a terrible feature will have a direct impact on your model. (7)

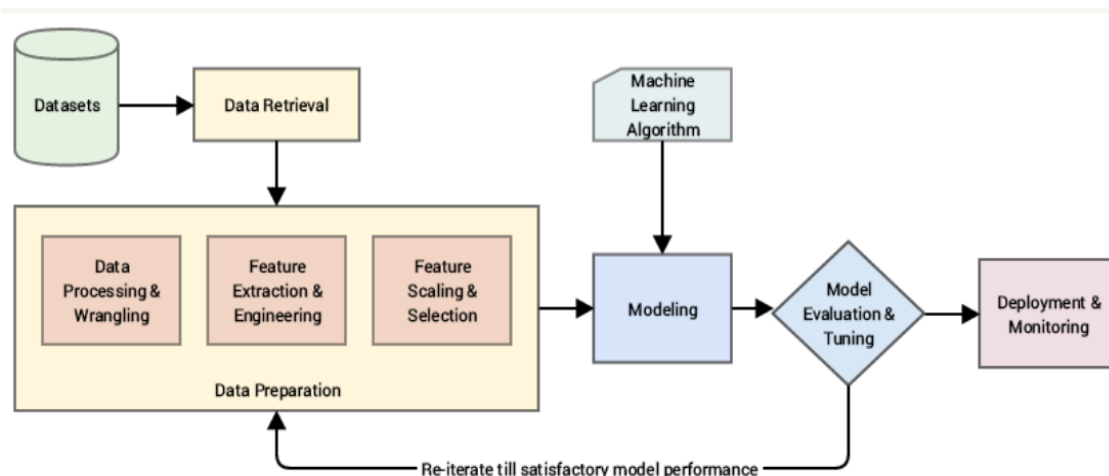


Figure 27 Feature engineering

Steps for feature engineering

Data Preparation:

Exploratory Analysis:

Benchmark:

6.2 Techniques for feature engineering

Imputation

Handling Outliers

Log Transform

Binning

One hot coding

Feature split

Although feature engineering helps in increasing the accuracy and performance of the mode.

6.3 Reading the dataset.

Here, after we have imported the dataset, we have read the dataset and we have only selected a few columns which are necessary for the target variable. Below is just the sample image of the required columns. These columns have both numbers and words. These columns are called categorical columns, so for the model building, we need just the numbers. So, here we do the encoding and change the words to numbers, by giving them some values, which is explained further in this document.

Unnamed: 0	Rif. Bando	Data Inserimento	Scadenza	Importo (€)	Oneri (€)	Onorario (€)	Fonti	Ente Appaltante	Categorie	
0	0	13350922	2021-11-17	2021-11-23 00:00:00	12000	NaN	NaN	ALBO	153	T-SANITA, O-SFW
1	1	13351176	2021-11-17	2021-12-17 00:00:00	1498598	3648	NaN	ALBO	638	O-SFW, O-MSW, O-GEST
2	2	13325837	2021-10-22	2021-11-29 00:00:00	430000	NaN	NaN	ALBO, CEE, GURI, RET_ALBO	126	O-SIC, O-SFW
3	3	13341901	2021-11-17	2021-12-06 00:00:00	1000000	NaN	1000000	ALBO, CEE, GURI	1512	T-SANITA, O-SFW, O-MSW, O-GEST, F-SEG, F-ASSO,...
4	4	13349952	2021-11-17	2021-12-01 00:00:00	138520	300	NaN	ALBO	2075	O-SFW, O-ELAB, O-APPA

Figure 28 Columns in Dataset.

7 PREDICTING THE IMPORTO VALUES:

In this model we are predicting the Importo which have NS as a value. The objective is to predict importo and importo feature is continuous, so RMSE and MAE are the correct metrics to evaluate them and to do the training by comparing different models performance to pick the best model on certain conditions.

Data is filtered based on importo value where each data is divided into 2 parts of 80 percent training data and 20 percent validation data to predict the data samples of unknown importo values so in this section mostly comparing performance with different conditions and picking the best model to predict the unseen data or unknown importo values.

7.1 TEXT CLASSIFICATION:

Text classification is a machine learning technique that assigns a set of predefined categories to open-ended text. Text classifiers can be used to organize, structure, and categorize pretty much any kind of text from documents, files, and all over the web.

Text is one of the most prevalent types of unstructured data, making up an estimated 80% of all information. The chaotic nature of language makes it difficult and time-consuming to analyze, comprehend, arrange, and sort text data. A text classifier can take this phrase as an input, analyze its content, and then automatically assign relevant tags, such as UI and Easy to Use.

Feature extraction, which involves turning each text into a numerical representation in the form of a vector, is the initial stage in training a machine learning NLP classifier. Bag of words is one of the most used methods, where a vector indicates a word's frequency inside a predetermined vocabulary of terms.

Then, the machine learning algorithm is fed with training data that consists of pairs of feature sets and tags. Once it's trained with enough training samples, the machine learning model can begin to make accurate predictions. The same feature extractor is used to transform unseen text to feature sets, which can be fed into the classification model to get predictions on tags. (8)

7.2 TEXT SUMMARIZATION:

Text summarization is the problem of reducing the number of sentences and words of a document without changing its meaning. There are different techniques to extract information from raw text data and use it for a summarization model, overall, they can be categorized as Extractive and Abstractive.

Extractive methods select the most important sentences within a text without necessarily understanding the meaning, therefore the result summary is just a subset of the full text.

Abstractive models use advanced NLP (i.e., word embeddings) to understand the semantics of the text and generate a meaningful summary. Consequently, Abstractive techniques are much harder to train from scratch as they need a lot of parameters and data. (9)

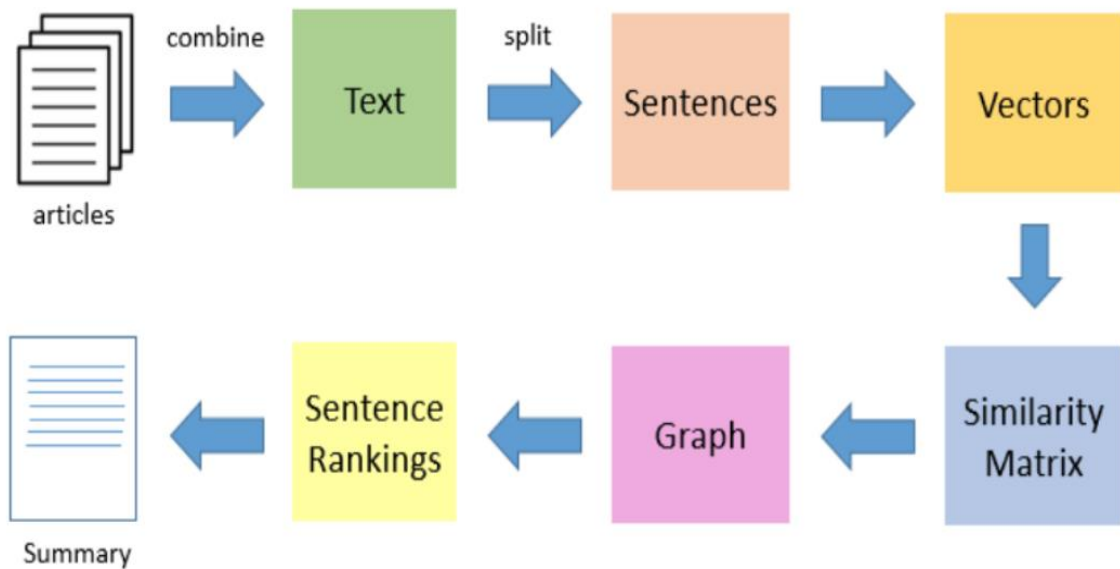


Figure 29 Process for Text summarization.

7.3 NAMED ENTITY RECOGNITION:

The term entity recognition (NER), sometimes known as entity chunking, extraction, or identification, is frequently used. It involves locating and classifying important textual information (entities). Any word or group of words that constantly refers to the same item is an entity. Each recognized object is put into a specific category. When reading a text, we automatically recognize identified entities like characters, concepts, places, and so forth. However, for computers to classify entities, we must first assist them in recognize.

In order to provide relevant training data for a NER model, we must first establish entity categories like as Name, Location, Event, Organization, etc. Then, eventually teach the NER model how to detect entities on its own by associating some word and phrase samples with their respective entities. Here we are using the Oggetto column from the data.

7.4 Label encoding

In machine learning, we frequently work with datasets that have several labels in a single column or numerous columns. These designations may be written in words or represented by numbers. The training data is frequently labelled in words to make it human readable or intelligible.

Label encoding is the process of transforming labels into a numeric form so that they may be read by machines. The operation of those labels can then be better determined by machine

learning techniques. It is a significant supervised learning pre-processing step for the structured dataset. (10)

8 Text Feature Extraction:

Machine Learning and deep learning algorithms don't understand text so need to represent a numerical vector that is nothing but feature extraction or feature conversion and, in this case, study mainly following 2 techniques one is semantic based, another is frequency based and those are Average Word2vec and Term frequency inverse document frequency.

8.1 Term Frequency Inverse Document frequency (TF-IDF):

In the name of the technique only its saying about it is based on word frequency and this technique is most widely used in search engines to search any document over the entire corpus and it contains 2 important equations one is term frequency, and another is inverse document frequency.

Term Frequency:

$TF(X) = (\text{No of times a word "X" appears in the Document}) / (\text{No of words present in the document})$

Term frequency is calculated based on document level

Inverse Document Frequency:

$IDF(X) = \log (\text{No of documents present in the corpus} / \text{No of documents where word "X" appears})$

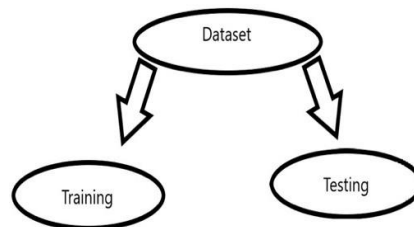
Term frequency inverse document frequency of word (X) is $TF(X) * IDF(X)$

While converting each document to vector format TFIDF of each word in the document is calculated based on the above formula and remaining words from vocabulary that are not present in the document their term frequency is zero for that word.

8.2 Average Word2vec:

This technique is based on semantics of each word and each document in the corpus is converted to a certain dimensional vector and vector shape can be customizable. Most used shape in industry is 50 or 100 so in this case study used 50-dimensional vector representation.

9 DATA MODELING



How to use the existing data is one of the initial considerations to be made when beginning a modeling project. The data can be divided into two groups, which are commonly referred to as the training and testing sets.

The training set is used to create feature sets and models; they are the building blocks for parameter estimation, model comparison, and all other processes necessary to create a final model.

The test set can only be used to estimate an objective, final evaluation of the model's performance if all these procedures have been finished. If you have fewer training data, the variance of your parameter estimations will be bigger.

Additionally, if you have fewer testing data, the performance statistic's variation will be bigger.

The best way to divide the data so that neither is excessively high will depend more on how much data you have. Once the preprocessing is done next step is applying machine learning algorithms on the preprocessed data by dividing data into 2 parts one is training another part testing with a ratio of 80 : 20 and the problem statement is predicting the continuous value that is important so regression algorithms can be applied like Linear Regression , Decision Tree Regressor , Random Forest Regressor , K Nearest Neighbors Regressor and Artificial Neural Network and doing hyper parameter tuning to find the best hyper parameter of the machine learning algorithm.

10 EVALUATION METRICS:

Once training is done the next step is evaluating models based on metrics that model is giving good results or not in case of regression most popular metrics are Root Mean Squared Error and Mean Absolute Error.

Root Mean Squared Error:

Predicted value means predicted from the algorithm and actual means ground truth

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}}$$

Mean Absolute Error:

It is the ratio of sum of all errors to No of samples

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - x|$$

11 METHODOLOGIES

11.1 Decision Tree.

One of the strongest and most well-liked algorithms is the decision tree. The supervised learning algorithms group includes the decision-tree algorithm. It works with output variables that are categorized and continuous.

- The first node, known as the root node, represents the entire population or sample and is then partitioned into two or more homogenous sets.
- Splitting: Splitting a node into two or more sub-nodes is the procedure.
- Decision Node: A decision node is a sub-node that divides into additional sub-nodes.
- Leaf/Terminal Node: Leaf or Terminal nodes are nodes that do not split.
- Pruning: Pruning is the process of removing sub-nodes from a decision node. You could describe it as the opposite of splitting.
- Branch / Sub-Tree: A branch or sub-tree is a smaller portion of the overall tree.
- Parent and Child Node: A node that has sub-nodes is referred to as the parent node of sub-nodes, and sub-nodes are the children of the parent node. (11)

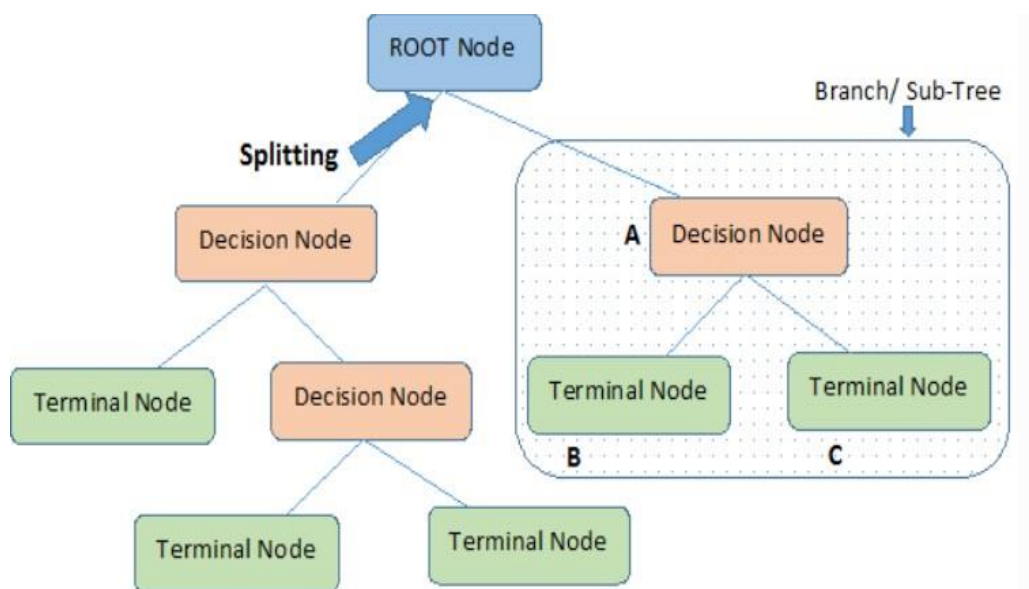


Figure 30 Structure of Decision trees

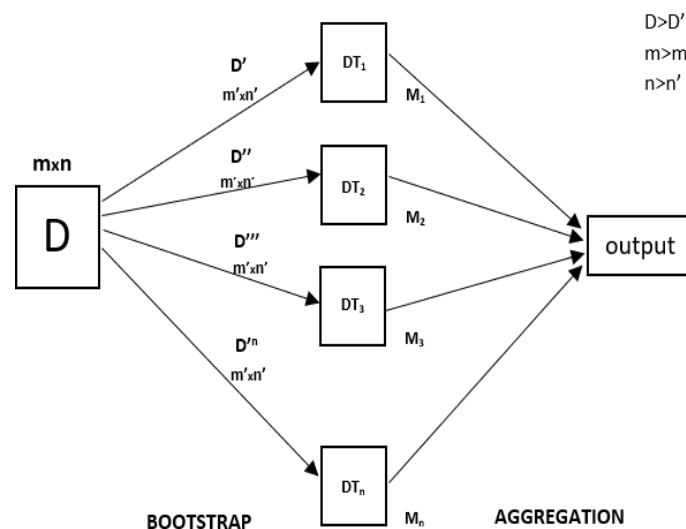
How is Splitting Decided for Decision Trees?

The choice to make strategic splits has a significant impact on a tree's accuracy. Regression and classification trees have different decision criteria. Mean squared error (MSE) is typically used in decision trees regression to determine whether to divide a node into two or more sub-nodes. If binary tree is used, the method will first select a value and divide the data into two subsets. It will compute the MSE independently for each subset. The value that produces the least MSE value is the one the tree selects.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2$$

11.2 Random forest Regression.

Every decision tree has a significant variance, but when we mix them all in parallel, the variance is reduced since each decision tree is perfectly trained using that specific sample of data, and as a result, the output is dependent on numerous decision trees rather than just one. The majority voting classifier is used to determine the final output in a classification challenge. The final output in a regression problem is the mean of every output. Aggregation is this section. (12)



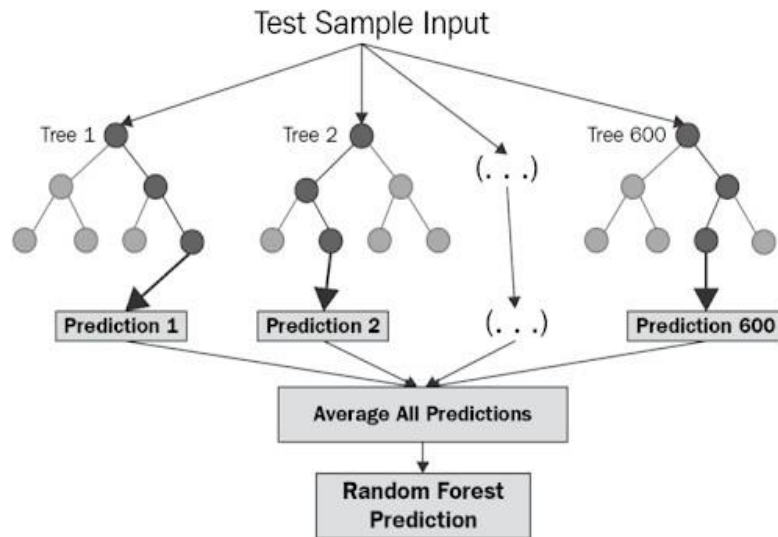


Figure 31 Structure of Random Forest regression.

An ensemble learning technique is used for classification and regression in the supervised learning algorithm known as random forest.

In contrast to boosting techniques, random forest is a bagging approach. In random forests, the trees grow in parallel, therefore there is no interaction between them as they grow.

In order to perform classification or regression, random forest builds many decision trees during training and then outputs the class that represents the mean of the predictions made by each tree. (12)

A random forest, which aggregates many decision trees with certain useful alterations, is a meta- estimator (i.e., it combines the outcome of multiple forecasts).

- A certain portion of the total number of features that can be separated at each node (which is known as the hyper-parameter). This restriction makes that the ensemble model uses all potentially predictive features fairly and does not rely excessively on any one feature.
- To add additional randomization and avoid overfitting, each tree selects a random sample from the original data set while creating its divides.

11.3 KNN (K-Nearest Neighbour)

One of the simplest machine learning algorithms, based on the supervised learning method, is K- Nearest Neighbour. The K-NN algorithm assumes that the new case and the existing cases are comparable, and it places the new instance in the category that is most like the existing categories. A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored. This means that utilizing the K-NN method, fresh data can be quickly and accurately sorted into a suitable category. Although the K-NN approach is most frequently employed for classification problems, it can also be utilized for regression. (13)

Since K-NN is a non-parametric technique, it makes no assumptions about the underlying data. It is also known as a lazy learner algorithm since it saves the training dataset rather than learning from it immediately. Instead, it uses the dataset to perform an action when classifying data. The KNN method simply saves the information during the training phase, and when it receives new data, it categorizes it into a category that is quite like the new data. (13)

Why do we need a K-NN Algorithm?

Which category does the new data point, x_1 , belong in if there are two categories, Category A and Category B? A K-NN algorithm is necessary to handle this kind of problem. Finding the category or class of a given dataset is made simple by K-NN. (13)

How does K-NN work?

The following algorithm can be used to describe how the K-NN works:

Step 1: Select the number K of the neighbors.

Step 2: Calculate the Euclidean distance between K neighbors in step two.

Step 3: Based on the determined Euclidean distance, select the K nearest neighbors.

Step 4: Count the number of data points in each category among these k neighbors.

Step 5: Assign the new data points to the category that has the greatest number of neighbors.

Step 6: Our model is complete. Calculating the average of the K nearest neighbors' numerical target is an easy way to implement KNN regression. An alternative method makes use of the K nearest neighbors' inverse distance-weighted

average. The same distance functions are used in KNN regression as in KNN classification. (14)

$$\text{Euclidean} = \sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

$$\text{Manhattan} = \sum_{i=1}^k |x_i - y_i|$$

$$\text{Minkowski} = \left(\sum_{i=1}^k (|x_i - y_i|)^q \right)^{\frac{1}{q}}$$

Only continuous variables can be used with the three-distance metrics mentioned above. The Hamming distance, which is a measurement of the number of instances in which corresponding symbols are different in two strings of identical length, must be used in the case of categorical variables.

Hamming Distance

$$D_H = \sum_{i=1}^k |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

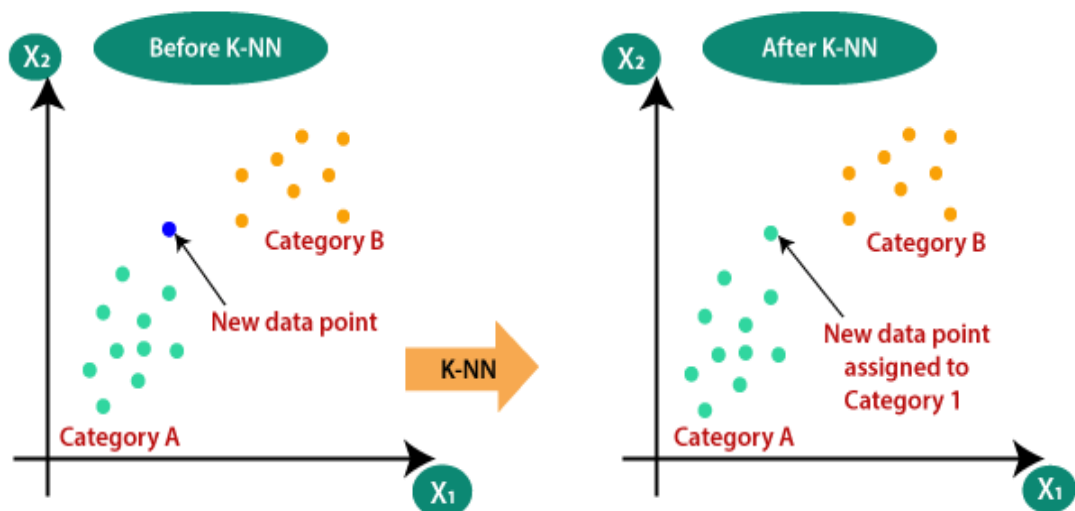


Figure 32 KNN performance.

11.4 Regression

Simple linear regression and multiple linear regression are the two fundamental types of regression procedures. Non-linear regression techniques like polynomial regression are used for more complex data and analysis. In contrast to multiple linear regression, which uses two or more independent variables to construct the prediction outcome, simple linear regression only uses one independent variable to predict the dependent variable Y . The general equation of a regression algorithms are as follows: (16)

1. Simple linear regression: $Y = a + b * X + u$
2. Multiple linear regression: $Y = a + b_1 * X_1 + b_2 * X_2 + b_3 * X_3 + \dots + b_t * X_t + u$

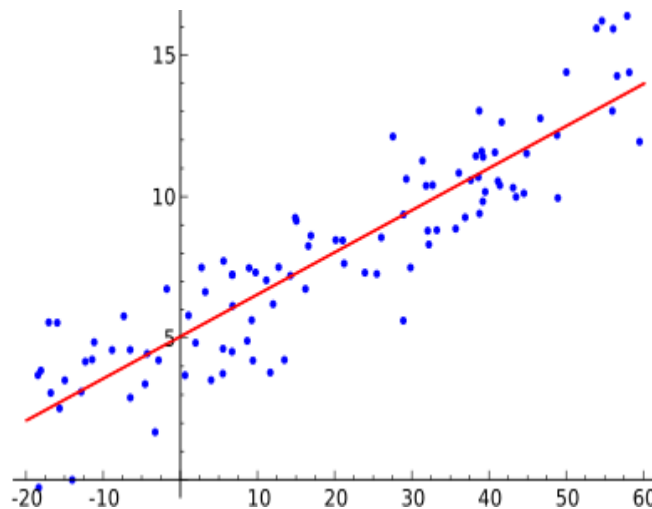


Figure 33 Simple Regression.

11.5 Deep learning Techniques.

Deep learning is a sort of machine learning that teaches a computer to carry out human-like functions including speech recognition, image recognition, and prediction. Deep learning puts up basic parameters about the data and teaches the computer to learn on its own by spotting patterns utilizing several layers of processing, as opposed to structuring the data to run through predefined equations. (17)

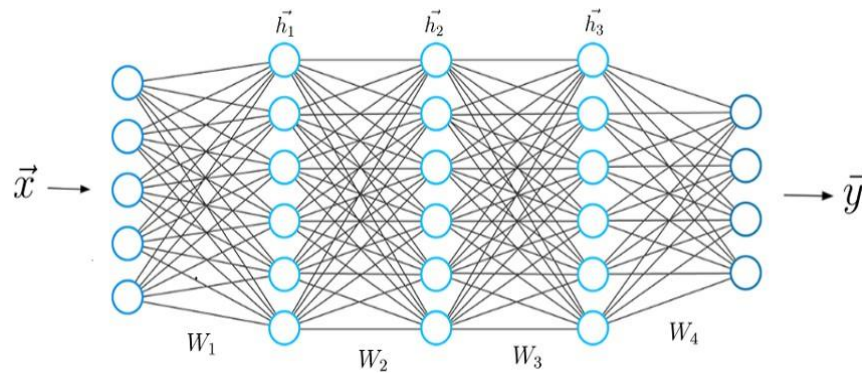


Figure 34 Deep hidden layers

11.5.1 Artificial Neural Networks

One deep learning algorithm that mimics the actions of neurons in the human brain is artificial neural networks. Vanilla neural networks, recurrent neural networks, and convolutional neural networks are just a few examples of artificial neural networks. Only organized data can be handled by vanilla neural networks; in contrast, recurrent neural networks and convolutional neural networks excel at handling unstructured data. In this article, the regression analysis will be carried out using Artificial neural networks. (17)

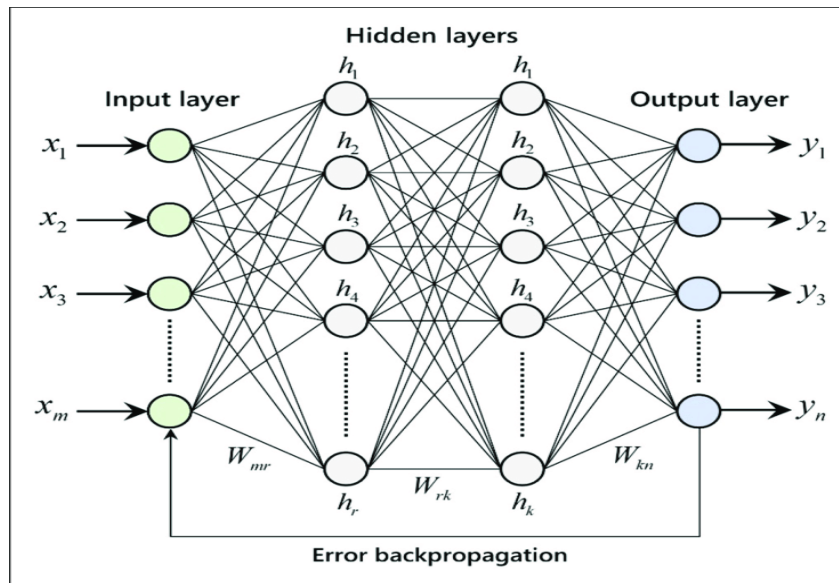


Figure 35 Structure of ANN

There are three layers in artificial neural networks: input, hidden, and output. There may be more than one concealed layer. A layer has n number of neurons in it. Each layer's neurons will each have an associated activation function. The function that introduces non-linearity into the relationship is the activation function. In our situation, a linear activation function is required in the output layer. Regularizers may also be connected to each layer. Regularizers are responsible for avoiding overfitting.

Artificial Neural Networks consists of two phases,

- Forward Propagation
- Backward Propagation

The practice of adding weights after multiplying them by each feature is known as forward propagation. The outcome also includes the bias. The process of updating the weights in the model is known as backward propagation. An optimization function and a loss function are needed for backward propagation. (18)

12 RESULTS FOR IMPORTO

12.1 TF-IDF:

1) Linear Regression

Train

Mean Squared Error.	414557.14496579504
Root Mean Squared Error	643.8611224214388
Mean Absolute Error	51.68039633207794

Test

Mean Squared Error	226094124772.85635
Root Mean Squared Error	475493.5591286767
Mean Absolute Error.	344437.68576645764

2) Decision Tree Regressor

Train

Mean Squared Error.	636991528.0957725
Root Mean Squared Error	25238.691093156405
Mean Absolute Error	20912.61485879611

Test

Mean Squared Error	693843877.7040668
Root Mean Squared Error	26340.916417316745
Mean Absolute Error.	21854.034807699547

3) Random Forest Regressor

Train

Mean Squared Error.	329870006.4406352
Root Mean Squared Error	18162.32381719463
Mean Absolute Error	14935.11918307689

Test

Mean Squared Error	586155356.5581557
Root Mean Squared Error	24210.64552130231
Mean Absolute Error.	19799.305505846783

4) KNN Regressor

Train

Mean Squared Error.	519421169.11181754
Root Mean Squared Error	22790.813261308107
Mean Absolute Error	18184.90928787879

Test

Mean Squared Error	766699167.3530688
Root Mean Squared Error	27689.333096935883
Mean Absolute Error.	22598.34833333333

Training dataset - The actual dataset that we use to train the model (weights and biases in the case of a Neural Network). The model sees and learns from this data.

Validation Dataset - The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

Based on the all-model performances (i.e., Decision Tree Regressor, Random Forest Regression, KNN, Linear Regression) we have calculated the Mean squared error, Root meansquared error, Mean Absolute error for all the models using tf-idf.

Term Frequency Inverse Document frequency (TF-IDF):

- **Mean Absolute Error:**

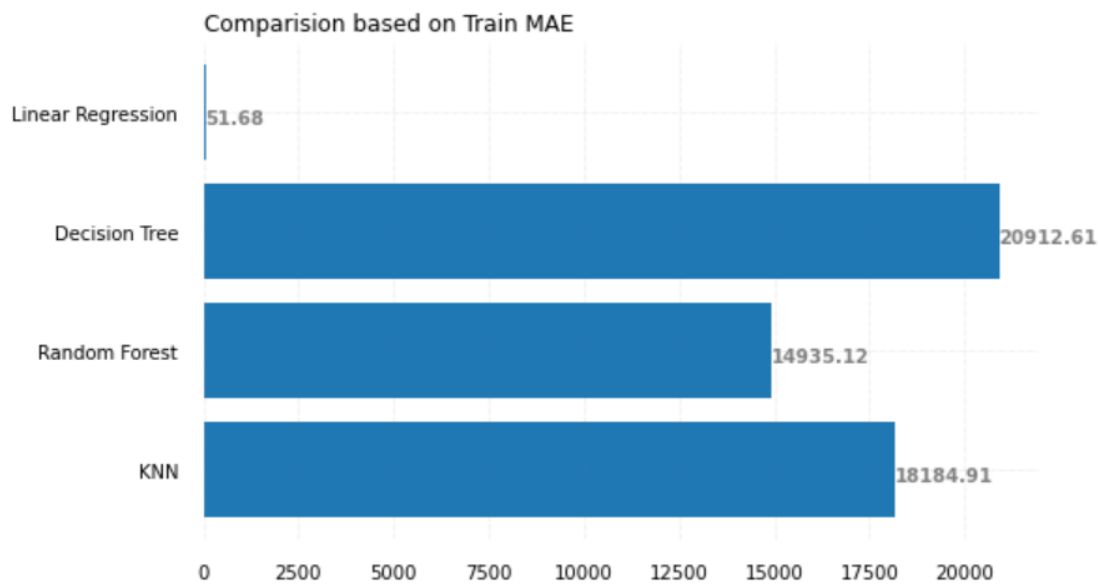


Figure 36 Bar plot representation for Train MAE.

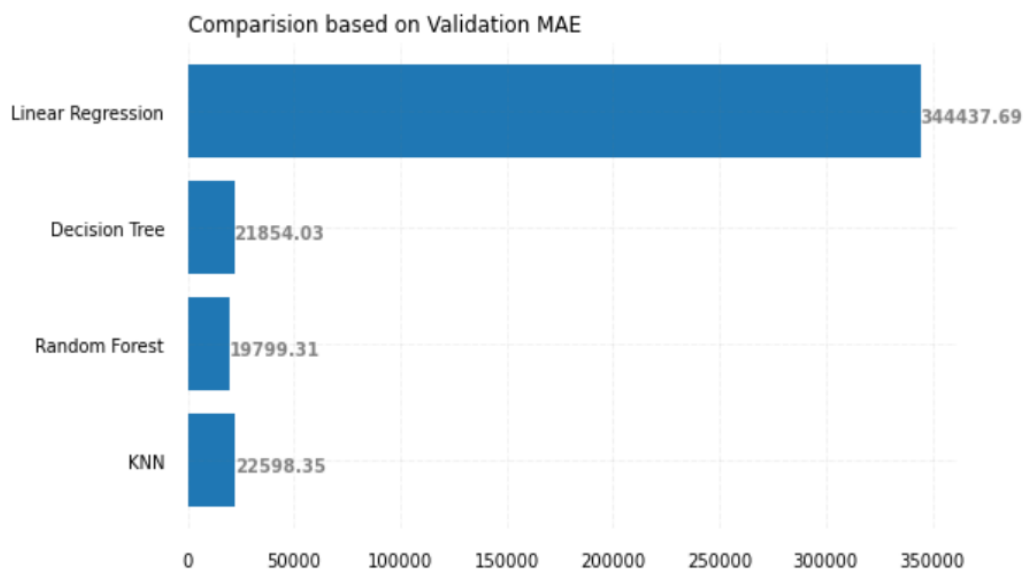


Figure 37 Bar plot representation for Validation MAE.

Random forest is giving good results compared to all other models in terms of train and validation without overfitting and underfitting

- **Root Mean Squared Error:**

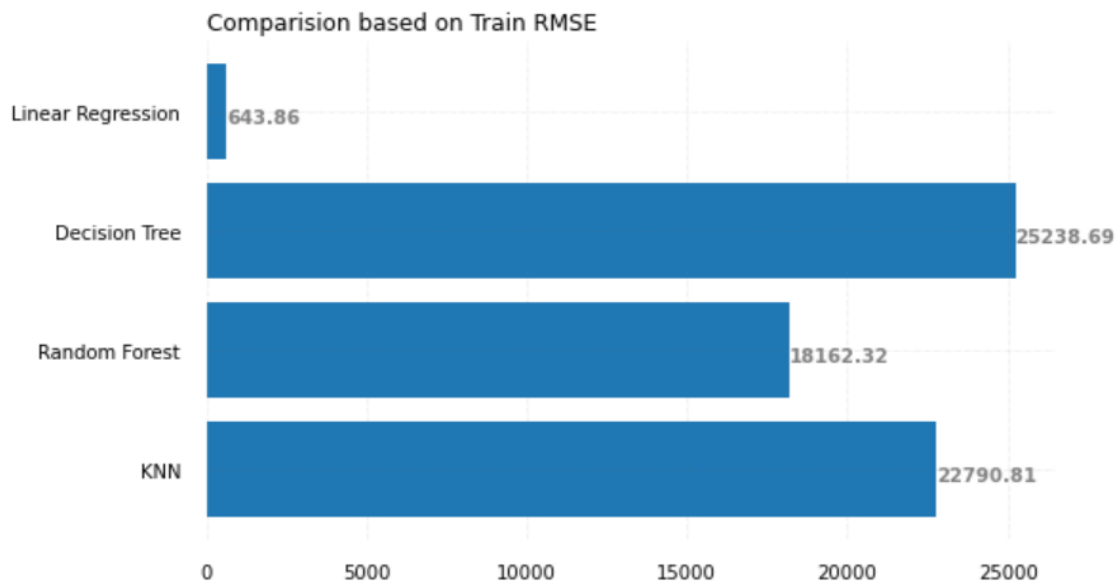


Figure 38 Bar plot representation for Train RMSE.

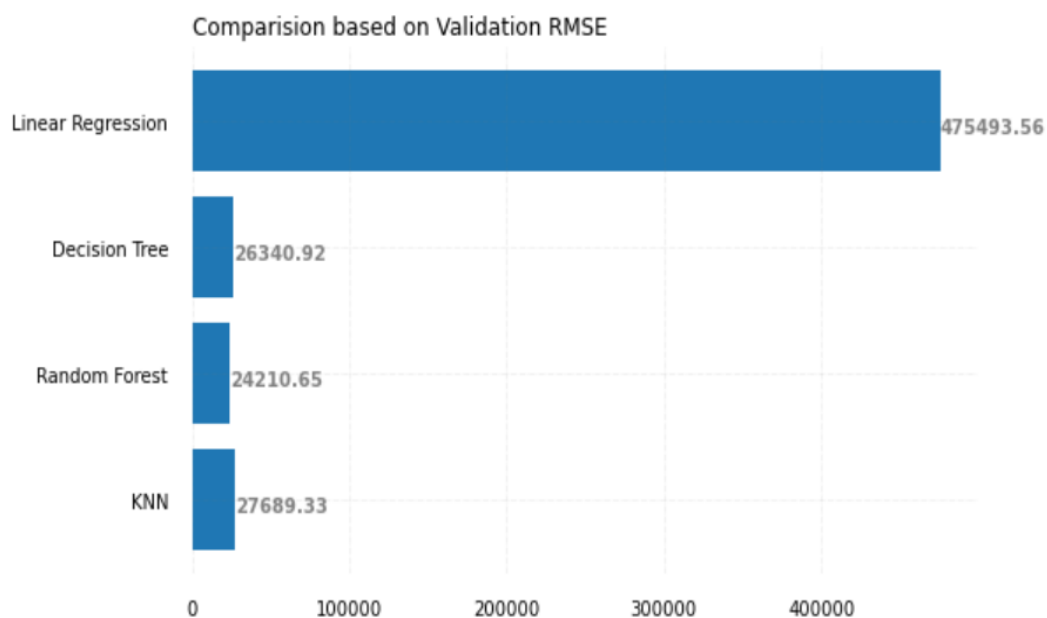


Figure 39 Bar plot representation for Test RMSE.

Random forest is giving good results compared to all other models in terms of train and validation without overfitting and underfitting.

	TRAIN MAE	TEST MAE	ALGORITHMS
0	51.68	344437.69	Linear Regression
1	20912.61	21854.03	Decision Tree
3	14935.12	19799.31	Random Forest
4	18184.91	22598.35	KNN

Train MAE is frequently lower than Test MAE since the model has previously seen the training set during training. Consequently, performing well on the practice set is easier. On the other hand, because it is more difficult to perform effectively on unobserved data, we frequently anticipate that the test MAE will be higher because the test set is unseen. However, it is not required that Train MAE be lower than Test MAE. The model might "by chance" perform better on the test set than the training set, which would lead to a lower Test MAE.

Therefore, considering the Test MAE, we see that Random Forest is the best model in overall performance as the Test MAE for Random Forest model is less when compared to other models.

12.2 Word2vec:

1) Linear Regression

Train

Mean Squared Error.	555600.694251
Root Mean Squared Error	23571.177648
Mean Absolute Error	18934.331818

Test

Mean Squared Error	714403.916679
Root Mean Squared Error	26728.326695
Mean Absolute Error.	22039.806004

2) Decision Tree Regressor

Train

Mean Squared Error.	601443.27489
Root Mean Squared Error	24524.347173
Mean Absolute Error	20036.885590

Test

Mean Squared Error	68877.93782
Root Mean Squared Error	26244.478790
Mean Absolute Error.	21609.014421

3) Random Forest Regressor

Train

Mean Squared Error.	114150.53763
Root Mean Squared Error	10684.164244
Mean Absolute Error	8466.241360

Test

Mean Squared Error	57826.901648
Root Mean Squared Error	24047.207195
Mean Absolute Error.	19109.862251

4) KNN Regressor

Train

Mean Squared Error.	56488.0427353
Root Mean Squared Error	23767.342806
Mean Absolute Error	19181.418047

Test

Mean Squared Error	694095.38594
Root Mean Squared Error	26345.686599
Mean Absolute Error	22148.442596

5) ANN

Train

Mean Squared Error.	62246.8645561
Root Mean Squared Error	24949.299687
Mean Absolute Error	20336.164192

Test

Mean Squared Error	68953.2988752
Root Mean Squared Error	26259.036660
Mean Absolute Error.	21871.354314

Training dataset - The actual dataset that we use to train the model (weights and biases in the case of a Neural Network). The model sees and learns from this data.

Validation Dataset - The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

Based on the all-model performances (i.e., Decision Tree Regressor, Random Forest Regression, KNN, Linear Regression, ANN) we have calculated the Mean squared error, Root mean squared error, Mean Absolute error for all the models using word2vec.

Word2vec:

- **Mean Absolute Error:**

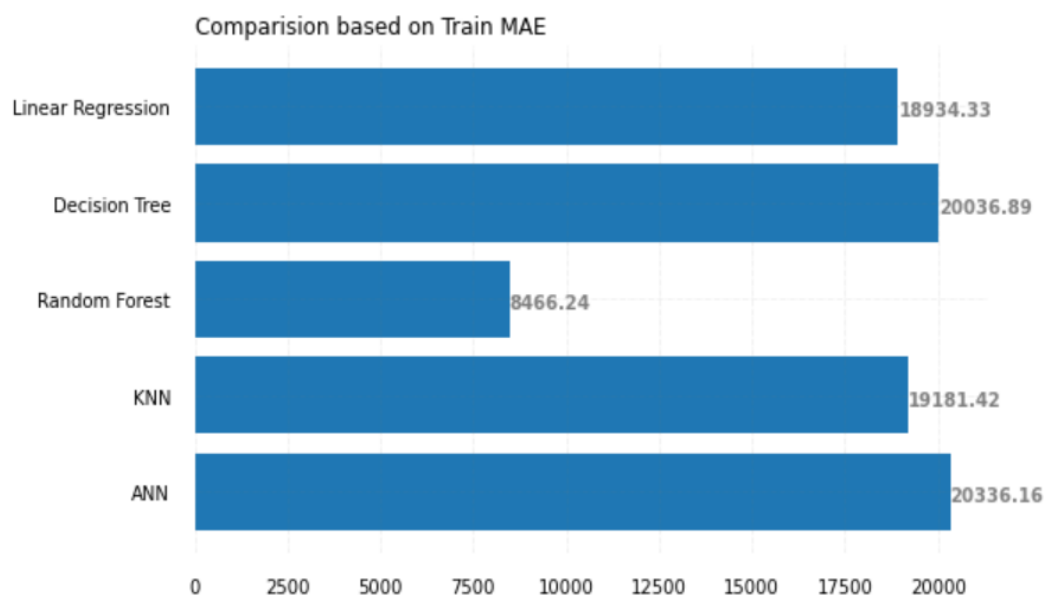


Figure 40 Bar plot representation for Train MAE.

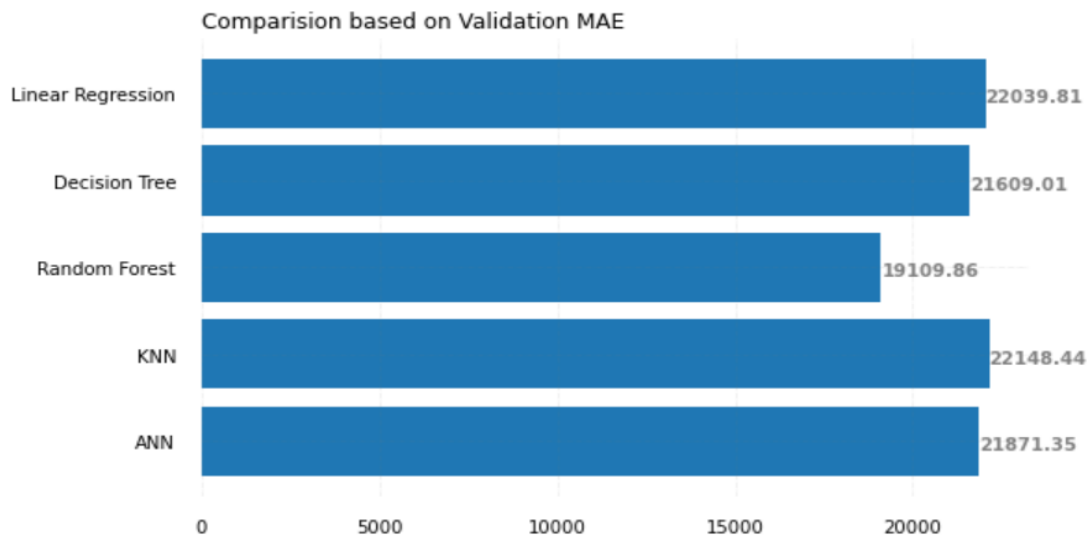


Figure 41 Bar plot representation for Validation MAE.

Random forest is giving good results compared to all other models in terms of train and validation without overfitting and underfitting

- **Root Mean Squared Error:**

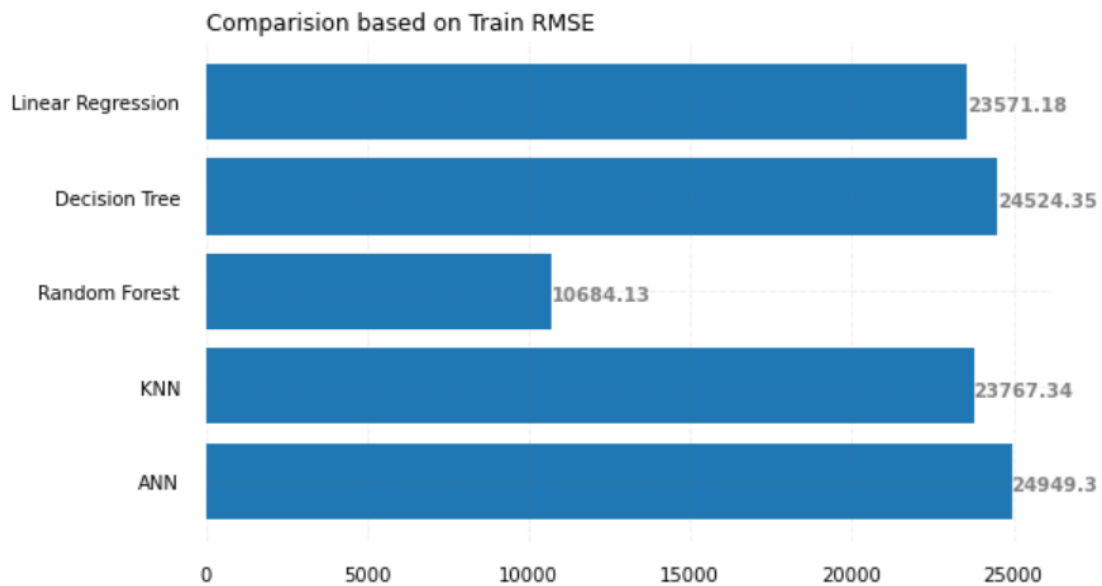


Figure 42 Bar plot representation for Train RMSE.

	TRAIN MAE	TEST MAE	ALGORITHMS
Linear Regression	18934.331818	22039.806004	Linear Regression
Decision Tree	20036.885590	21609.014421	Decision Tree
Random Forest	8466.241360	19109.862251	Random Forest
KNN	19181.418047	22148.442569	KNN
ANN	20336.164192	21871.354314	ANN

Random forest is giving good results compared to all other models in terms of train and validation without overfitting and underfitting.

Train MAE is frequently lower than Test MAE since the model has previously seen the training set during training. Consequently, performing well on the practice set is easier. On the other hand, because it is more difficult to perform effectively on unobserved data, we frequently anticipate that the test MAE will be higher because the test set is unseen. However, it is not required that Train MAE be lower than Test MAE. The model might "by chance" perform better on the test set than the training set, which would lead to a lower Test MAE.

Therefore, considering the Test MAE, we see that Random Forest is the best model in overall performance as the Test MAE for Random Forest model is less when compared to other models.

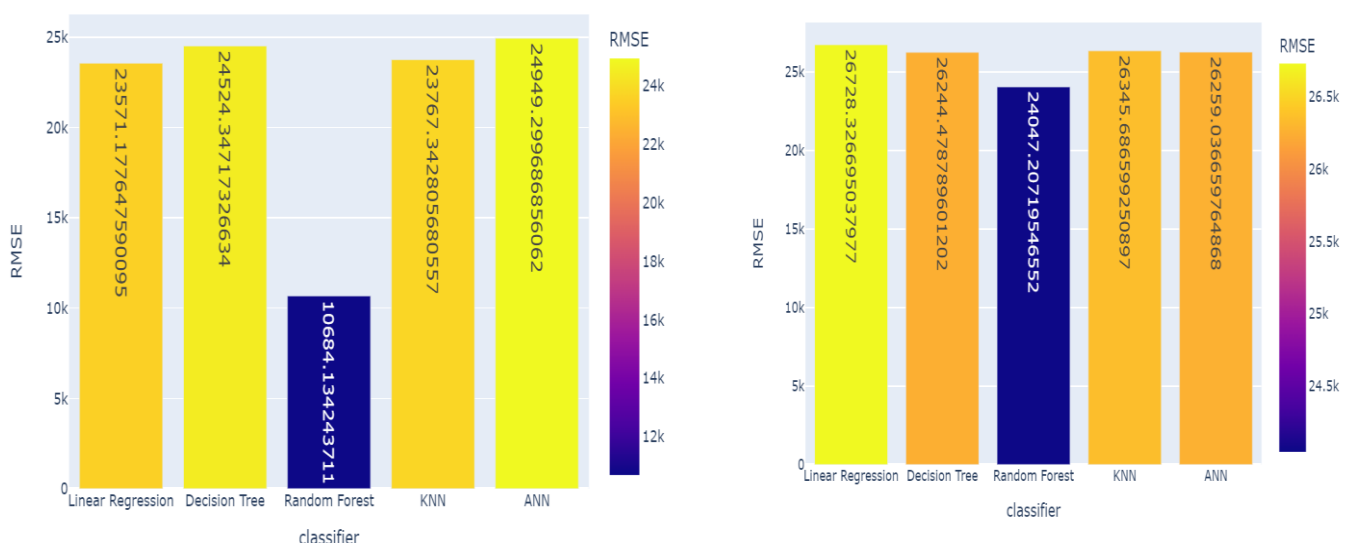


Figure 43 Bar plot representation for Train RMSE and Validation RMSE.

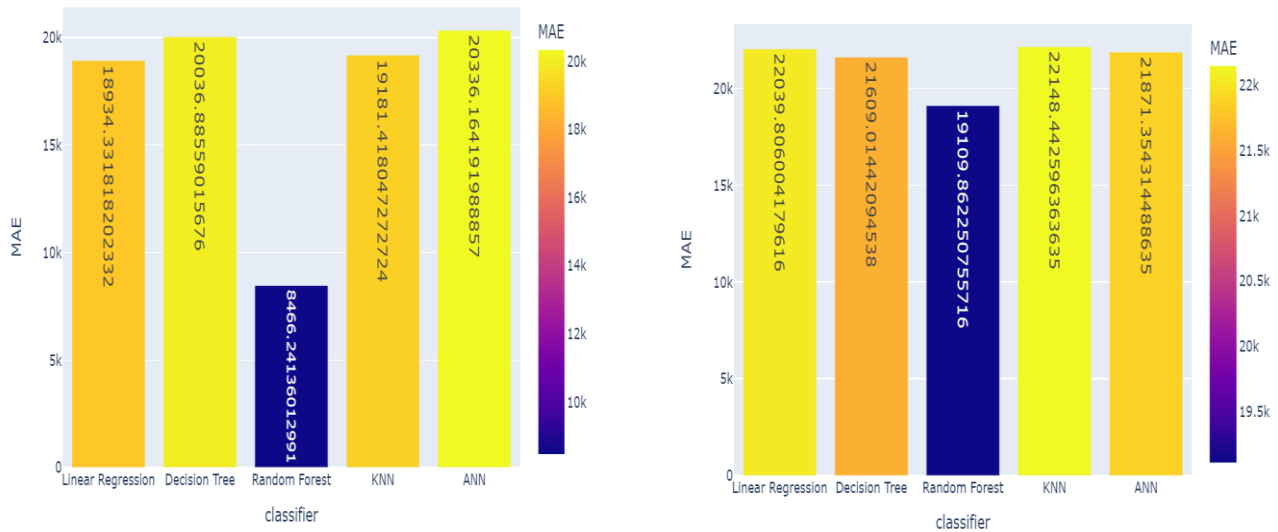


Figure 44 Bar plot representation for Train MAE and Validation

13 Comparison of different Methods:

METHOD	BEST ML TECHNIQUE	TRAIN MAE	VALIDATION MAE	TRAIN RMSE	VALIDATION RMSE
TF-IDF	Random Forest	14935.1191	19799.3055	18162.3238	24210.6455
Word2vec	Random Forest	8466.241360	14109.86225	10684.1342	24047.2071

Word2vec models is giving good results compared to TF - IDF with random forest regressor on Dataset. Random Forest regressor with average word2vec algorithms is giving compared to all other ml techniques with different preprocessing techniques like TF IDF so based on these results best model is random forest regressor with Average word2vec to predict unseen or unknown Importo values.

14 RESULTS OF TEXT CLASSIFICATION:

Training dataset - The actual dataset that we use to train the model.

Validation Dataset - The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

Based on the all-model performances (i.e., Decision Tree Regressor, Random Forest Regression, KNN, Linear Regression).

14.1 Logistic Regression:

classification report

	precision	recall	f1-score	support
0	0.14	0.58	0.22	186
1	0.98	0.87	0.92	5118
accuracy			0.86	5304
macro avg	0.56	0.72	0.57	5304
weighted avg	0.95	0.86	0.90	5304

auc score 0.8045693108617851

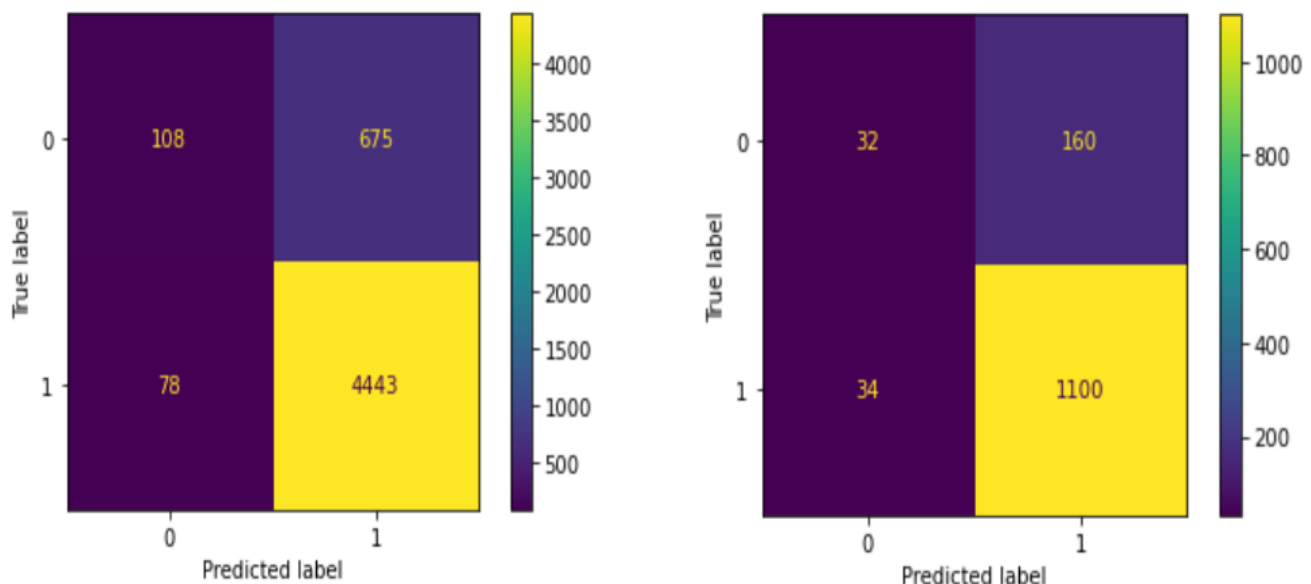


Figure 44 Confusion matrix representation for Train and Test.

```

classification report
              precision    recall  f1-score   support

     0       0.17         0.48         0.25         66
     1       0.97         0.87         0.92        1260

 accuracy          0.85         1326
 macro avg         0.57         0.68         0.58         1326
 weighted avg      0.93         0.85         0.89         1326

```

auc score 0.7931134259259259

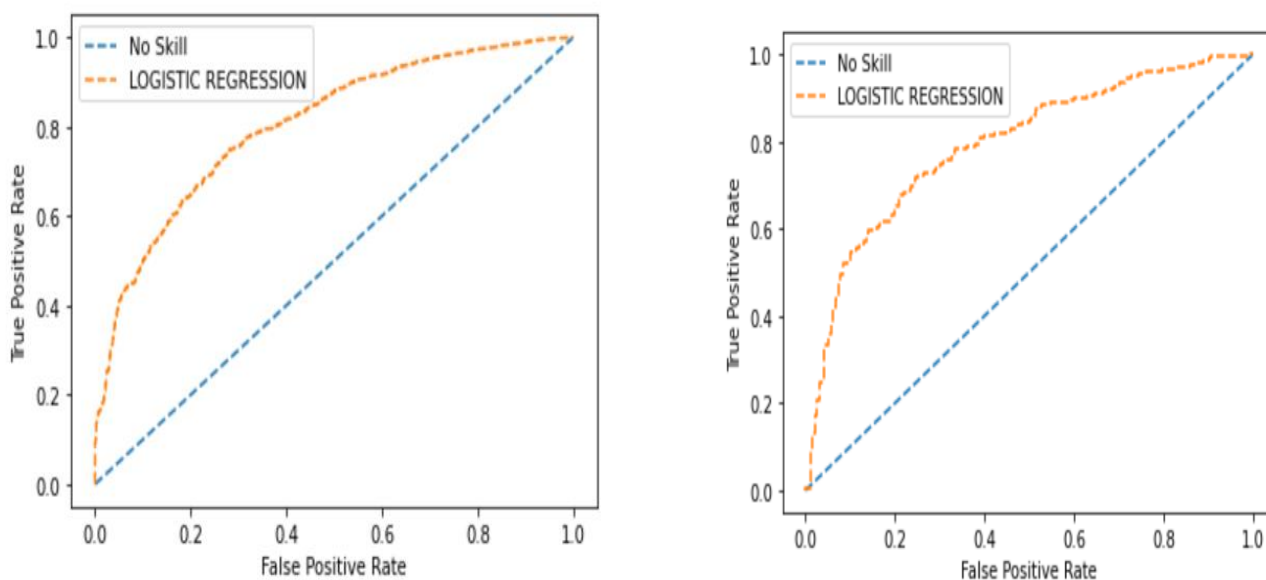


Figure 45 ROC curve representation for Train and Test.

14.2 Decision Tree Classifier:

```

classification report
              precision    recall  f1-score   support

     0       0.20         0.83         0.32        185
     1       0.99         0.88         0.93       5119

 accuracy          0.88         5304
 macro avg         0.59         0.85         0.62         5304
 weighted avg      0.97         0.88         0.91         5304

```

auc score 0.8036820084391189

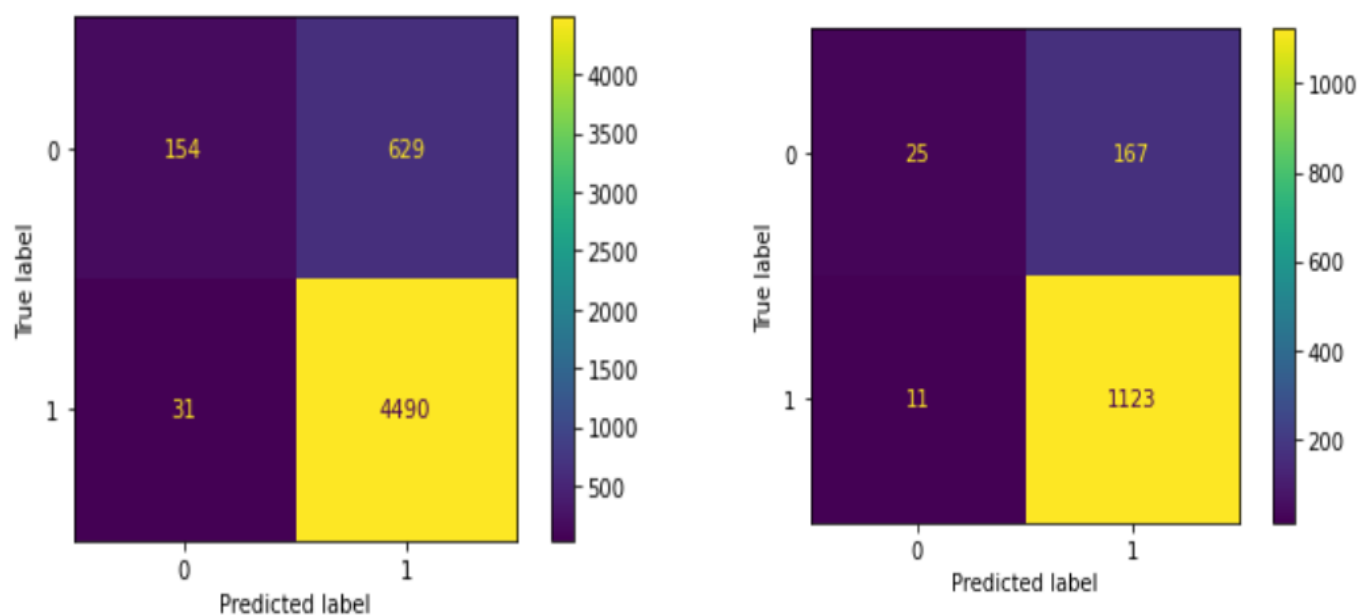
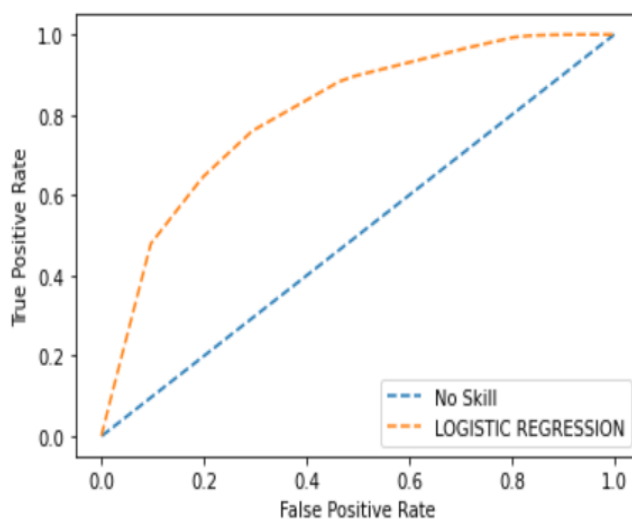


Figure 46 Confusion matrix representation for Train and Test.

classification report				
	precision	recall	f1-score	support
0	0.13	0.69	0.22	36
1	0.99	0.87	0.93	1290
accuracy			0.87	1326
macro avg	0.56	0.78	0.57	1326
weighted avg	0.97	0.87	0.91	1326

auc score 0.7661302175191066



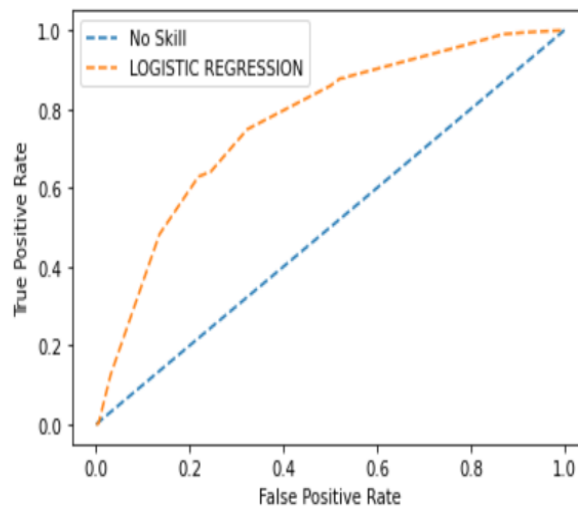


Figure 47 ROC curve representation for Train and Test.

14.3 Random Forest:

classification report				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	784
1	1.00	1.00	1.00	4520
accuracy			1.00	5304
macro avg	1.00	1.00	1.00	5304
weighted avg	1.00	1.00	1.00	5304

auc score 0.9999998587547878

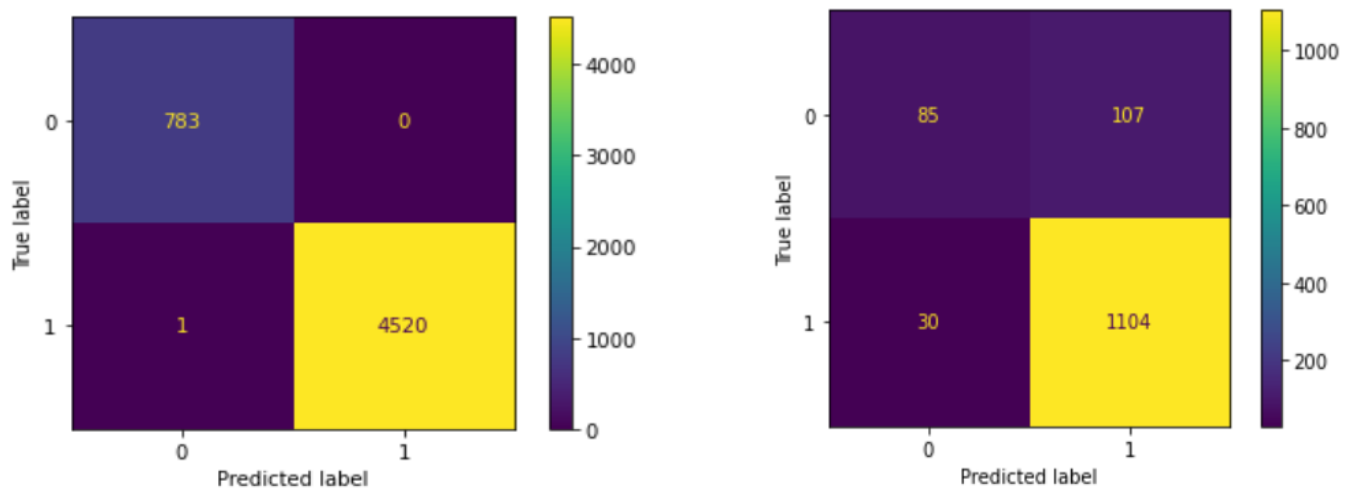


Figure 48 Confusion matrix representation for Train and Test

classification report				
	precision	recall	f1-score	support
0	0.44	0.74	0.55	115
1	0.97	0.91	0.94	1211
accuracy			0.90	1326
macro avg	0.71	0.83	0.75	1326
weighted avg	0.93	0.90	0.91	1326

auc score 0.9148088440623163

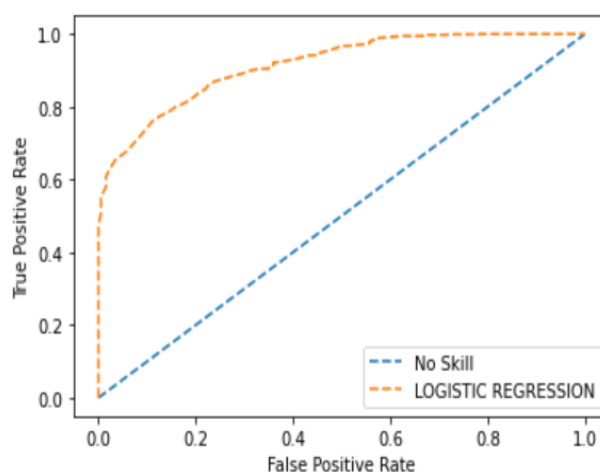
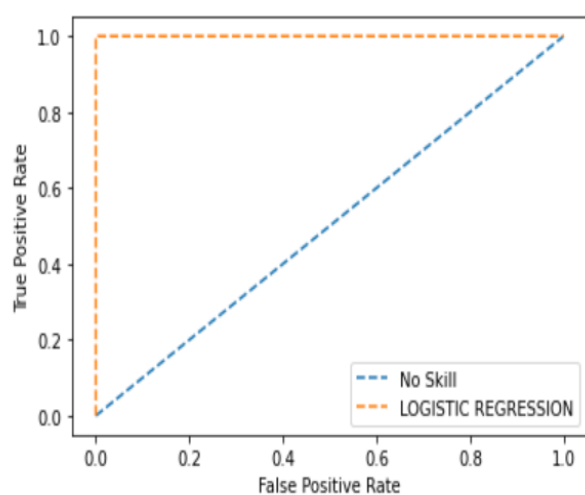


Figure 49 ROC curve representation for Train and Test.

14.4 KNN:

classification report				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	784
1	1.00	1.00	1.00	4520
accuracy			1.00	5304
macro avg	1.00	1.00	1.00	5304
weighted avg	1.00	1.00	1.00	5304

auc score 0.9999998587547878

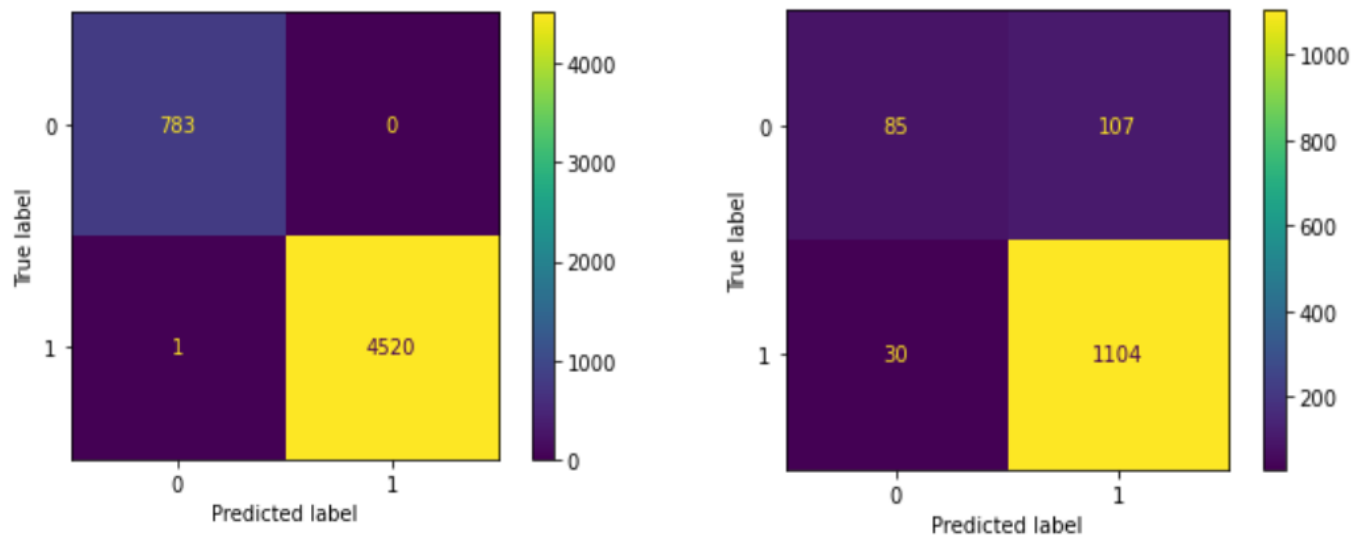


Figure 50 Confusion matrix representation for Train and Test.

```

classification report
              precision    recall  f1-score   support

     0       0.44         0.74         0.55         115
     1       0.97         0.91         0.94        1211

 accuracy          0.90         1326
 macro avg       0.71         0.83         0.75         1326
 weighted avg    0.93         0.90         0.91         1326

auc score 0.9148088440623163

```

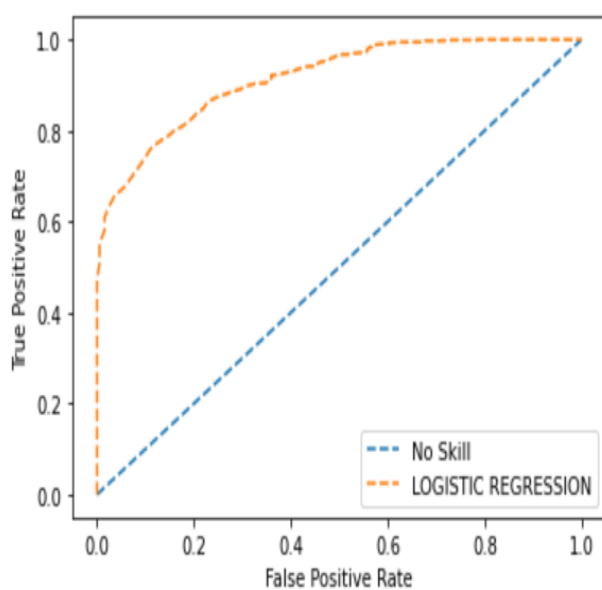
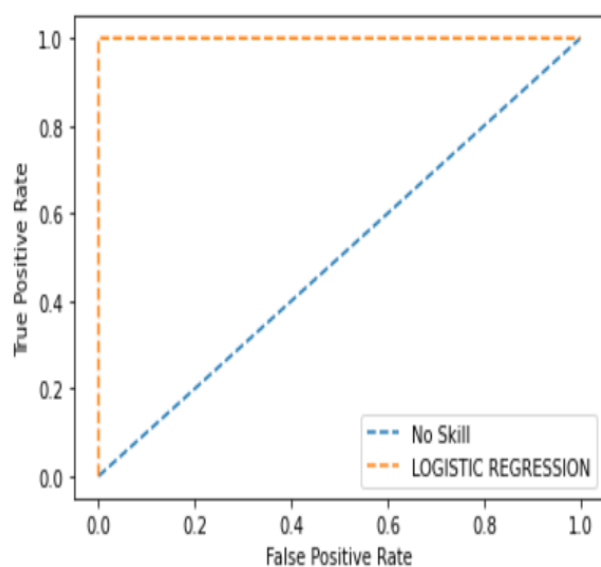


Figure 51 ROC curve representation for Train and Test.

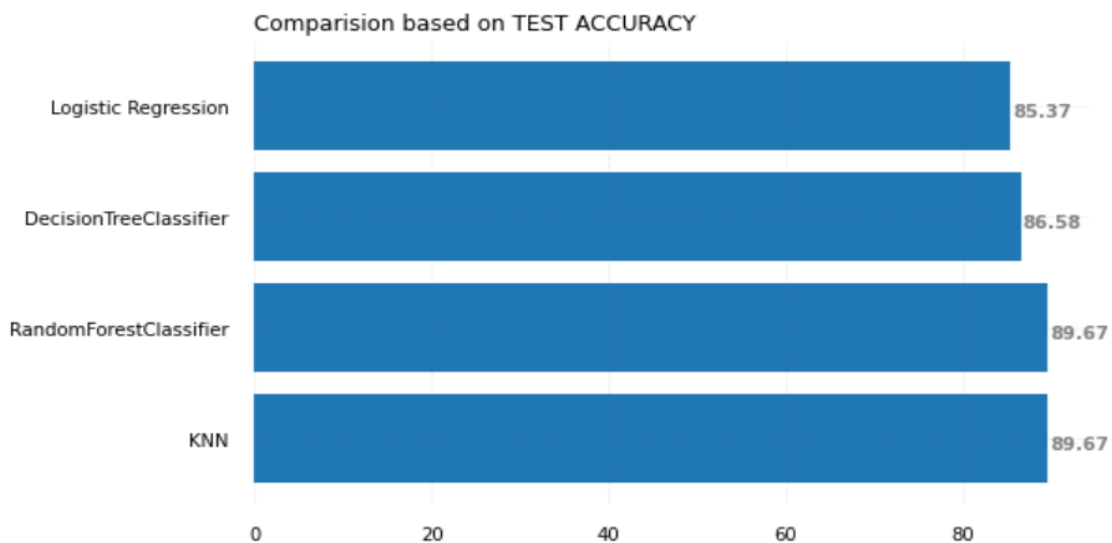
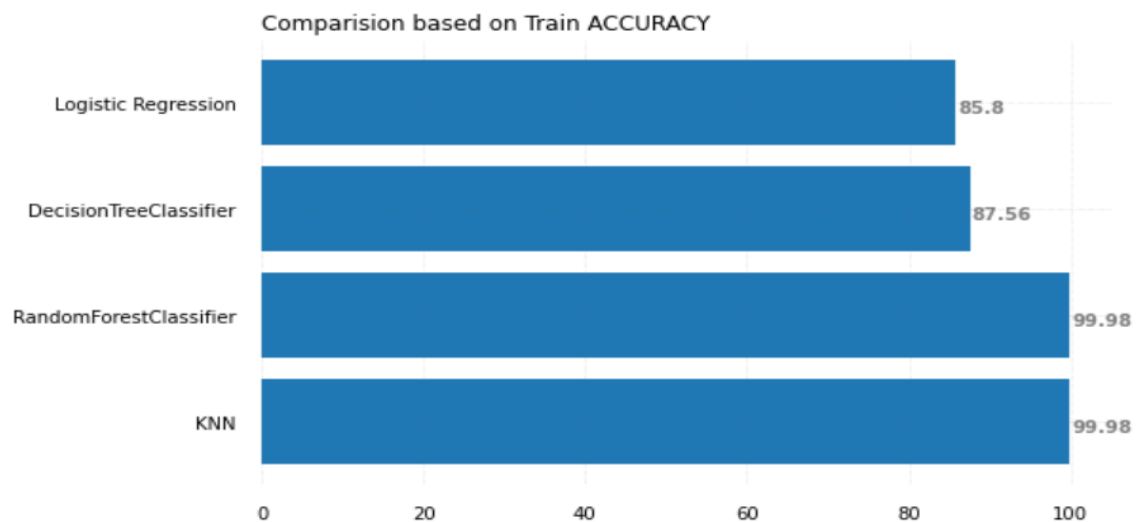
TRAIN:

	Accuracy (%)	AUC (%)	Precision	Classifier
0	85.803167	56.033911	0.868113	Logistic Regression
1	87.556561	59.491127	0.877124	Decision Tree Classifier
3	99.981146	99.988940	1.000000	Random Forest
4	99.981146	99.988940	1.000000	KNN

TEST:

	Accuracy (%)	AUC (%)	Precision	Classifier
0	85.369532	56.834215	0.873016	Logistic Regression
1	86.576169	56.025408	0.870543	Decision Tree Classifier
3	89.668175	70.812665	0.911643	Random Forest
4	89.668175	70.812665	0.911643	KNN

15 Comparison based on Train and Test Accuracy:



16 RESULTS OF TEXT SUMMARIZATION:

Text summarization is the process of creating shorter text without removing the semantic structure of text.

- Covert text to sentences: Converting a single text to list of sentences.
- Pre-process text: Clean the sentences by removing unnecessary words, stop words, punctuations, etc.
- Extract word vector embeddings: Word embeddings represents the vectors of words which are the mathematical form of a word. It represents words as real-valued vectors in a predefined vector space.

```
[('relativa', 0.9510765671730042), ('manutenzione', 0.9473053812980652), ('assistenza', 0.9458499550819397), ('tecnica', 0.9454094767570496), ('ordinaria', 0.9314870238304138), ('evolutiva', 0.9271891713142395), ('postazioni', 0.9173157215118408), ('sistemistica', 0.9150425791740417), ('aggiornamento', 0.9149183630943298), ('correttiva', 0.9077432155609131)]
```

=====

- Create similarity matrix: Similarity matrix represents the similarity between each sentence with every other sentences.
- Determine sentence rank: It is determined graphically by using page rank algorithm.

Original Text

Procedura Aperta Per La Fornitura Di Sistemi E Relativi Servizi, Materiale Di Consumo, Accessori E Kit Per Trattamenti Dialitici Ospedalieri E Domiciliari Per Le Necessita' Degli Enti Del Ssr Marche. Gara Europea A Procedura Aperta, In Modalita' Telematica, Condotta Da Asur In Qualita' Di Centrale Di Committenza, Per La Fornitura Di Sistemi E Relativi Servizi, Materiale Di Consumo, Accessori E Kit Per Trattamenti Dialitici Ospedalieri E Domiciliari Per Le Necessita' Dell'azienda Sanitaria Unica Regionale, Dell'azienda Ospedaliera Universitaria Ospedali Riuniti, Dell'azienda Ospedaliera Marche Nord E Dell'istituto Nazionale Riposo E Cura Anziani Di Ancona . Fornitura, In Noleggio, Dell'apparecchiatura, E L'acquisto Di Filtri E Di Materiale Di Consumo Per Trattamenti Dialitici Extracorporei Continui/intermittenti In Terapia Intensiva Per Pazienti Critici. Fornitura In Noleggio Dell'apparecchiatura E L'acquisto Di Materiali Di Consumo Per Trattamenti Di Plasmaexchange-cf (cascade Filtration), Plasma-adsorbimento . Acquisto Di Materiale Di Consumo Per Il Trattamento Di Pazienti In Emodialisi Domiciliare Frequente/ Quotidiana, Omni comprensivo Del Costo Delle Apparecchiature A Corredo. Acquisto Di Materiale Di Consumo Per Il Trattamento Di Pazienti In Emodialisi Domiciliare. Fornitura, In Noleggio, Dell'apparecchiatura E L'acquisto Di Materiali Di Consumo Per L'effettuazione Della Metodica Afb Con K Variabile E Hdx Per Pazienti Diabetici. Fornitura, In Noleggio, Di Apparecchiatura, E L'acquisto Di Materiali Di Consumo Per Trattamenti Extracorporei Per Rimozione Di Mediatori Dell'infiammazione Per Adsorbimento In Corso Di Shock Settico. Fornitura, In Noleggio, Di Apparecchiatura, E L'acquisto Di Materiale Di Consumo Per Trattamenti Per Pazienti Dislipidici Refrattari Alla Terapia Farmacologica Massimale. Fornitura Di Materiali Di Consumo Per L'effettuazione Di Trattamenti Hfr (emodiafiltrazione Con Reinfusione Endogena) Con Filtro E Cartuccia Adsorbente. Fornitura, In Noleggio, Di Un Sistema Di Telemedicina Per La Cura, Monitoraggio Ed Addestramento A Distanza Mediante Connessione Audio E Video. Fornitura, In Noleggio, Di Apparecchiature Per Osmosi Portatile Comprensivo

Ranking Sentences Based ON Scores in Descending Orders (Selecting Only top -5)

(0.083933547182691499, "Fornitura, In Noleggio, Dell'apparecchiatura E L'acquisto Di Materiali Di Consumo Per L'effettuazione Della Metodica Afb Con K Variabile E Hdx Per Pazienti Diabetici")

(0.08390895168966625, "Fornitura, In Noleggio, Di Apparecchiatura, E L'acquisto Di Materiale Di Consumo Per Trattamenti Extracorporei Per Rimozione Di Mediatori Dell'infiammazione Per Adsorbimento In Corso Di Shock Settico")

- **Generate Summary:** Generate a summary by extracting the sentences having higher ranks.

Ranking Sentences Based ON Scores in Descending Orders (Selecting Only top -5)

(0.5, "S21023 - Procedura Aperta Per La Conclusione Di Contratti Attuativi Basati Su Accordo Quadro Con Un Solo Operatore Economico, Di Cui All'art")

(0.5, "50/2016, Aveni Ad Oggetto L'affidamento Dei Servizi Di Manutenzione, Assistenza, Conduzione Ed Evoluzione Dei Sistemi Di E-government Per Imprese, Professionisti Ed Operatori Degli Enti Dell'area Metropolitana Di Bari, In Esecuzione Delle Determinazione Dirigenziale Innovazione Tecnologica Sistemi Informativi E Tlc N")

Summary

S21023 - Procedura Aperta Per La Conclusione Di Contratti Attuativi Basati Su Accordo Quadro Con Un Solo Operatore Economico, Di Cui All'art. 50/2016, Aveni Ad Oggetto L'affidamento Dei Servizi Di Manutenzione, Assistenza, Conduzione Ed Evoluzione Dei Sistemi Di E-government Per Imprese, Professionisti Ed Operatori Degli Enti Dell'area Metropolitana Di Bari, In Esecuzione Delle Determinazione Dirigenziale Innovazione Tecnologica Sistemi Informativi E Tlc N.

17 RESULTS OF NAMED ENTITY RECOGNITION:

NER systems have been created that use linguistic grammar-based techniques as well as statistical models such as machine learning.

Spacy is a multipurpose open-source Natural Language Processing library. It contains mechanisms for named entity recognition. A quick statistical entity recognition system is available in Spacy. Spacy is a simple tool to utilize for NER tasks. The spacy model generally works well for all sorts of text data, even though we frequently need to train our own data.

Spacy. load() is essentially a convenience wrapper that reads the pipeline's config. cfg file, creates a language (it_core_news_sm) object using the language and pipeline information, loads the model data, weights, and returns it.

In this NER one of the major challenges in identifying named entities is language. Recognizing words which can have multiple meanings or words that can be a part of different sentences. Another major challenge is classifying similar words from texts.

Selezione Esperto Progettazione - 6 - 37 - PER - Named person or family.

La Realizzazione Di - 42 - 61 - MISC - Miscellaneous entities, e.g. events, nationalities, products or works of art

Spazi Laboratoriali - 62 - 81 - LOC - Non-GPE locations, mountain ranges, bodies of water

La Dotazione Di Strumenti Digitali - 88 - 122 - MISC - Miscellaneous entities, e.g. events, nationalities, products or works of art

L'apprendimento Delle Stem - 127 - 153 - MISC - Miscellaneous entities, e.g. events, nationalities, products or works of art

Miur Prot - 162 - 171 - PER - Named person or family.

Del 15 Maggio 2021 - 184 - 203 - MISC - Miscellaneous entities, e.g. events, nationalities, products or works of art

Orario Euro - 211 - 222 - PER - Named person or family.

18 CONCLUSION

Tender Market analysis is one of the most important aspects for the organization as it helps in understanding the data that is present in the dataset i.e., Importo, Category, Fonti, Oggetto, Zone and other attributes. These factors help companies in making a well-informed decision which is highly crucial for business.

The study of unprocessed datasets to make assumptions about the information they contain is known as data analytics. It enables us to recognize patterns in the raw data and draw valuable conclusions from it. Applications using simulation, automated systems, and machine learning algorithms may be employed in data analytics processes and techniques. Systems and algorithms utilize the unstructured data for human use.

Data visualization tools like, Kibana are a great way to represent the data visually. They help the data to understand in a meaningful way. The aim of this project was to test the ability of machine learning techniques in predicting the Importo values in the dataset. Our target values were Importo, which has different approaches in prediction due to most of the missing entries.

From all model performances (i.e., Decision tree regressor, Random Forest regression, KNN, Linear regression and ANN) we have calculated the Mean squared error, Root mean squared error, and Mean Absolute error for all the models. In the results section, we have calculated the Train MAE and Test MAE for all the models. All these lead to accuracy performance. The lower values are the best performance.

However, when we are using real world data it's not always possible that the predictions turn out to be true. There might be limitations in the data which go unseen (when we are handling big data).

For future work, we can use different columns for text classification to show more accurate results as I have considered category as the main column to do the classification. Language is one of the biggest obstacles to named entity identification.

recognizing words that can be used in different phrases or that can have multiple meanings. Sorting similar terms out of texts is another difficult task.

19 ACKNOWLEDGEMENT:

We owe considerable thanks to our mentors Flavio Venturini, Juan Carlos Martinez Perez and Antonella Lipari, from Iconsulting for discussions during the development of the systems, and for data preparation and system implementation. We would like to thank Professor. Ioannis Chatzigiannakis my thesis Supervisor from La Sapienza university di Roma, helped me with my research work.

We are grateful to Maria Laura Bongiovanni, HR from Iconsulting for continuous support to communications and paper works.

This work was completed at Iconsulting, Via della Conciliazione, 10
00193 Roma (RM) - Italy.

20 Bibliography

1. <https://www.elastic.co/what-is/elasticsearch>.
2. <https://www.elastic.co/what-is/kibana>.
3. <https://www.elastic.co/guide/en/kibana/current/create-a-dashboard-of-panels-with-web-server-data.html>.
4. <https://www.elastic.co/guide/en/elasticsearch/reference/7.17/text.html>.
5. <https://www.elastic.co/guide/en/kibana/master/dashboard.html>.
6. <https://www.analyticssteps.com/blogs/feature-engineering-process-and-techniques>.
7. <https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10>.
8. <https://monkeylearn.com/text-classification/>.
9. <https://towardsdatascience.com/text-summarization-with-nlp-texttrank-vs-seq2seq-vs-bart-474943efeb09>.
10. <https://www.kaggle.com/code/vamsichennakesava/credit-card-fraud-detection>.
11. <https://medium.com/@rahulrastogi1104/decision-tree-regression-and-its-mathematical-implementation-c87352a8a2b0>.
12. <https://prutor.ai/random-forest-regression-in-python/>.
13. <https://github.com/shubham9793/K-Nearest-Neighbor-KNN-Algorithm-for-Machine-Learning>.
14. https://www.saedsayad.com/k_nearest_neighbors_reg.htm.
15. <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>.
16. <https://www.scribbr.com/statistics/simple-linear-regression/#:~:text=What%20is%20simple%20linear%20regression,Both%20variables%20should%20be%20quantitative>.
17. <https://www.expressanalytics.com/blog/neural-networks-prediction/#:~:text=Neural%20networks%20work%20better%20at,the%20way%20a%20human%20does>.
18. <https://www.analyticsvidhya.com/blog/2021/08/a-walk-through-of-regression-analysis-using-artificial-neural-networks-in-tensorflow/#:~:text=One%20of%20those%20techniques%20is,activation%20function%20in%20each%20layer>.

Reference papers

Han J, Kamber M (2011).. Data Mining: Concepts and Techniques[JJ]. Data Mining Concepts Models Methods & Algorithms Second Edition, 5 (4): 1-18. . . (n.d.).

Ranjbar et al. (2010). M. Ranjbar, S. Soleymani (2010). The Way of using artificial neural network. . . (n.d.).

Tan. (2017). K. Tan, Neural networks: An Introductory Review of Deep Learning for Prediction Models With Big Data. . (n.d.).

Books

- i. *Head First Statistics: A Brain-Friendly Guide.*
- ii. *Introduction to Machine Learning with Python: A Guide for Data Scientists.*
- iii. *Practical Statistics for Data Scientists.*
- iv. *Python for data analysis.*