



Co-funded by the  
Erasmus+ Programme  
of the European Union

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

IS “EINAUDI PARETO”  
PALERMO

ITALY EARTQUAKE



# Course Overview

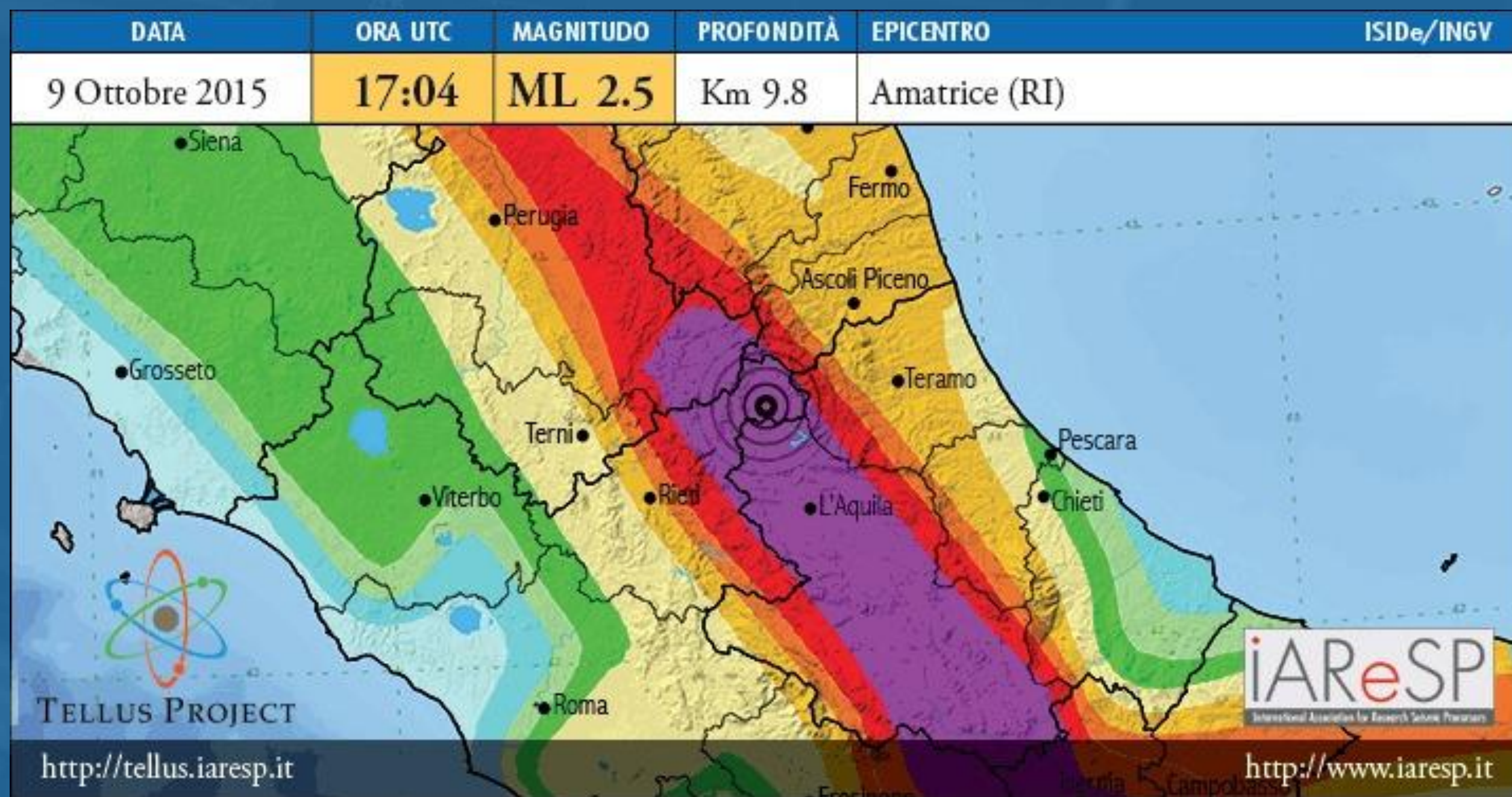
## List of topics

- DATA DESCRIPTION
- METHODOLOGY -
- RESULTS -
- CONCLUSION

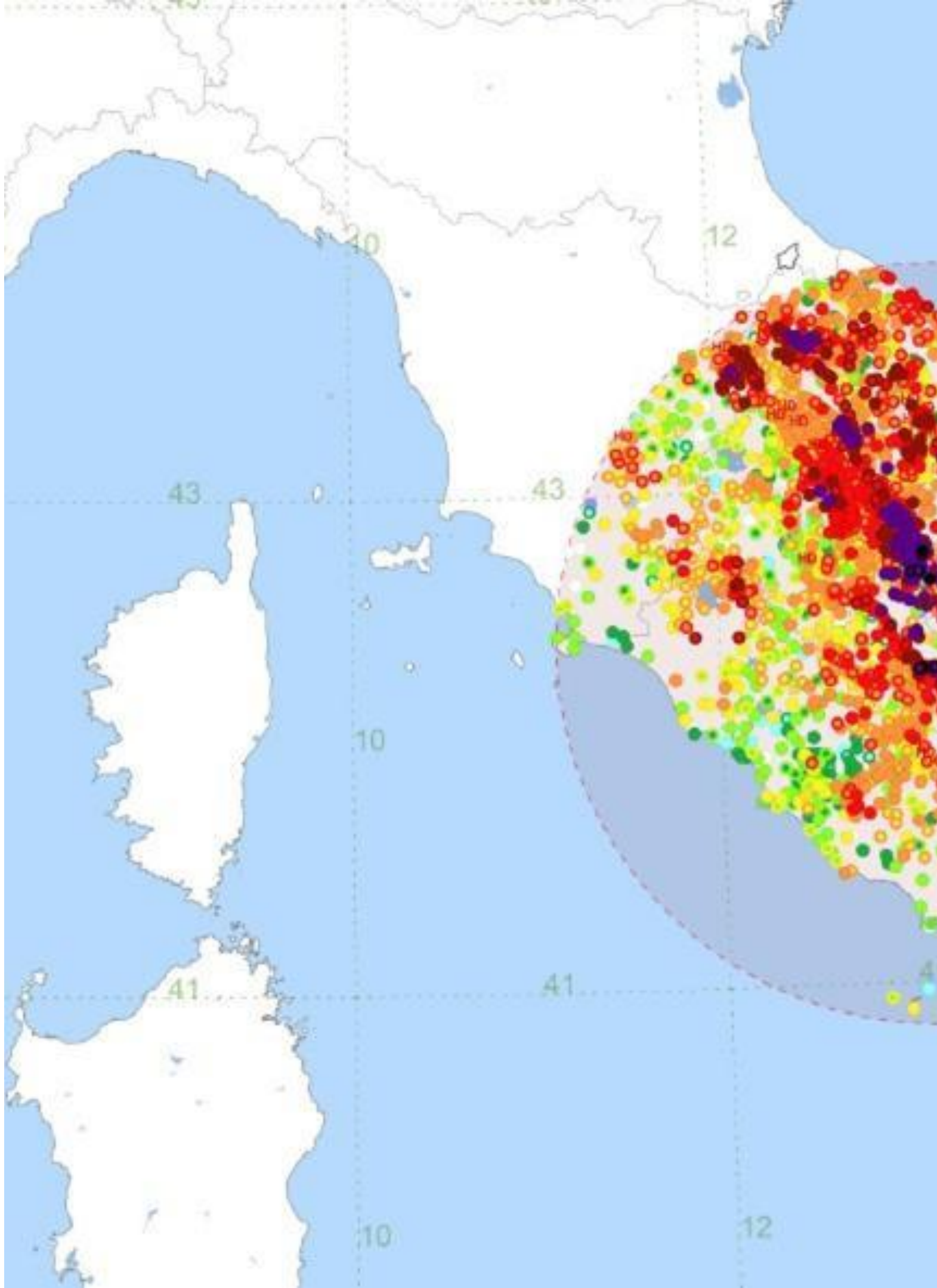




# DATA DESCRIPTION







•The seismic emergency on which research is focused begins in central Italy on 24 August 2016, with an earthquake of magnitude 6.0. Thousands of people were involved in the event, which caused 299 deaths, numerous injuries and serious damage on the territory.

•The emergency scenario, over the months is aggravated due to further strong earthquakes





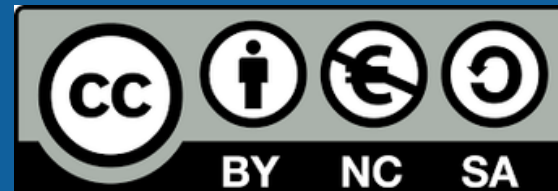
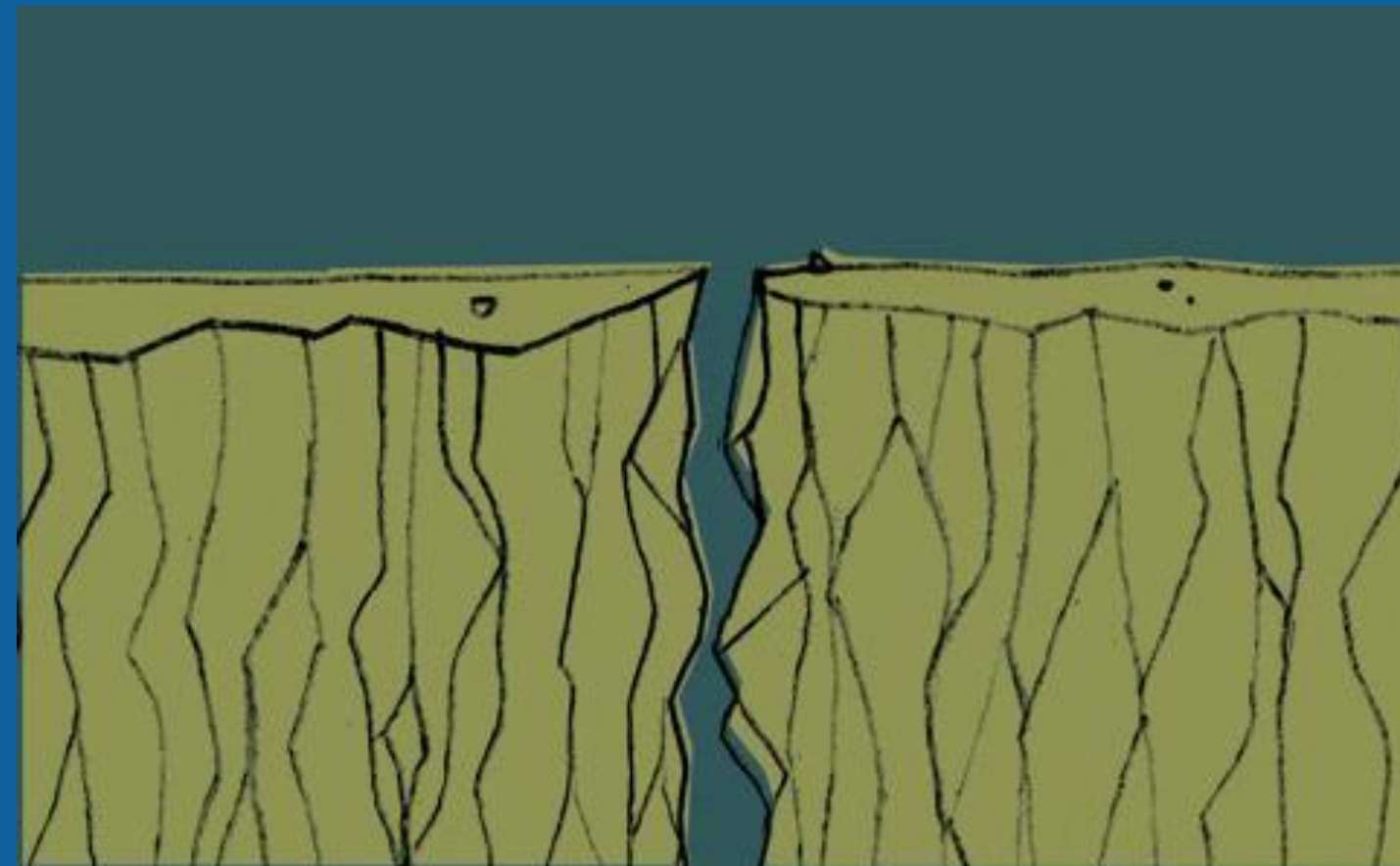


A haunting real image from the devastating earthquake that struck Italy in 2016



# Why this data?

"Seismic data analysis cannot prevent earthquakes or predict them with precision. However, it remains a vital tool in mitigating their devastating impact through informed strategies. Currently, medium-term forecasting appears to be the most achievable and realistic target."



The chosen dataset contains events from 2016-08-24 to 2016-11-30. It is a single csv file with the following header:

Time,latitude,longitude,depth/km,magnitude

# Dataset Source

The data set has been published on Kaggle and contains the official data of INGV (National Institute of Geophysics and Volcanology)

# Features

The data set contains 8087 rows (8086 of data + 1 header) For each event, the following properties are provided:

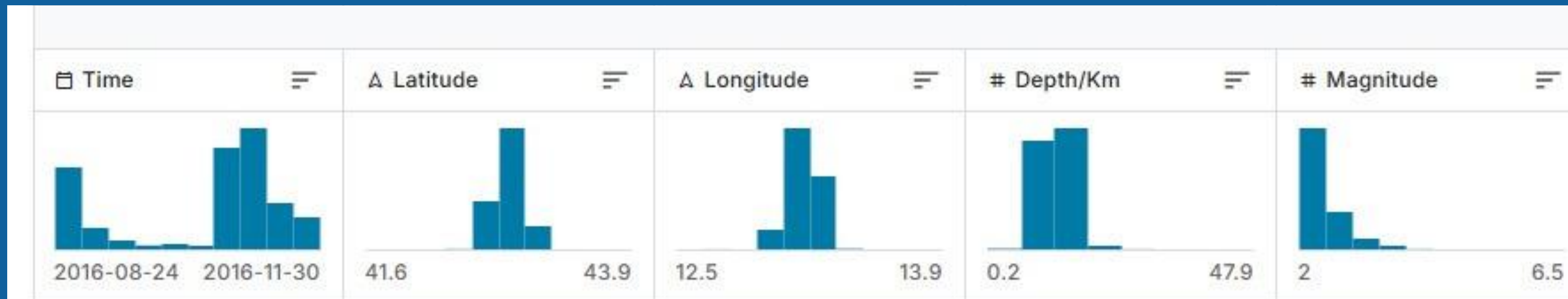
- the exact time of the event in the format "Ymd hh:mm:s. ms"
- the exact geographic coordinates of the event, in latitude and longitude
- the depth of the hypocenter in kilometres
- the magnitude value in the Richter scale



# The study aims to:

- Document seismic activity in Italy during the period indicated.
- Analyse the spatial and temporal distribution of earthquakes.
- Assess the frequency and intensity of seismic events.
- Support geophysical research and prevention of natural hazards.

The histograms provide a comprehensive representation of the data.





The data acquired has  
been analysed with  
Weka, but ...  
what is WEKA

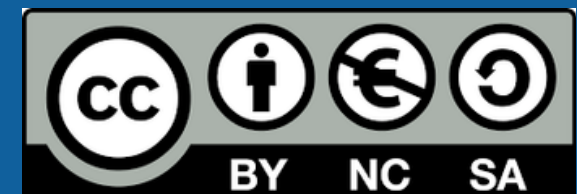




Weka (Waikato Environment for Knowledge Analysis) is an open-source data analysis and machine learning software developed by the University of Waikato, New Zealand.

### Main features:

- User-friendly graphical interface
- Extensive collection of machine learning algorithms
- Support for data pre-processing





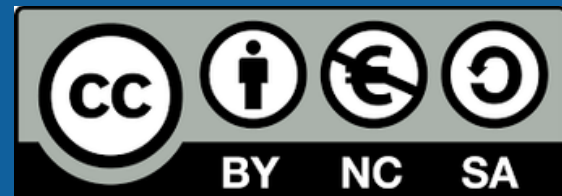
# What is Weka for?

Weka is used for:





- Data cleaning and transformation
- Data set display
- Construction and evaluation of machine learning

models (classification, regression, clustering)

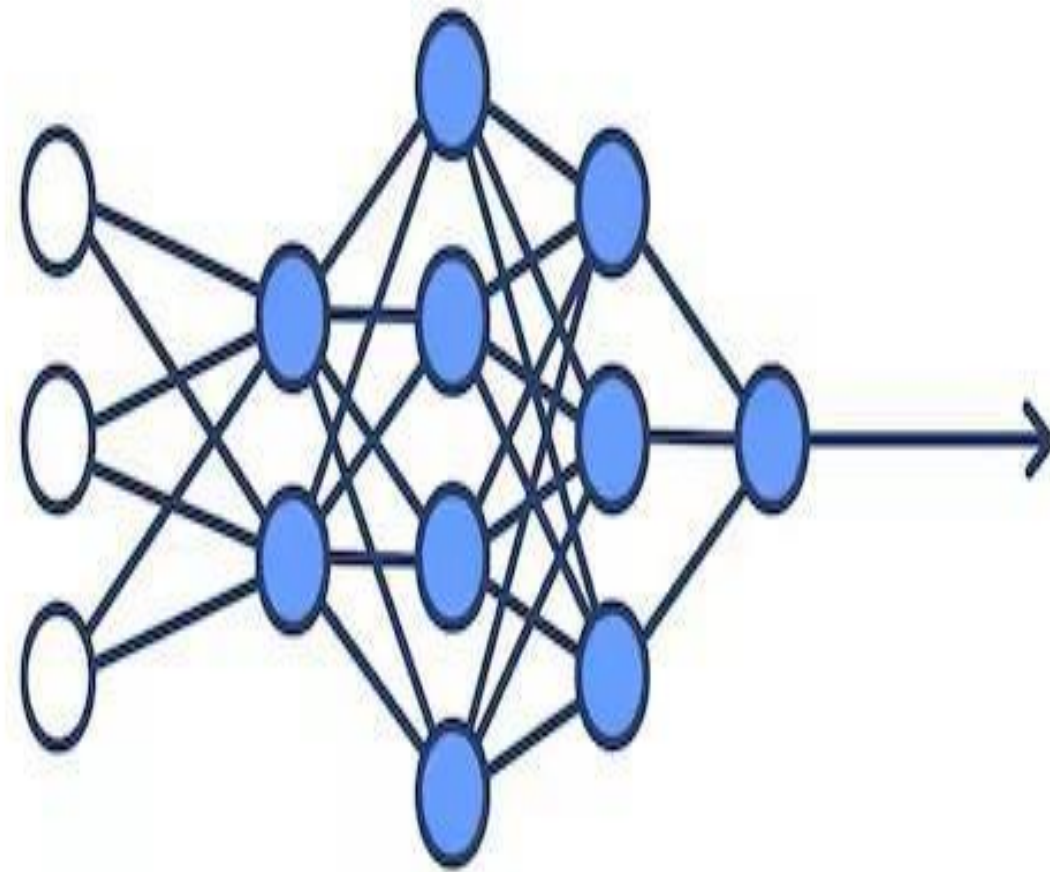
- Feature selection



# How does Weka work?

-  Import data (CSV, ARFF, database)
-  Pre-process data (normalization, management of missing values)
-  Apply an algorithm (e.g. decision tree, k-NN, Naive Bayes)
-  Evaluate performance (accuracy, ROC, confusion matrix)





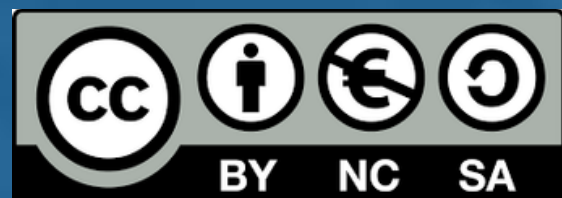
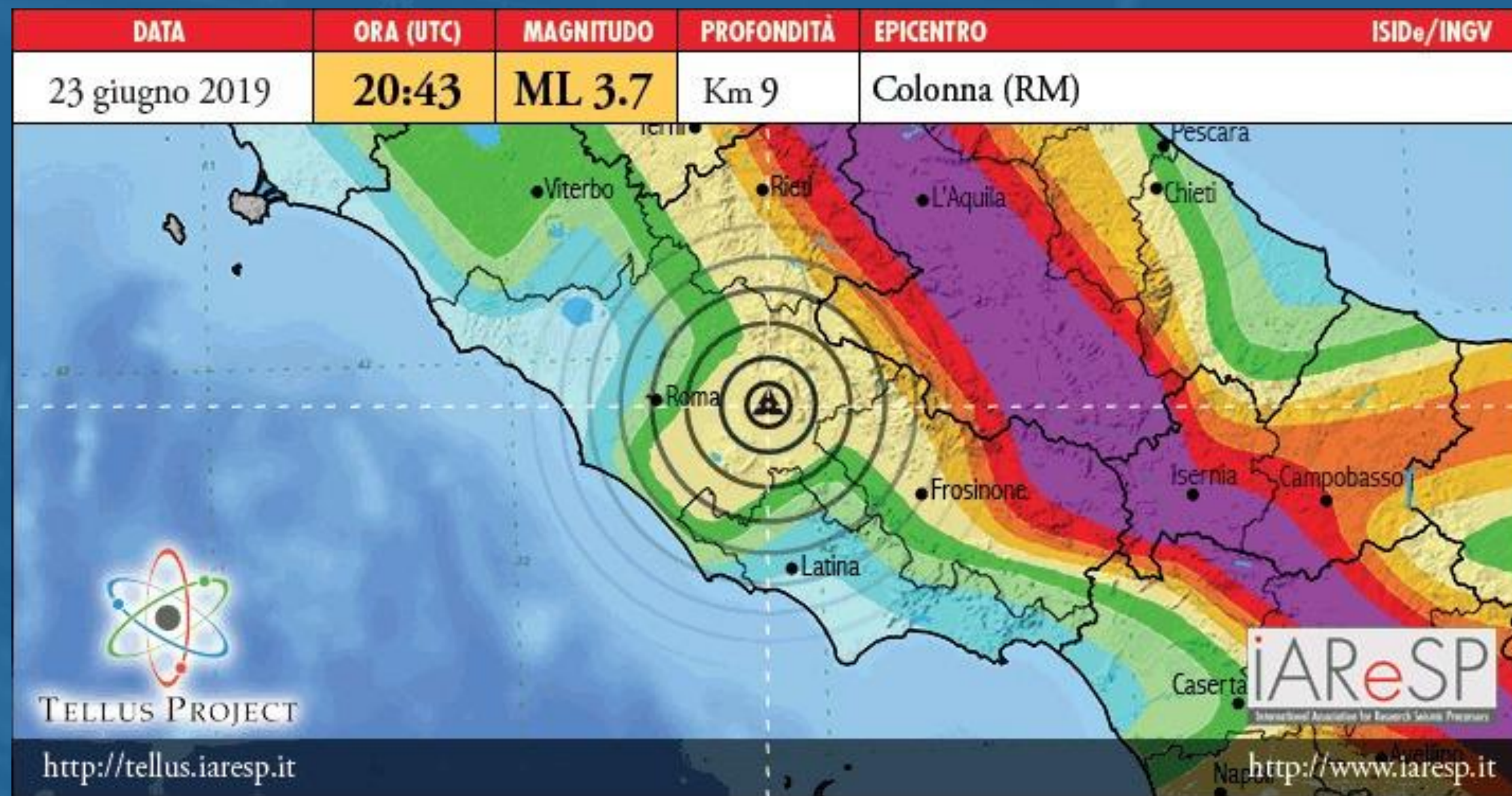
Fully connected neural  
networks (FCNNs)

# Limitations of data collection

The model obtained is not effective in predicting future earthquakes because at this time to make a reliable forecast large amounts of data are needed but in recent years thanks to machine learning have been made many advances in earthquake prediction.



# METHODOLOGY





# Data Processing



The file has been downloaded in csv. The data has been all converted to numeric except for the hours data which have been kept in their original format. The csv file was then modified and transformed into a useful format for Weka (arff).

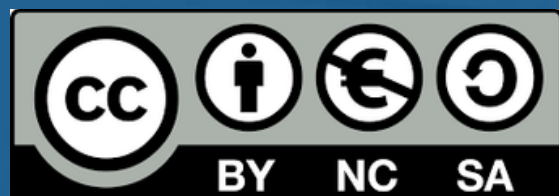
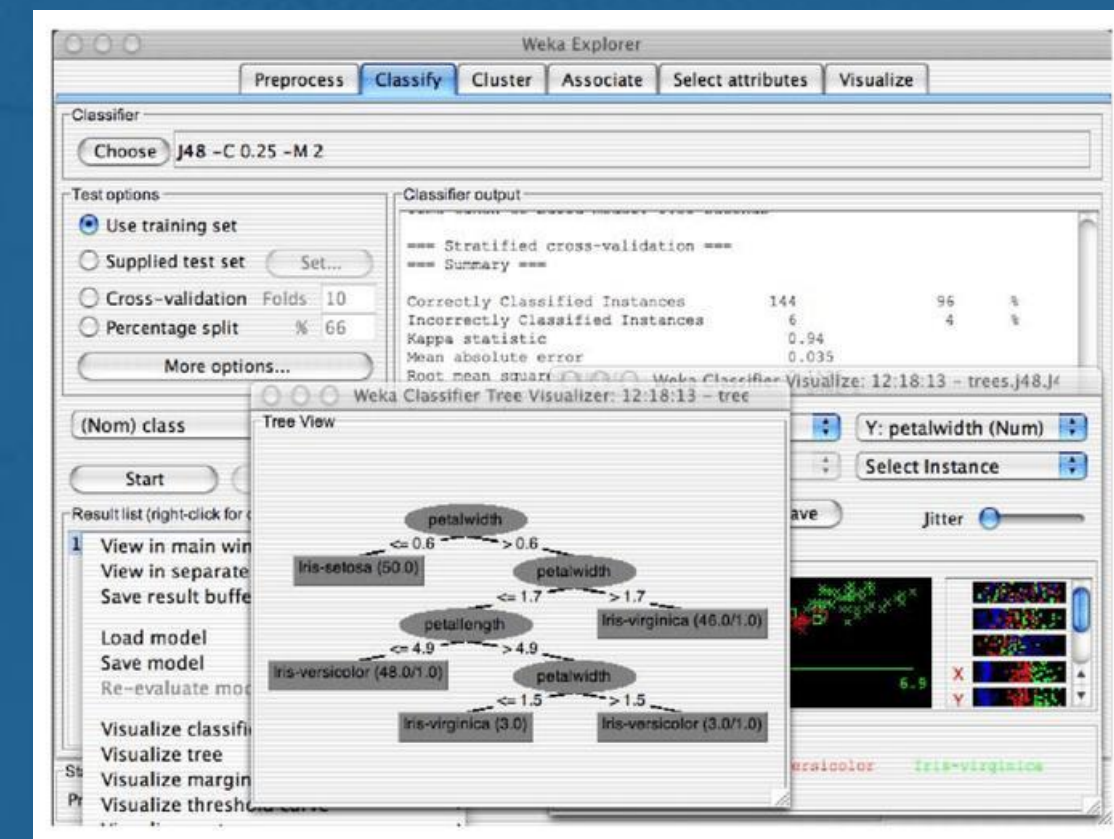
# THIS IMAGE REPRESENTS THE DEFINITION OF THE DATA IN THE ARFF FILE

```
@relation italy  
  
@attribute timestamp date yyyy-MM-dd HH:mm:ss.SSS  
@attribute latitude numeric  
@attribute longitude numeric  
@attribute depth(km) numeric  
@attribute magnitude numeric
```



# THE WEKA ALGORITHMS

APPLYING AN ALGORITHM ON WEKA MEANS USING ONE OF THE MACHINE LEARNING TOOLS PROVIDED BY WEKA TO ANALYZE A SET OF DATA AND BUILD A PREDICTIVE OR DESCRIPTIVE MODEL. THIS PROCESS ALLOWS FORECASTING OR SEGMENTATION OF DATA INTO SIGNIFICANT GROUPS



## THESE ARE THE ALGORITHMS APPLIED

Model	Type	Purpose
Linear regression	Linear model	Simple baseline
M5Rules	Trees with rules	Prediction with rules
Random Forest	Ensemble (trees)	Nonlinear prediction

THE ALGORITHMS WERE CHOSEN BECAUSE THEY TO BE THE MOST  
SUITABLE FOR THE DATA USED

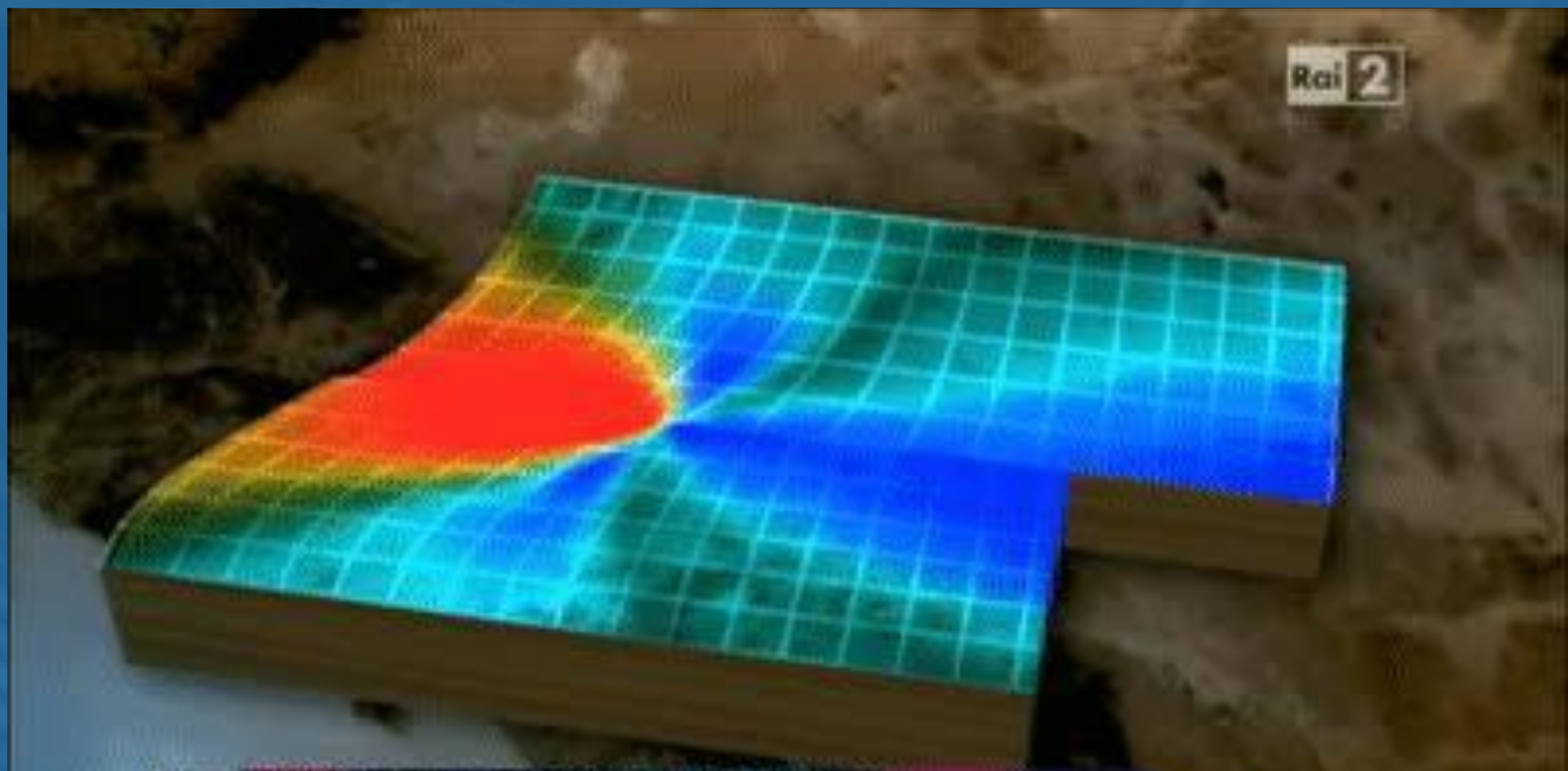


- LINEAR REGRESSION: TO PREDICT THE MAGNITUDE BASED ON DEPTH AND POSITION.
- M5RULES: DECISION TREE WITH LINEAR REGRESSION IN TERMINAL NODES.



“

## RESULTS



# Linear Regression

```
=== Cross-validation ===
```

```
=== Summary ===
```

Correlation coefficient	0.1124
Mean absolute error	0.317
Root mean squared error	0.4237
Relative absolute error	99.5694 %
Root relative squared error	99.3604 %
Total Number of Instances	8086

# M5Rules

```
=== Evaluation on training set ===
```

```
Time taken to test model on training data: 0.04 seconds
```

```
=== Summary ===
```

Correlation coefficient	0.4576
Mean absolute error	0.2802
Root mean squared error	0.3791
Relative absolute error	88.0144 %
Root relative squared error	88.9172 %
Total Number of Instances	8086





# Random Forest

```
=== Summary ===
```

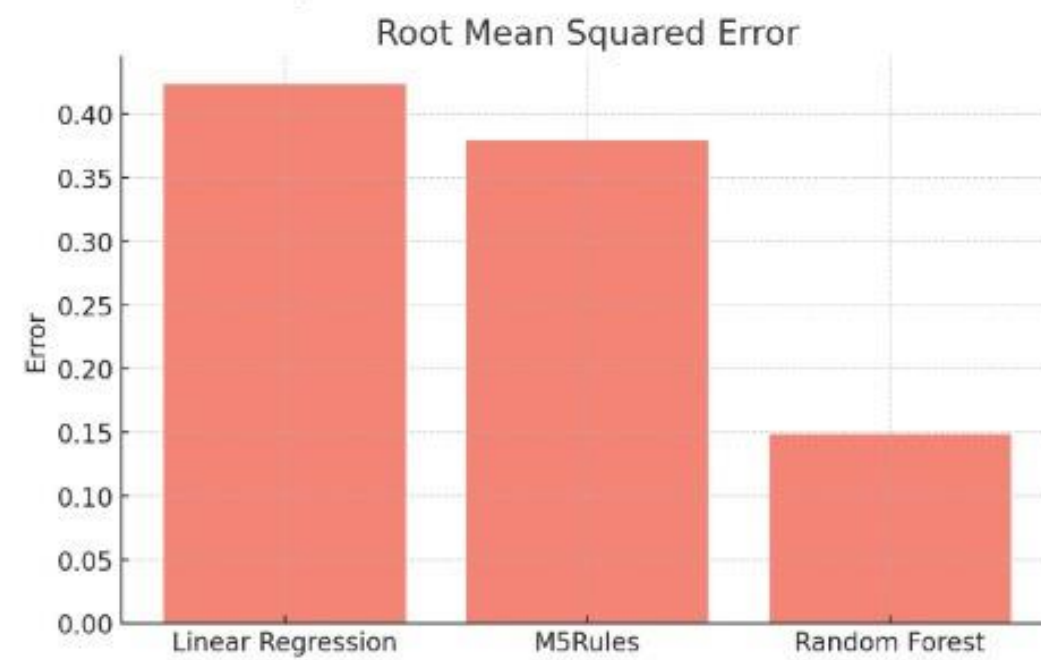
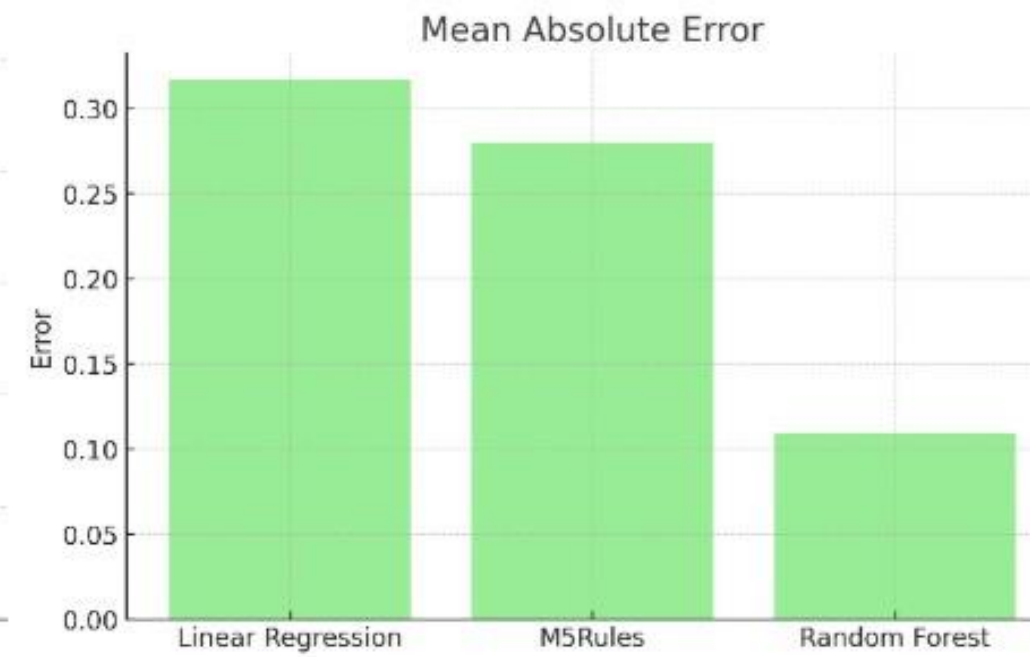
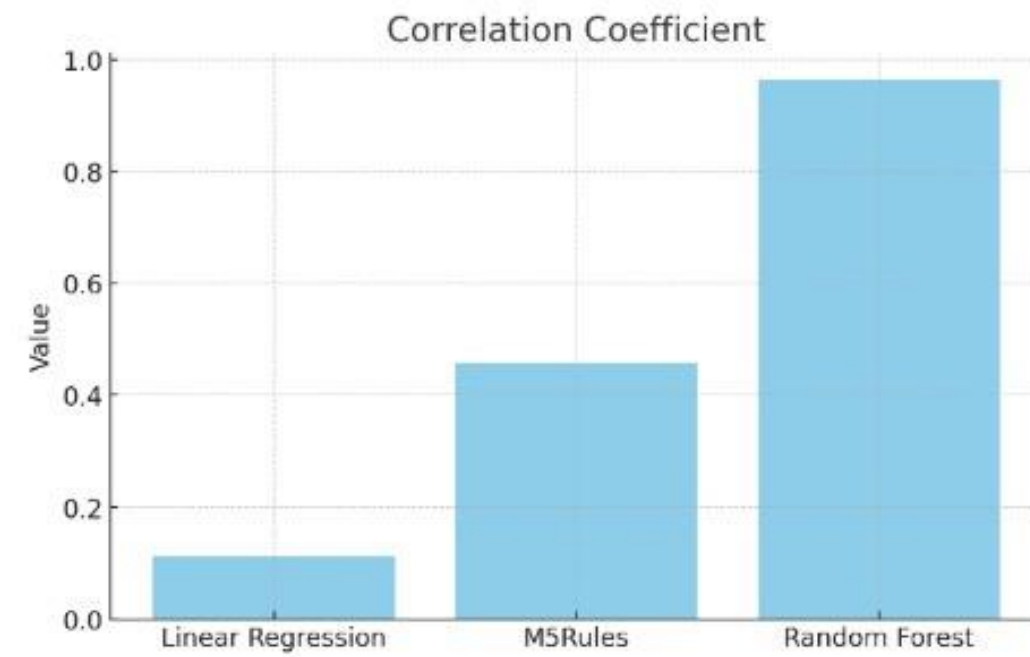
Correlation coefficient	0.964
Mean absolute error	0.1093
Root mean squared error	0.1484
Relative absolute error	34.3469 %
Root relative squared error	34.7961 %
Total Number of Instances	8086

# COMPARISON OF THE THREE ALGORITHMS APPLIED

Algorithm	Correlation	Mean Absolute Error	Root Mean Squared Error	Relative Error
Linear regression	0.1124	0.317	0.4237	99.57
M5Rules	0.4576	0.2802	0.3791	88.01
Random Forest	0.964	0.1093	0.1484	34.34

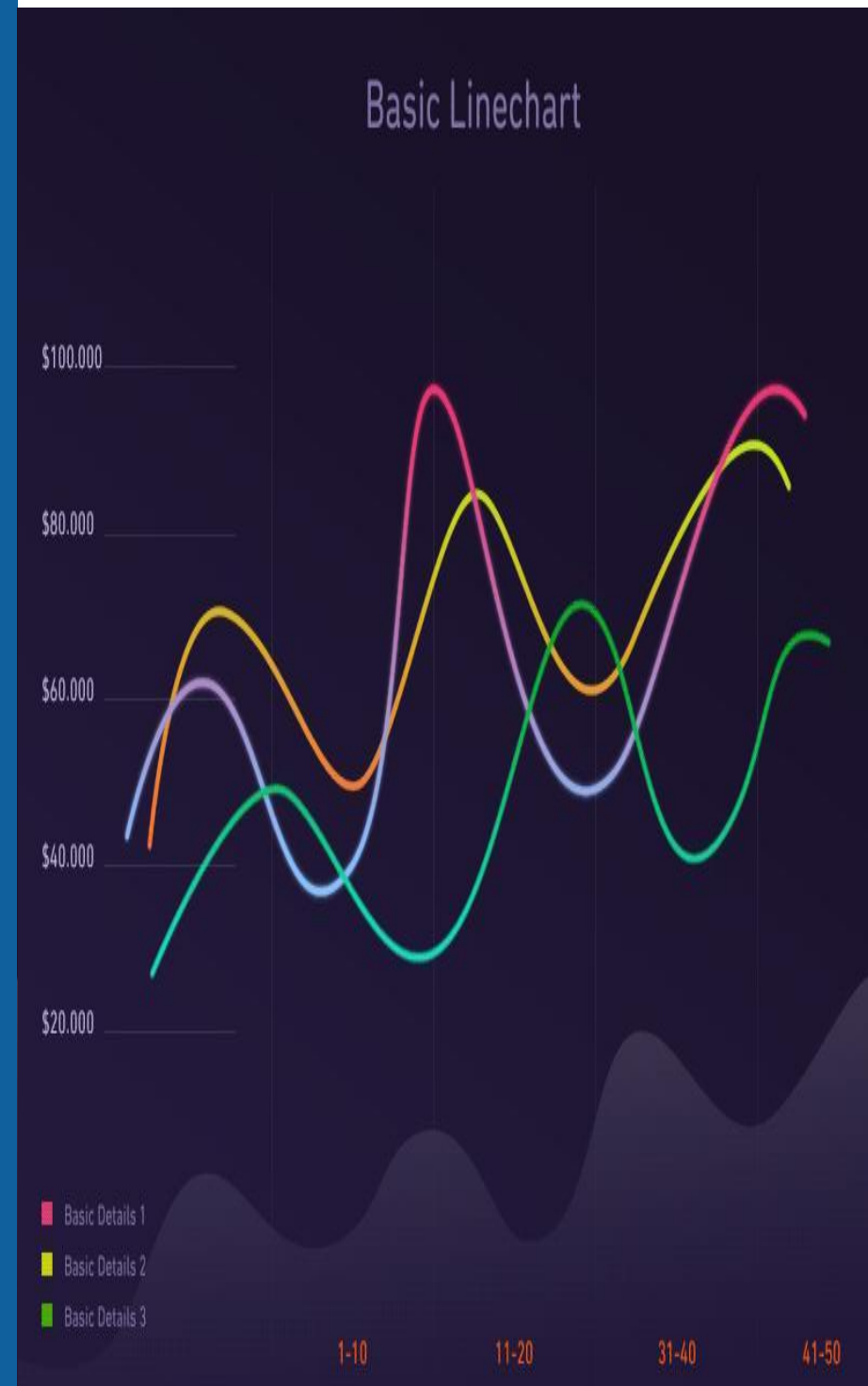


Performance Metrics by Algorithm



1.59: K

# CORRELATION COEFFICIENT



**MEASURE HOW WELL THE MODEL'S PREDICTIONS CORRELATE WITH REAL VALUES (VALUES CLOSER TO 1 INDICATE A BETTER CORRELATION)**

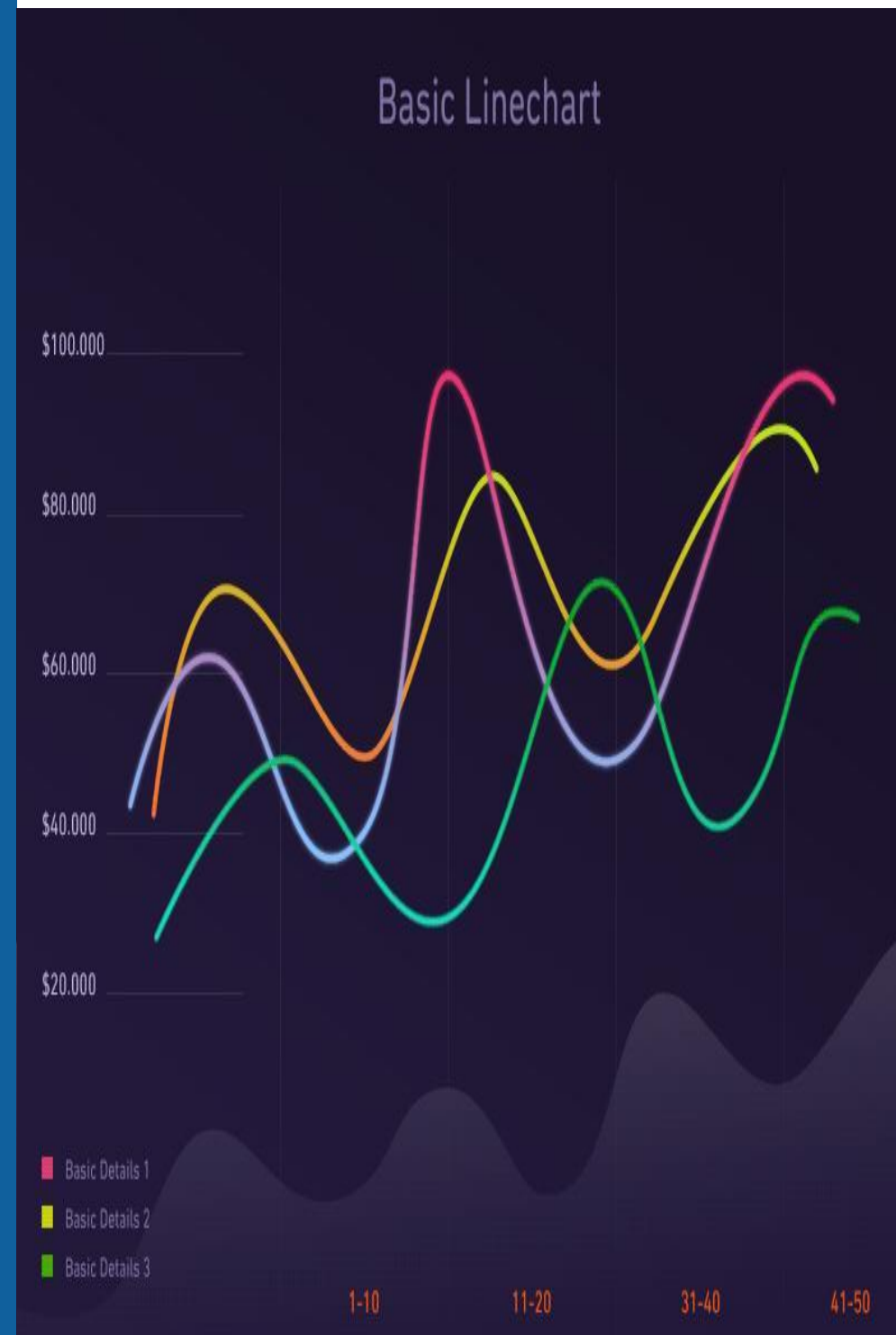
**Linear Regression: 0.1124 almost no correlation.**

**M5Rules: 0.4576 moderate correlation.**

**Random Forest: 0.964 strong correlation,**



# MEAN ABSOLUTE ERROR (MAE)



**ABSOLUTE MEAN ERROR BETWEEN  
ACTUAL AND EXPECTED VALUES  
(THE LOWER, THE BETTER)**

**Linear regression: 0.317**

**M5Rules: 0.2802**

**Random Forest: 0.1093**

# SQUARED ERROR ROOT MEAN (RMSE)



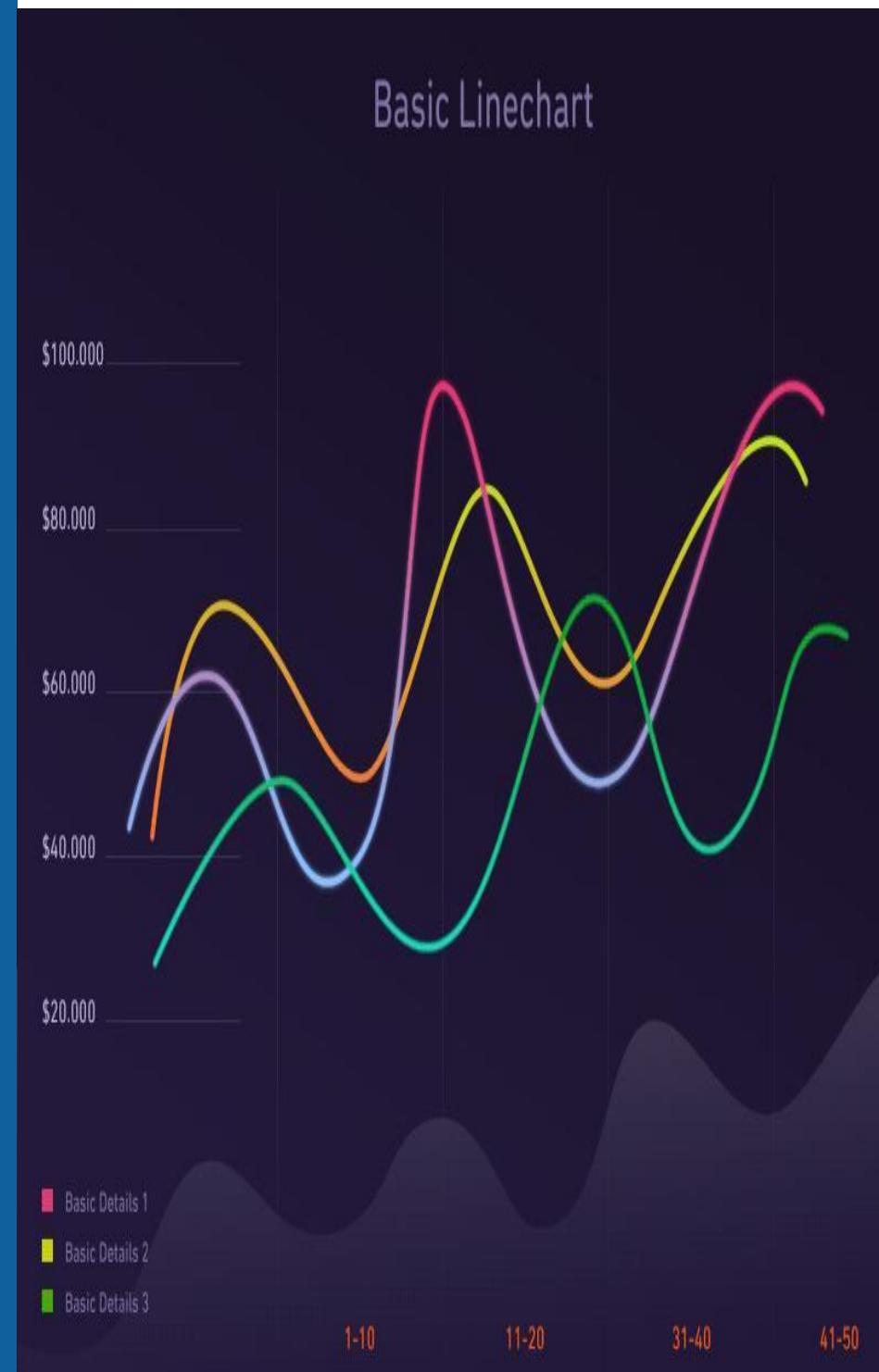
**SIMILAR TO THE MAE, BUT  
PENALIZES LARGER ERRORS  
MORE.**

**Linear regression: 0.4237**

**M5Rules: 0.3791**

**Random Forest: 0.1484**

# RELATIVE ERROR



**RELATIVE ERROR IN PERCENTAGE  
COMPARED TO A REFERENCE  
MODEL (THE LOWER, THE  
BETTER).**

**Linear regression: 99.57%**

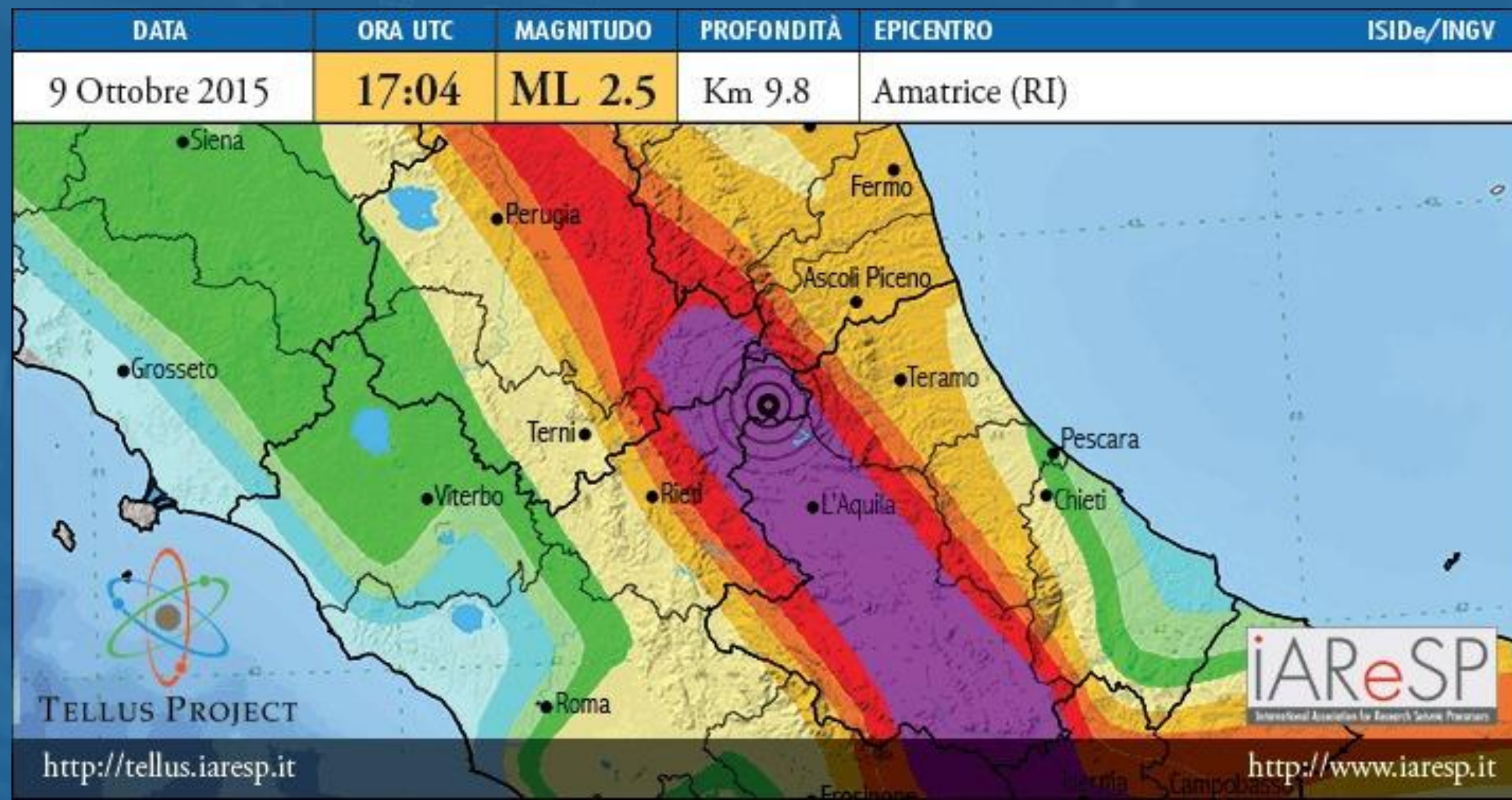
**M5Rules: 88.01%**

**Random Forest: 34.34%**



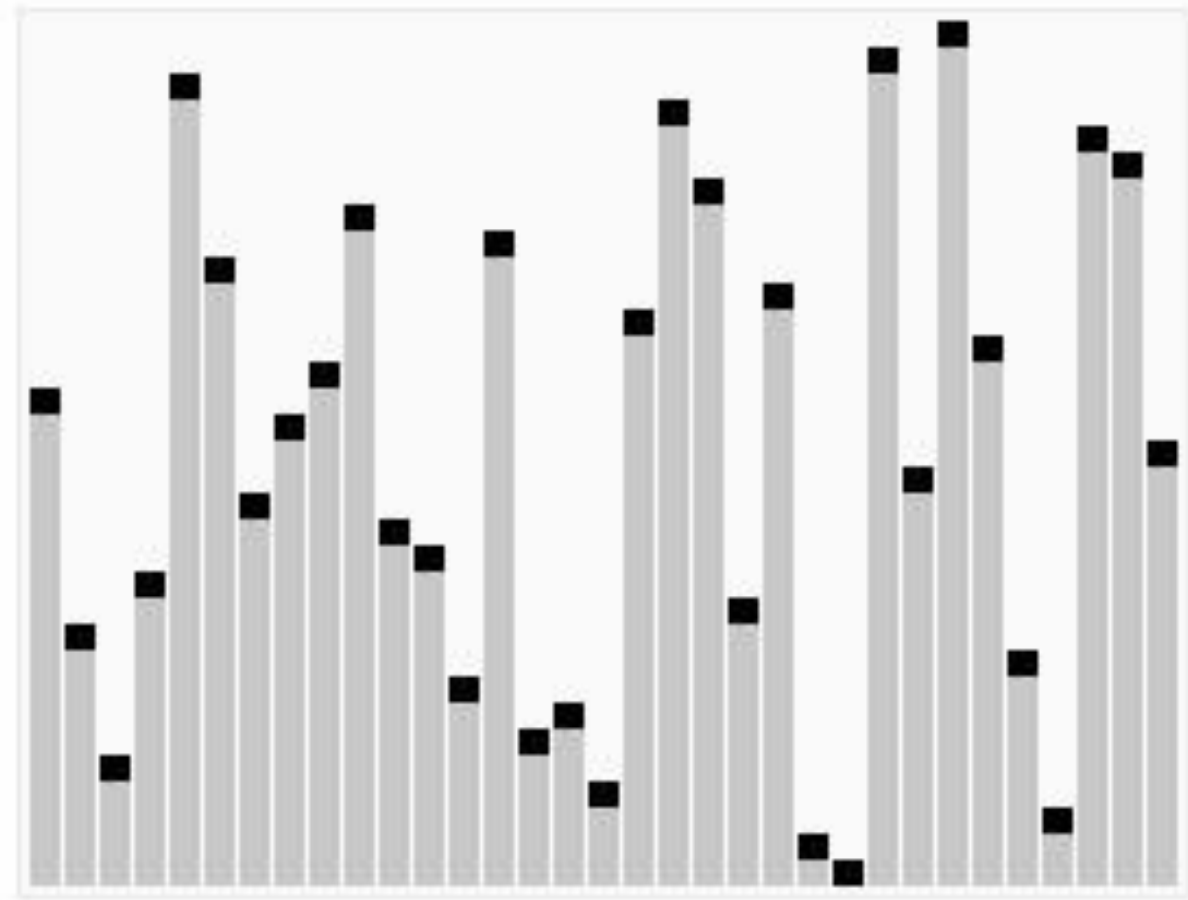


# CONCLUSIONS



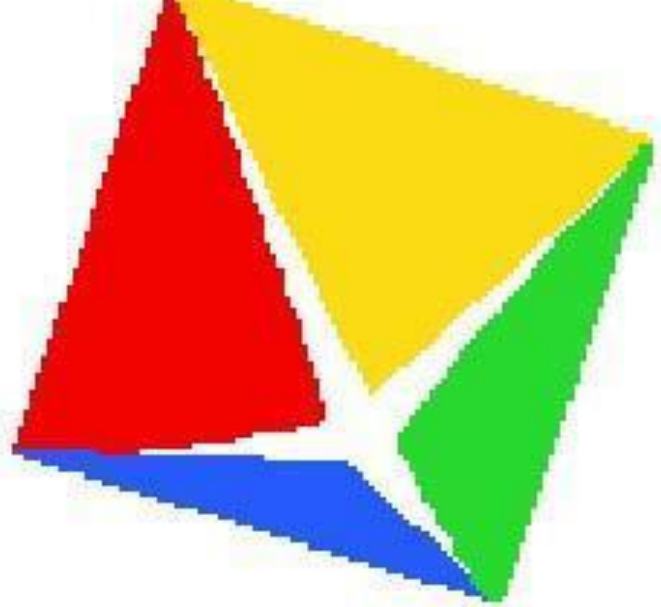
Metric	Best algorithm
Correlation coeff.	Random Forest
MAE	Random Forest
RMSE	Random Forest
Relative Error	Random Forest





- **RANDOM FOREST** IS CLEARLY THE BEST ALGORITHM FOR THIS DATASET, WITH SUPERIOR PERFORMANCE IN ALL METRICS.
- **LINEAR REGRESSION** PERFORMS VERY BADLY (ALMOST RANDOM)
- **M5RULES** IS AN IMPROVEMENT, BUT STILL FAR FROM THE RANDOM FOREST.





# ISTITUTO DI ISTRUZIONE SUPERIORE EINAUDI PARETO



Co-funded by the  
Erasmus+ Programme  
of the European Union

