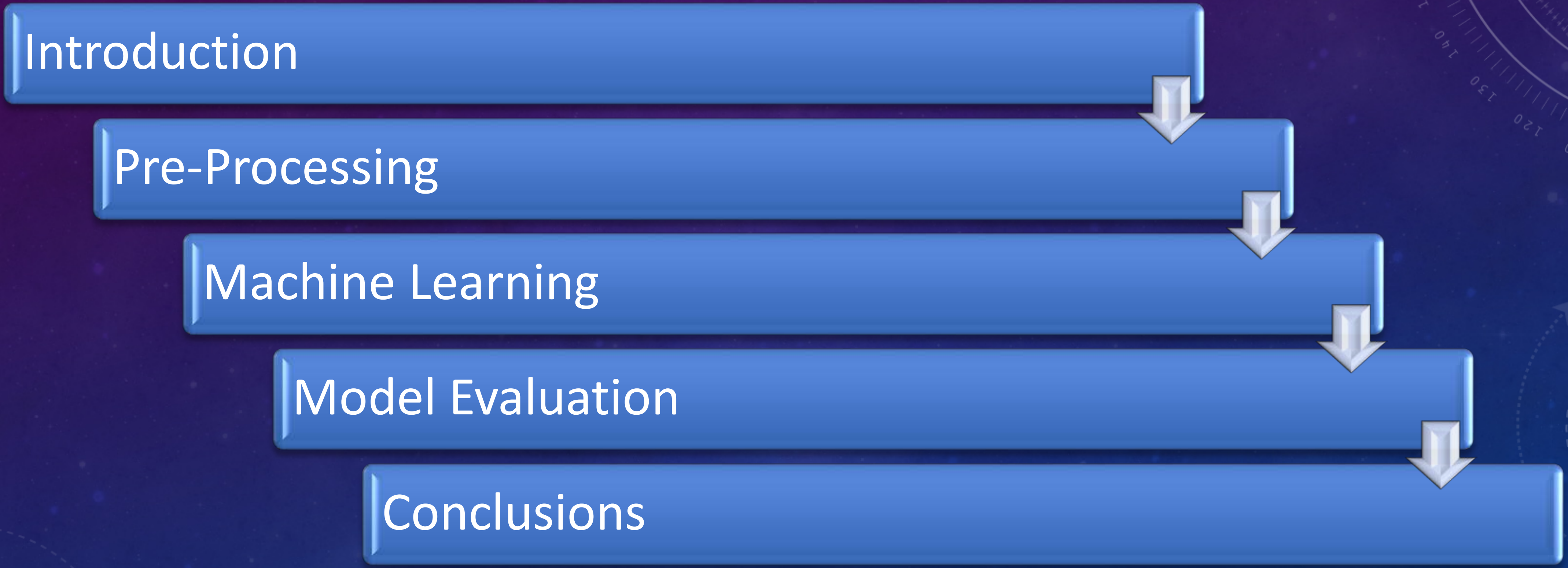
Machine learning methods for the analysis of data of an Electricity Distribution Network Operator

MASTER'S THESIS

Aristotle University of Thessaloniki, Faculty of Sciences, Department of Informatics
Supervisor: Dr. Eleftherios Angelis; Thesis Committee: Grigorios Tsoumakas, Ioannis Vlahavas

LAYOUT



Introduction

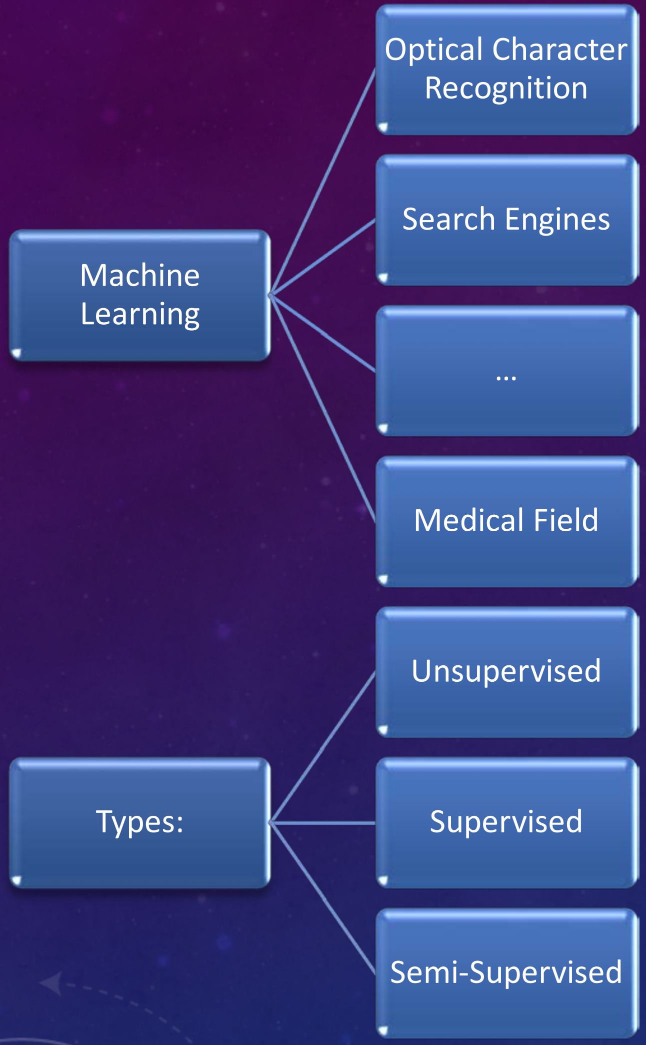
Pre-Processing

Machine Learning

Model Evaluation

Conclusions

INTRODUCTION [1/2]



Machine Learning

Optical Character Recognition

Search Engines

…

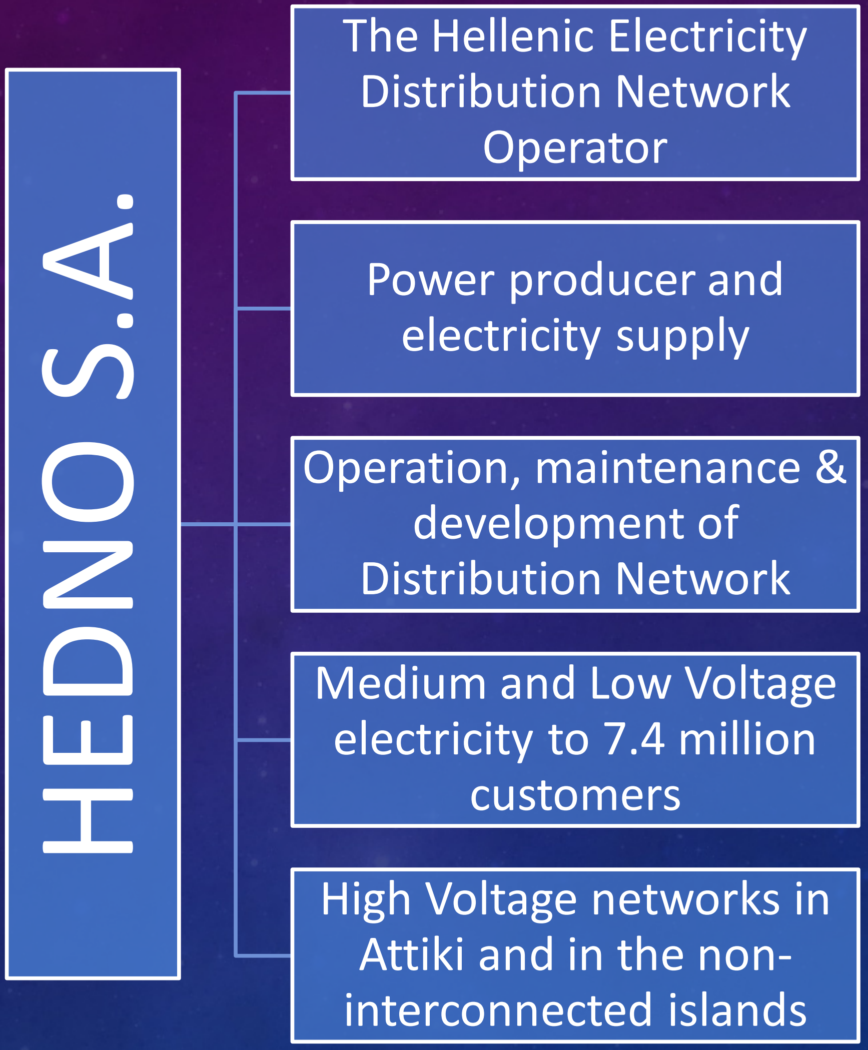
Medical Field

Types:

Unsupervised

Supervised

Semi-Supervised



HEDNO S.A.

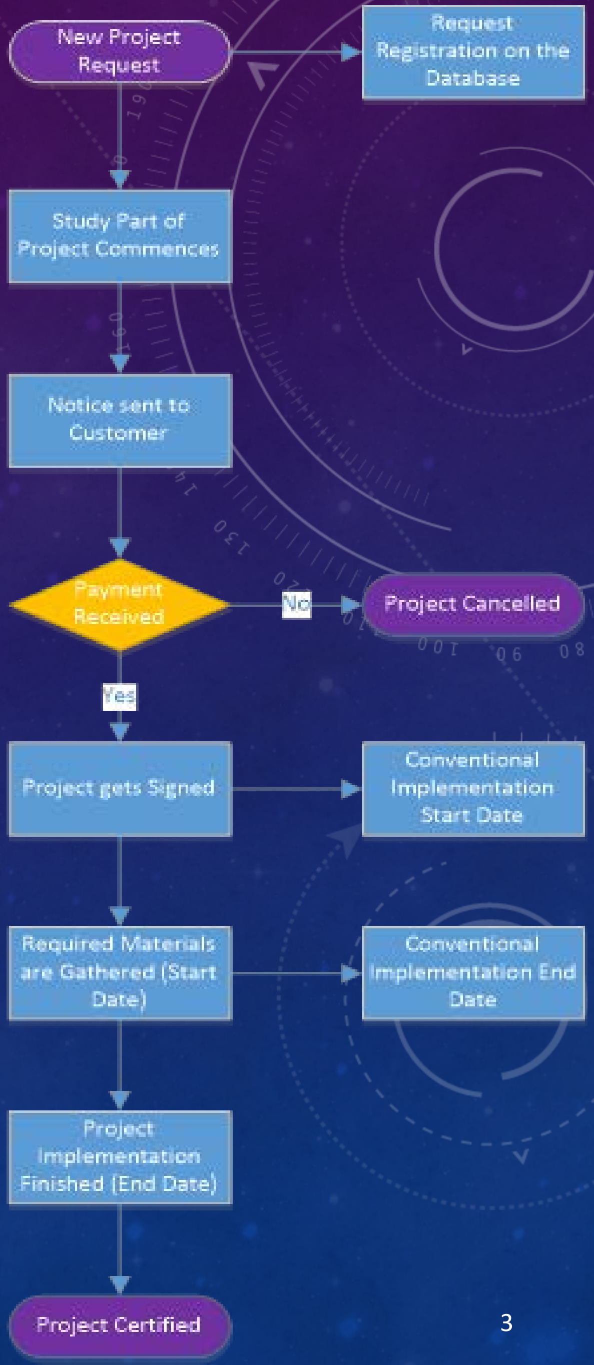
The Hellenic Electricity Distribution Network Operator

Power producer and electricity supply

Operation, maintenance & development of Distribution Network

Medium and Low Voltage electricity to 7.4 million customers

High Voltage networks in Attiki and in the non-interconnected islands



New Project Request

Request Registration on the Database

Study Part of Project Commences

Notice sent to Customer

Payment Received

No

Project Cancelled

Yes

Project gets Signed

Conventional Implementation Start Date

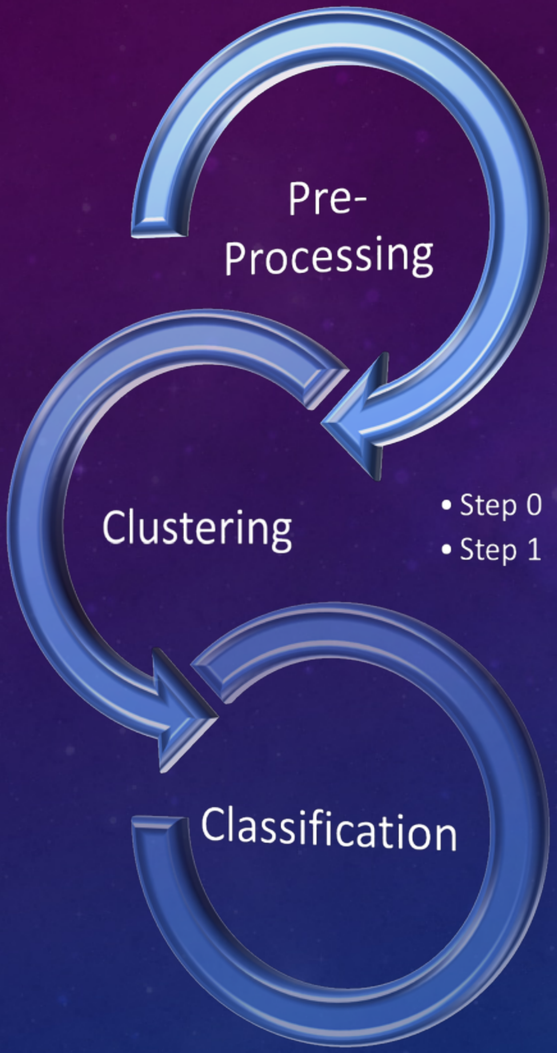
Required Materials are Gathered (Start Date)

Conventional Implementation End Date

Project Implementation Finished (End Date)

Project Certified

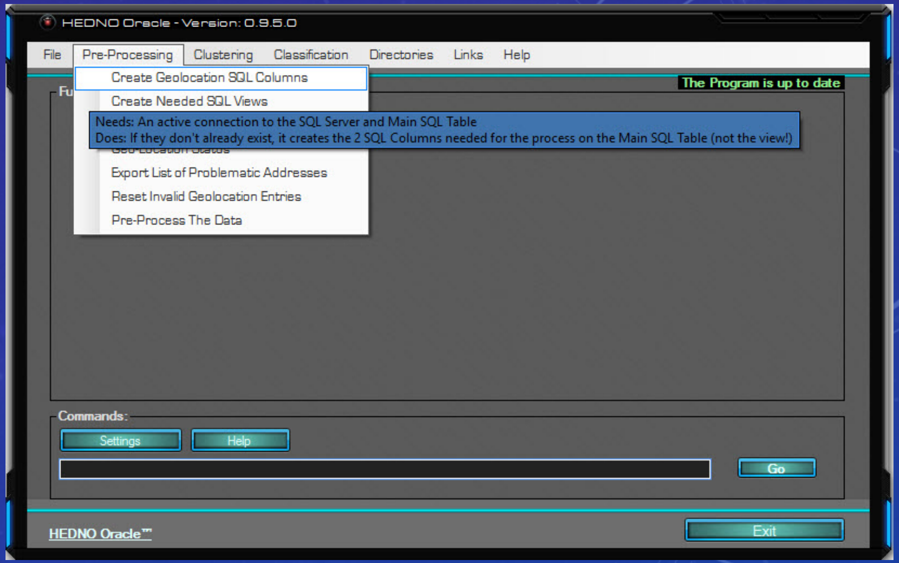
INTRODUCTION [2/2]



- Geolocation
- Check for Problematic Entries
- Pre-Process

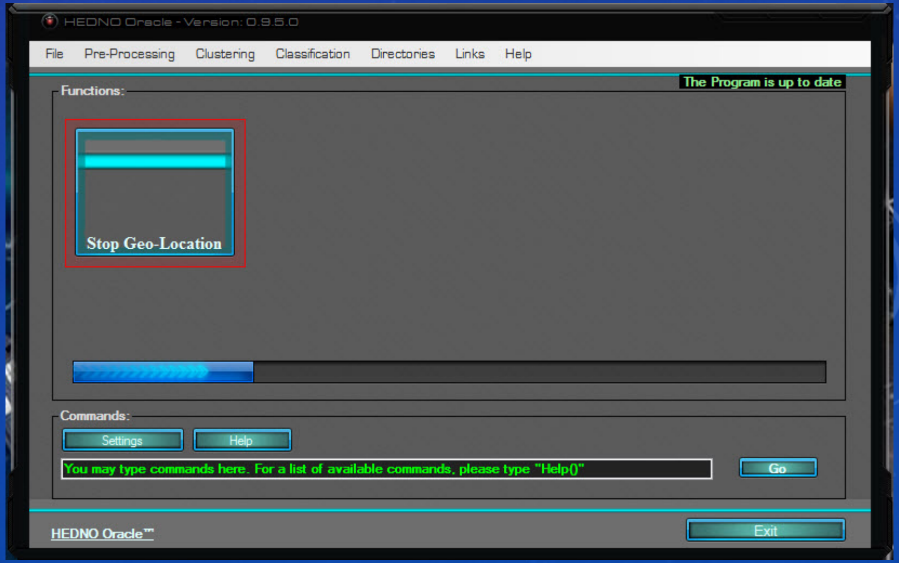
- Step 0
- Step 1

- Form Training & Testing Sets
- Model Training
- Predictions & Statistics



The Program is up to date

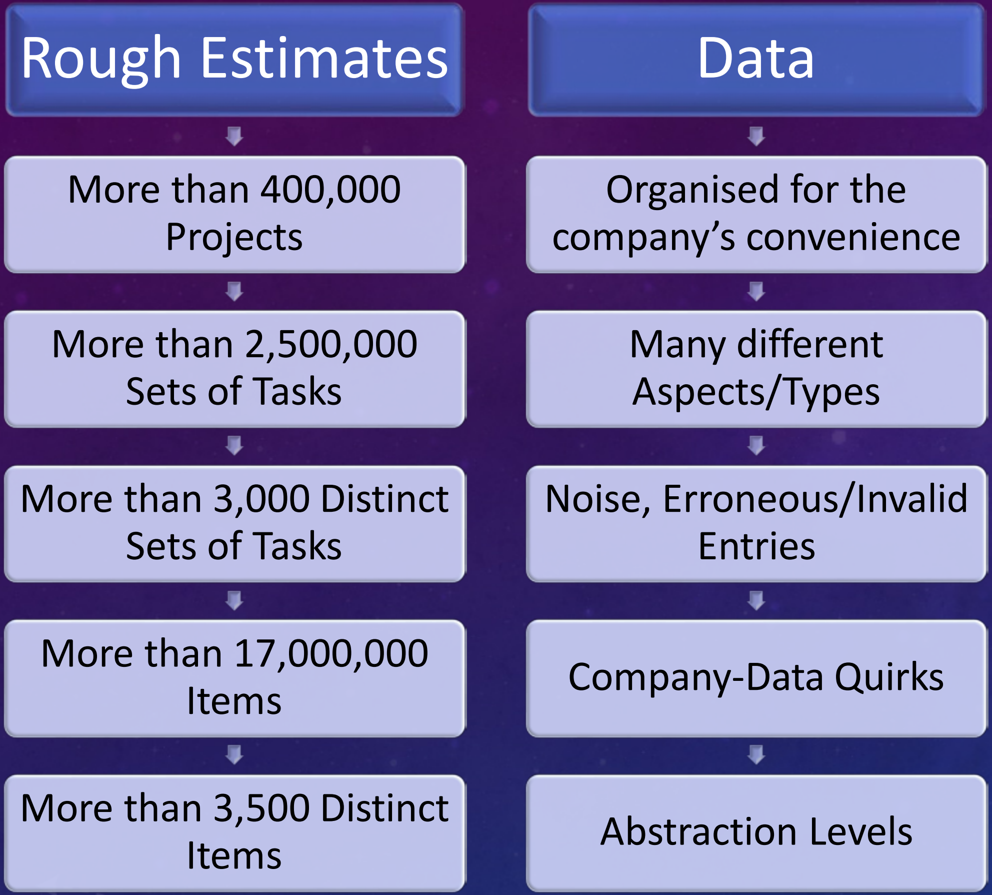
Needs: An active connection to the SQL Server and Main SQL Table
Does: If they don't already exist, it creates the 2 SQL Columns needed for the process on the Main SQL Table (not the view!)



The Program is up to date

You may type commands here. For a list of available commands, please type "Help()"

PRE-PROCESSING [1/2]



Rough Estimates

More than 400,000 Projects

More than 2,500,000 Sets of Tasks

More than 3,000 Distinct Sets of Tasks

More than 17,000,000 Items

More than 3,500 Distinct Items

Data

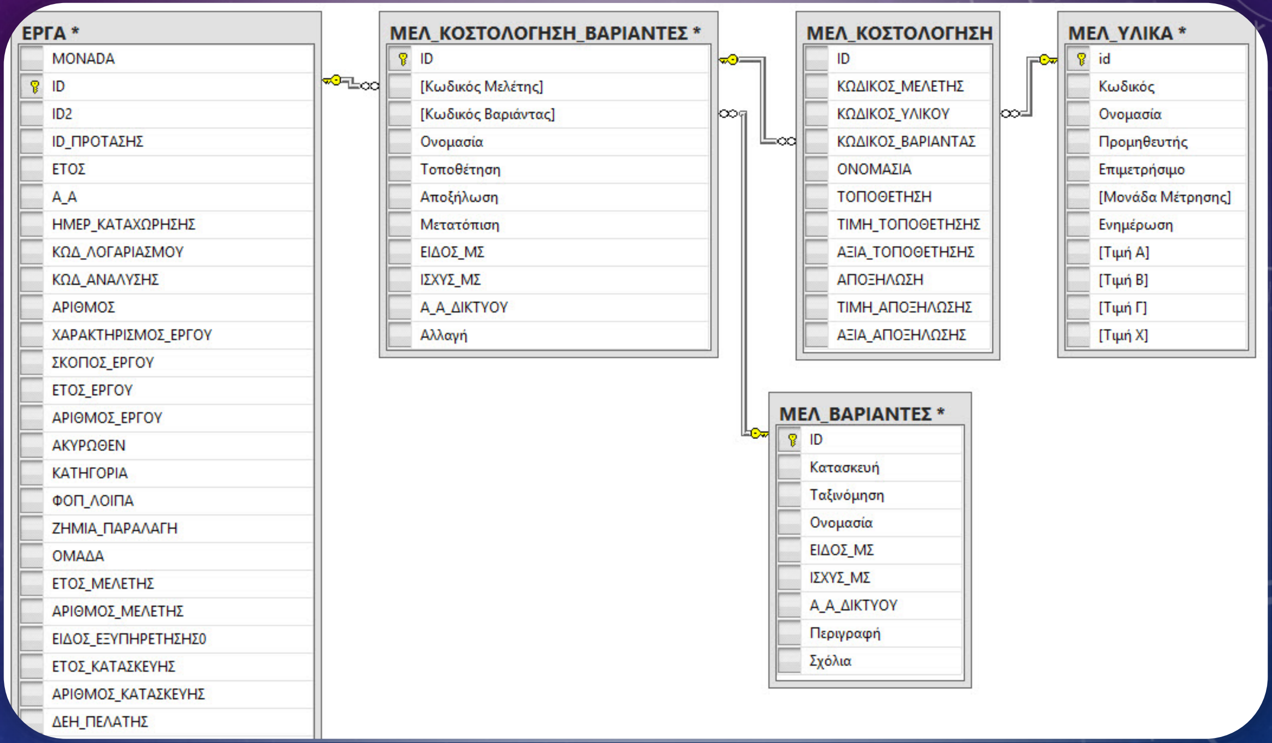
Organised for the company’s convenience

Many different Aspects/Types

Noise, Erroneous/Invalid Entries

Company-Data Quirks

Abstraction Levels



ΕΡΓΑ *
- ΜΟΝΑDΑ
- ID
- ID2
- ID_ΠΡΟΤΑΣΗΣ
- ΕΤΟΣ
- Α_Α
- ΗΜΕΡ_ΚΑΤΑΧΩΡΗΣΗΣ
- ΚΩΔ_ΛΟΓΑΡΙΑΣΜΟΥ
- ΚΩΔ_ΑΝΑΛΥΣΗΣ
- ΑΡΙΘΜΟΣ
- ΧΑΡΑΚΤΗΡΙΣΜΟΣ_ΕΡΓΟΥ
- ΣΚΟΠΟΣ_ΕΡΓΟΥ
- ΕΤΟΣ_ΕΡΓΟΥ
- ΑΡΙΘΜΟΣ_ΕΡΓΟΥ
- ΑΚΥΡΩΘΕΝ
- ΚΑΤΗΓΟΡΙΑ
- ΦΟΠ_ΛΟΙΠΑ
- ΖΗΜΙΑ_ΠΑΡΑΛΑΓΗ
- ΟΜΑΔΑ
- ΕΤΟΣ_ΜΕΛΕΤΗΣ
- ΑΡΙΘΜΟΣ_ΜΕΛΕΤΗΣ
- ΕΙΔΟΣ_ΕΞΥΠΗΡΕΤΗΣΗΣ0
- ΕΤΟΣ_ΚΑΤΑΣΚΕΥΗΣ
- ΑΡΙΘΜΟΣ_ΚΑΤΑΣΚΕΥΗΣ
- ΔΕΗ_ΠΕΛΑΤΗΣ

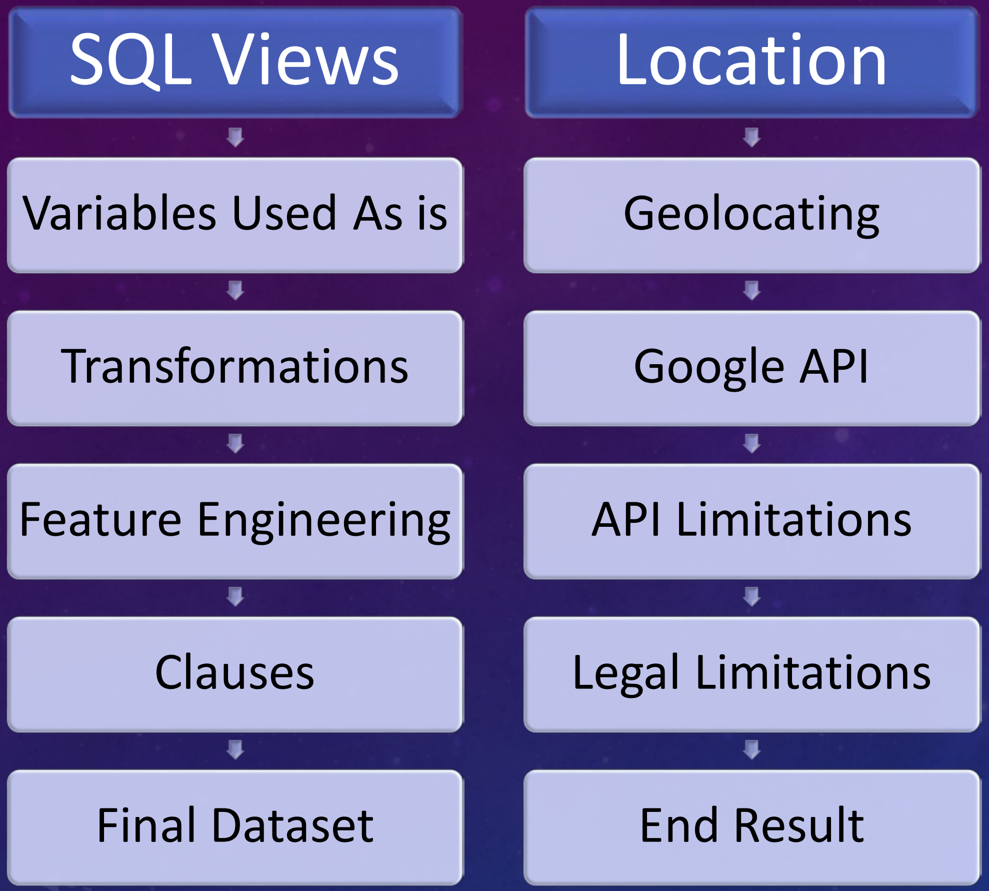
ΜΕΛ_ΚΟΣΤΟΛΟΓΗΣΗ_ΒΑΡΙΑΝΤΕΣ *
- ID
- [Κωδικός Μελέτης]
- [Κωδικός Βαριάντας]
- Ονομασία
- Τοποθέτηση
- Αποξήλωση
- Μετατόπιση
- ΕΙΔΟΣ_ΜΣ
- ΙΣΧΥΣ_ΜΣ
- Α_Α_ΔΙΚΤΥΟΥ
- Αλλαγή

ΜΕΛ_ΚΟΣΤΟΛΟΓΗΣΗ
- ID
- ΚΩΔΙΚΟΣ_ΜΕΛΕΤΗΣ
- ΚΩΔΙΚΟΣ_ΥΛΙΚΟΥ
- ΚΩΔΙΚΟΣ_ΒΑΡΙΑΝΤΑΣ
- ΟΝΟΜΑΣΙΑ
- ΤΟΠΟΘΕΤΗΣΗ
- ΤΙΜΗ_ΤΟΠΟΘΕΤΗΣΗΣ
- ΑΞΙΑ_ΤΟΠΟΘΕΤΗΣΗΣ
- ΑΠΟΞΗΛΩΣΗ
- ΤΙΜΗ_ΑΠΟΞΗΛΩΣΗΣ
- ΑΞΙΑ_ΑΠΟΞΗΛΩΣΗΣ

ΜΕΛ_ΥΛΙΚΑ *
- id
- Κωδικός
- Ονομασία
- Προμηθευτής
- Επιμετρήσιμο
- [Μονάδα Μέτρησης]
- Ενημέρωση
- [Τιμή Α]
- [Τιμή Β]
- [Τιμή Γ]
- [Τιμή Χ]

ΜΕΛ_ΒΑΡΙΑΝΤΕΣ *
- ID
- Κατασκευή
- Ταξινόμηση
- Ονομασία
- ΕΙΔΟΣ_ΜΣ
- ΙΣΧΥΣ_ΜΣ
- Α_Α_ΔΙΚΤΥΟΥ
- Περιγραφή
- Σχόλια

PRE-PROCESSING [2/2]



SQL Views

Variables Used As is

Transformations

Feature Engineering

Clauses

Final Dataset

Location

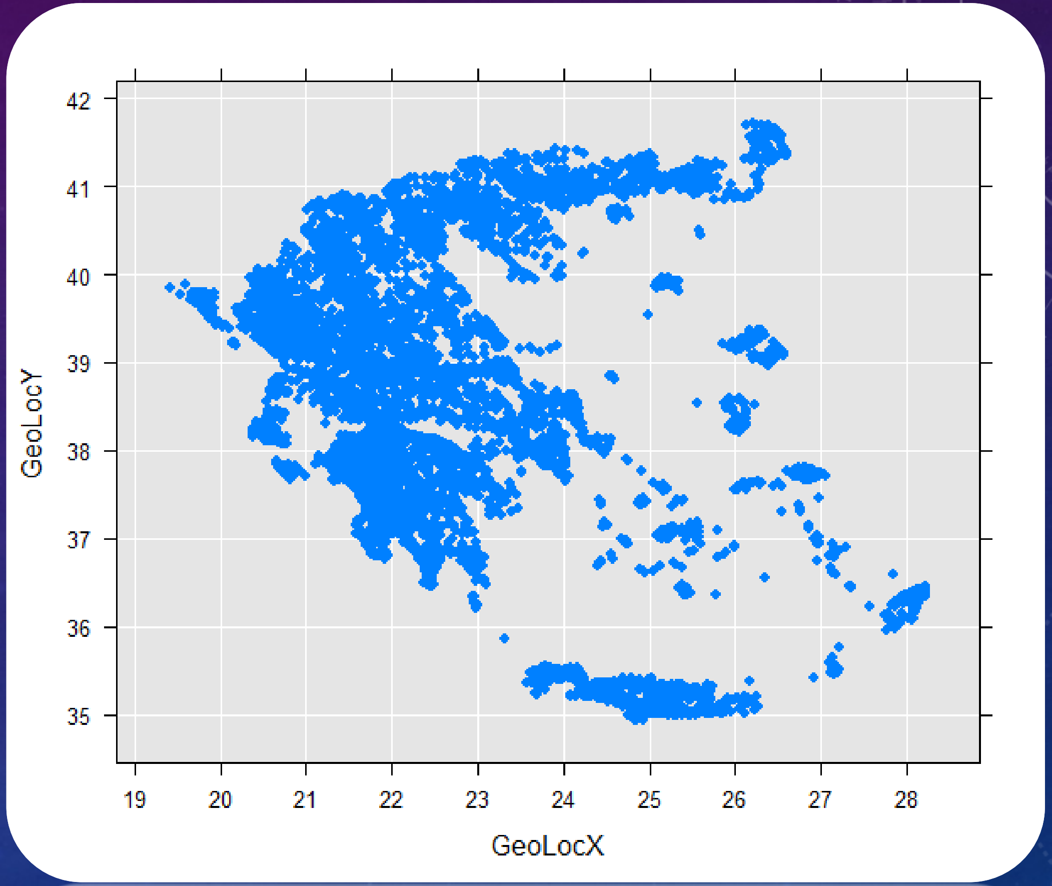
Geolocating

Google API

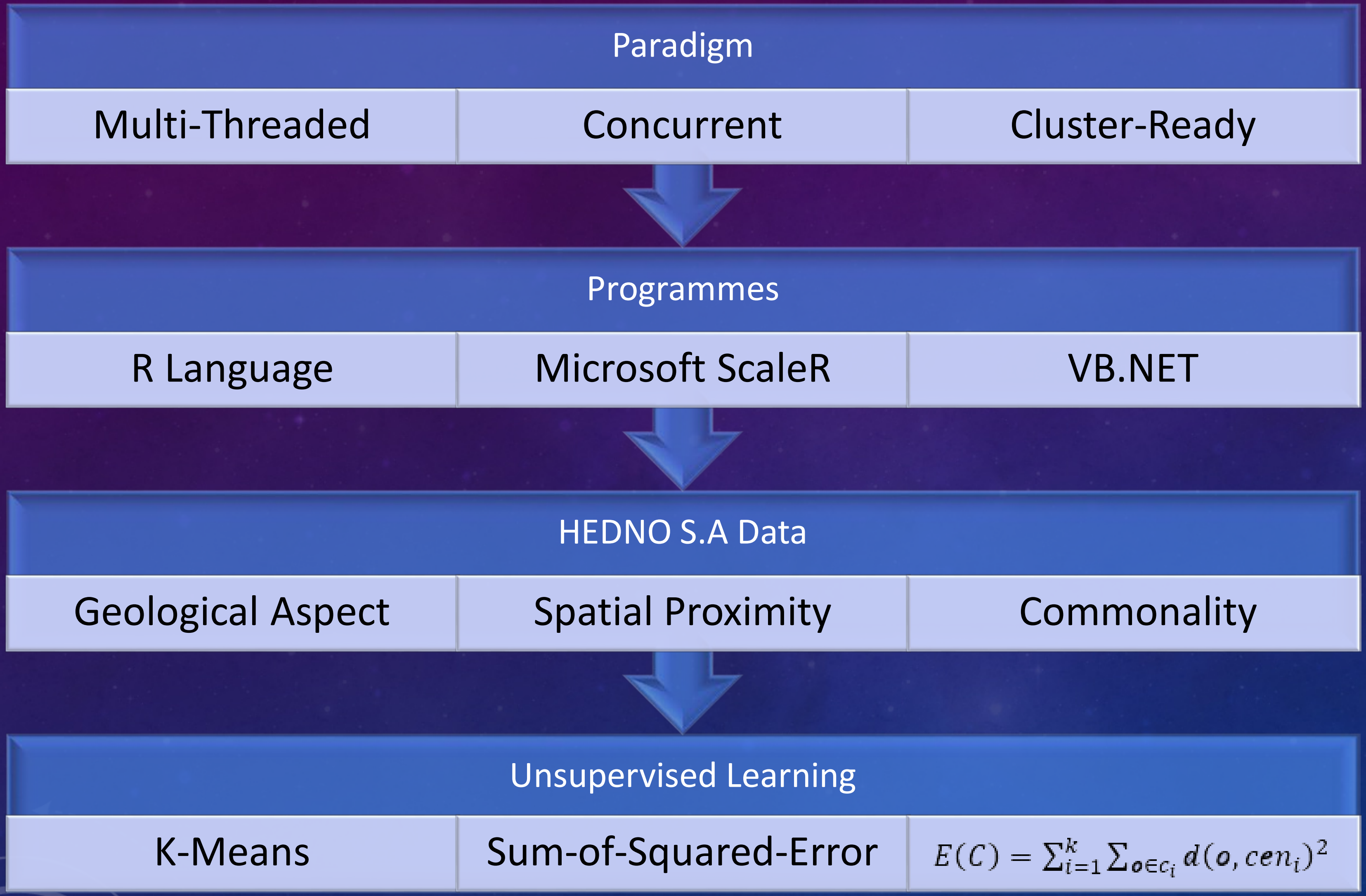
API Limitations

Legal Limitations

End Result



MACHINE LEARNING [1/3]



Paradigm

Multi-Threaded | Concurrent | Cluster-Ready

Programmes

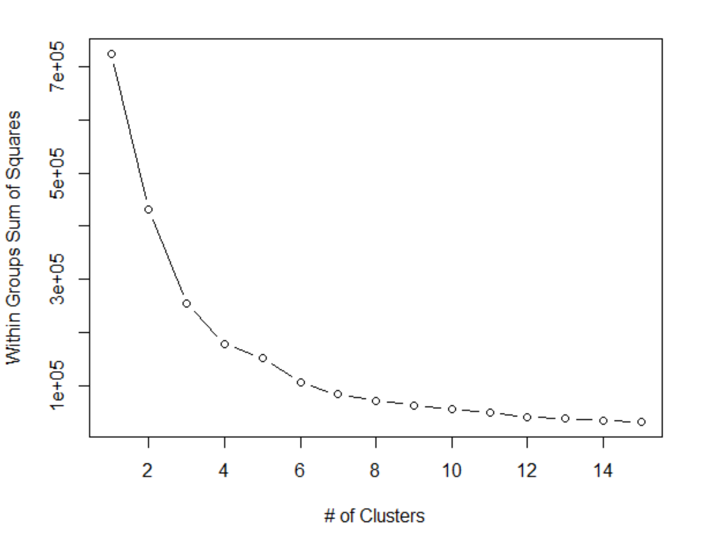
R Language | Microsoft ScaleR | VB.NET

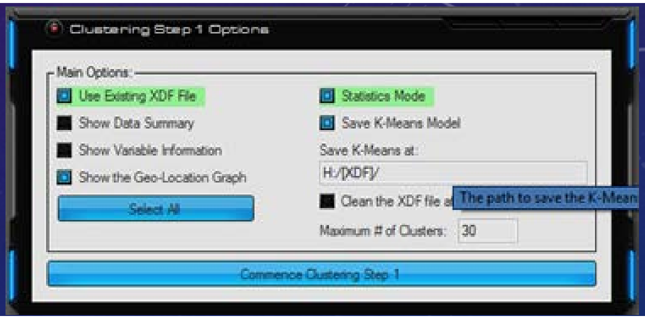
HEDNO S.A Data

Geological Aspect | Spatial Proximity | Commonality

Unsupervised Learning

K-Means | Sum-of-Squared-Error | $E(C) = \sum_{i=1}^{k} \sum_{o \in c_i} d(o, cen_i)^2$





Clustering Step 1 Options

Main Options:

☑ Use Existing XDF File

☐ Show Data Summary

☐ Show Variable Information

☑ Show the Geo-Location Graph

☑ Statistics Mode

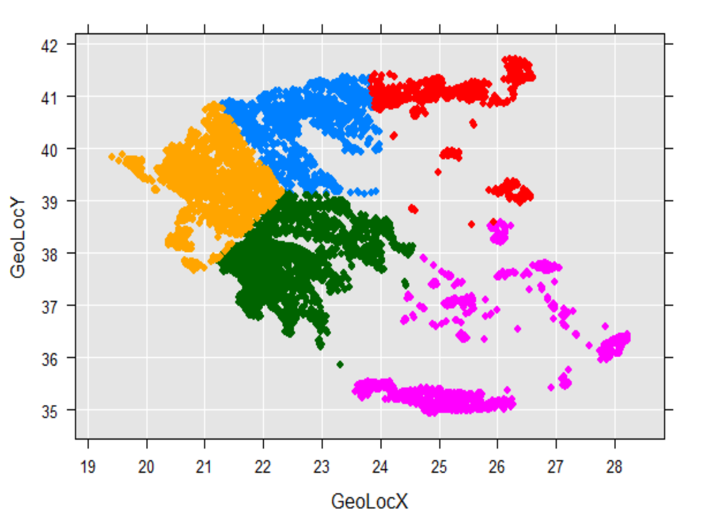
☑ Save K-Means Model

Save K-Means at: H:/[XDF]/

☐ Clean the XDF file a

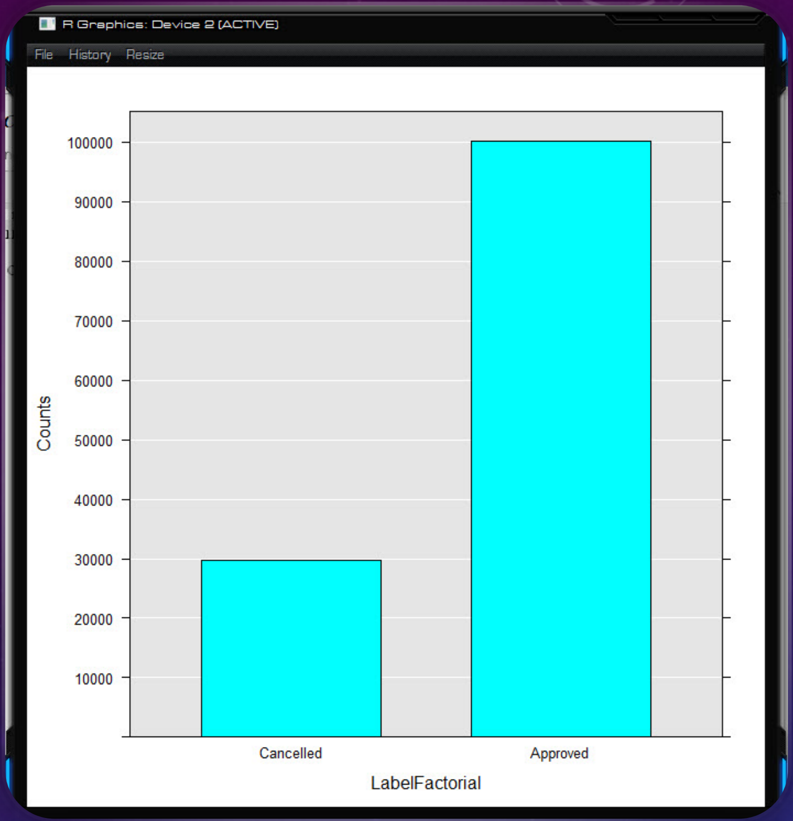
The path to save the K-Mean

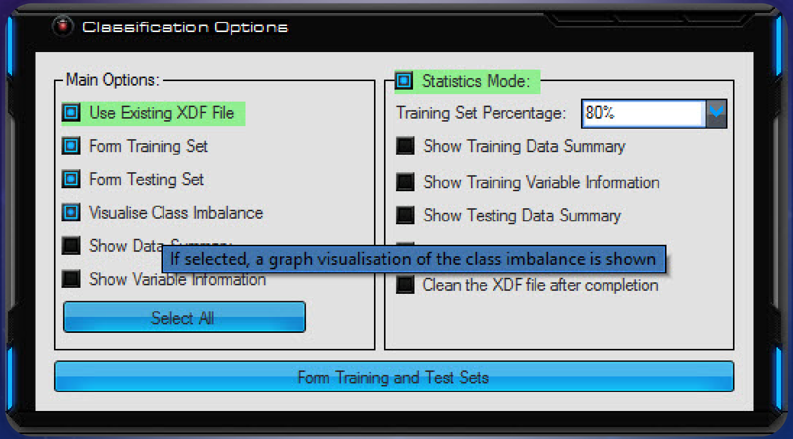
Maximum # of Clusters: 30



Ioannis Mamalikidis, UID: 633

MACHINE LEARNING [2/3]





Main Options:

☑ Use Existing XDF File
☑ Form Training Set
☑ Form Testing Set
☑ Visualise Class Imbalance
☐ Show Data
☐ Show Variable Information

☑ Statistics Mode:

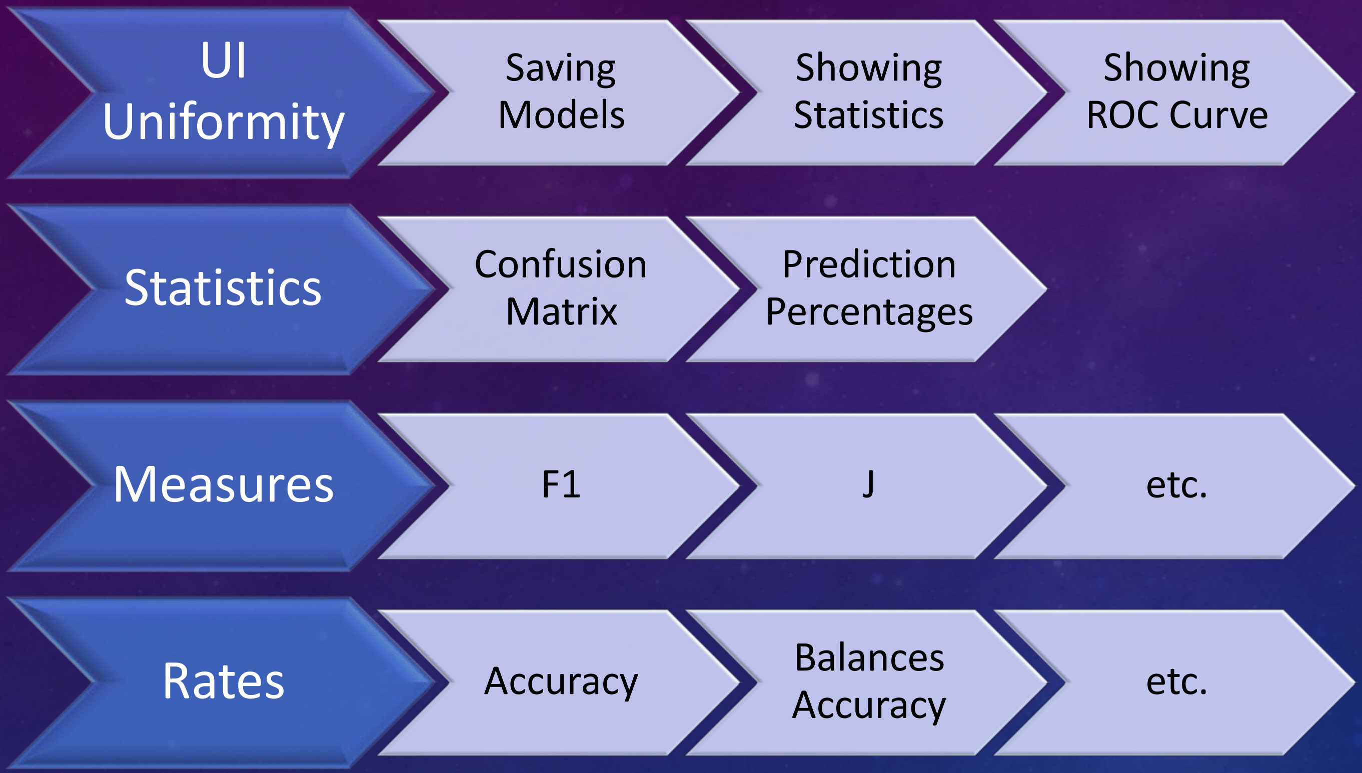
Training Set Percentage: 80%
☐ Show Training Data Summary
☐ Show Training Variable Information
☐ Show Testing Data Summary
☐ Clean the XDF file after completion

If selected, a graph visualisation of the class imbalance is shown

Statistics Mode

- Training Set Percentage
- Data Summary
- Variable Information
- Visualise Class Imbalance

MACHINE LEARNING [3/3]

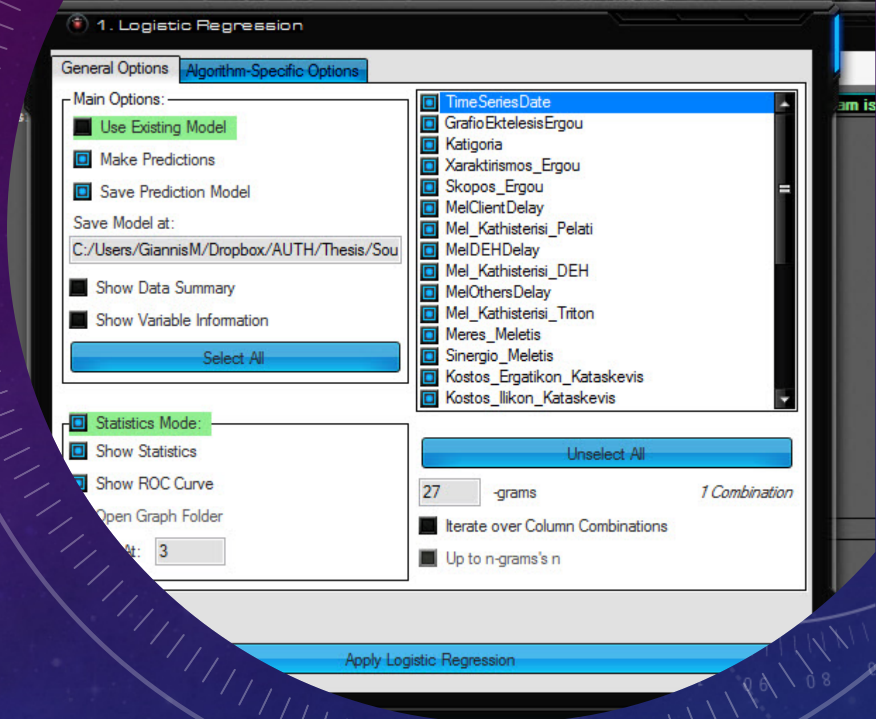


UI Uniformity → Saving Models → Showing Statistics → Showing ROC Curve

Statistics → Confusion Matrix → Prediction Percentages

Measures → F1 → J → etc.

Rates → Accuracy → Balances Accuracy → etc.



Main Options:

☐ Use Existing Model

☑ Make Predictions

☑ Save Prediction Model

Save Model at: C:/Users/GiannisM/Dropbox/AUTH/Thesis/Sou

☐ Show Data Summary

☐ Show Variable Information

☑ Statistics Mode:

☑ Show Statistics

☑ Show ROC Curve

Open Graph Folder

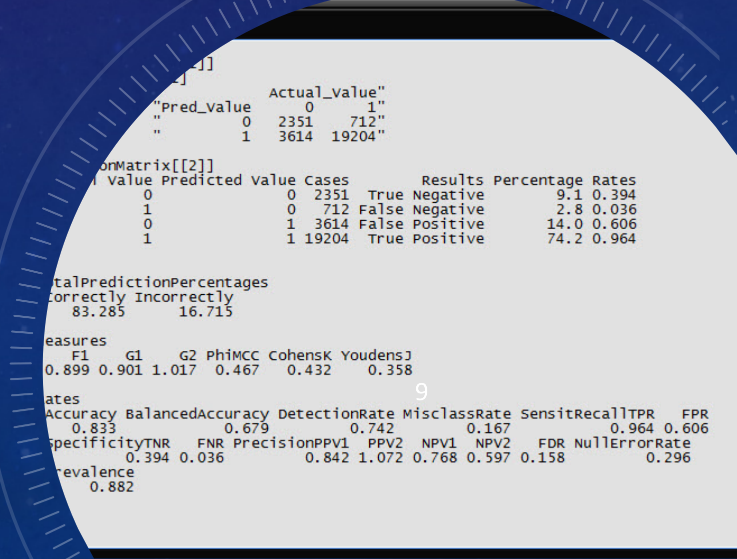
At: 3

☑ TimeSeriesDate
☑ GrafioEktelesisErgou
☑ Katigoria
☑ Xaraktirismos_Ergou
☑ Skopos_Ergou
☑ MelClientDelay
☑ Mel_Kathisterisi_Pelati
☑ MelDEHDelay
☑ Mel_Kathisterisi_DEH
☑ MelOthersDelay
☑ Mel_Kathisterisi_Triton
☑ Meres_Meletis
☑ Sinergio_Meletis
☑ Kostos_Ergatikon_Kataskevis
☑ Kostos_Ilikon_Kataskevis

27 -grams *1 Combination*

☐ Iterate over Column Combinations

☐ Up to n-grams's n

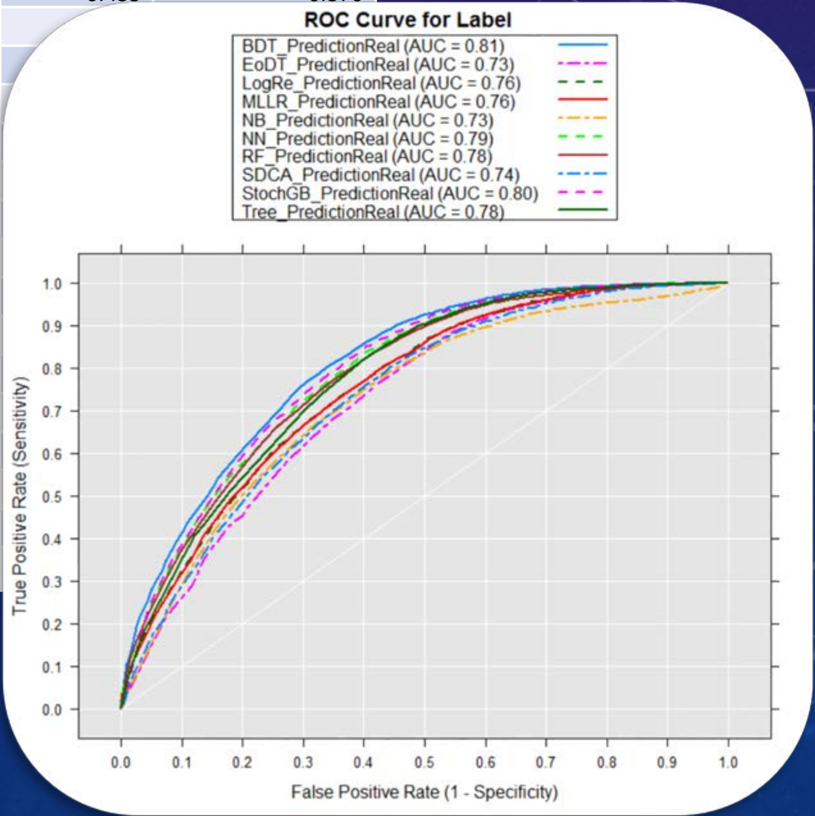


```
2]]
]
                  Actual_Value"
  "Pred_Value      0      1"
  "            0   2351    712"
  "            1   3614  19204"

onMatrix[[2]]
l Value Predicted Value Cases        Results Percentage Rates
       0               0  2351  True Negative        9.1 0.394
       1               0   712 False Negative        2.8 0.036
       0               1  3614 False Positive       14.0 0.606
       1               1 19204  True Positive       74.2 0.964

talPredictionPercentages
orrectly Incorrectly
  83.285      16.715

easures
   F1    G1    G2 PhiMCC CohensK YoudensJ
0.899 0.901 1.017  0.467   0.432    0.358

ates
Accuracy BalancedAccuracy DetectionRate MisclassRate SensitRecallTPR   FPR
   0.833            0.679         0.742        0.167           0.964 0.606
SpecificityTNR   FNR PrecisionPPV1  PPV2  NPV1  NPV2   FDR NullErrorRate
         0.394 0.036         0.842 1.072 0.768 0.597 0.158         0.296
revalence
    0.882
```

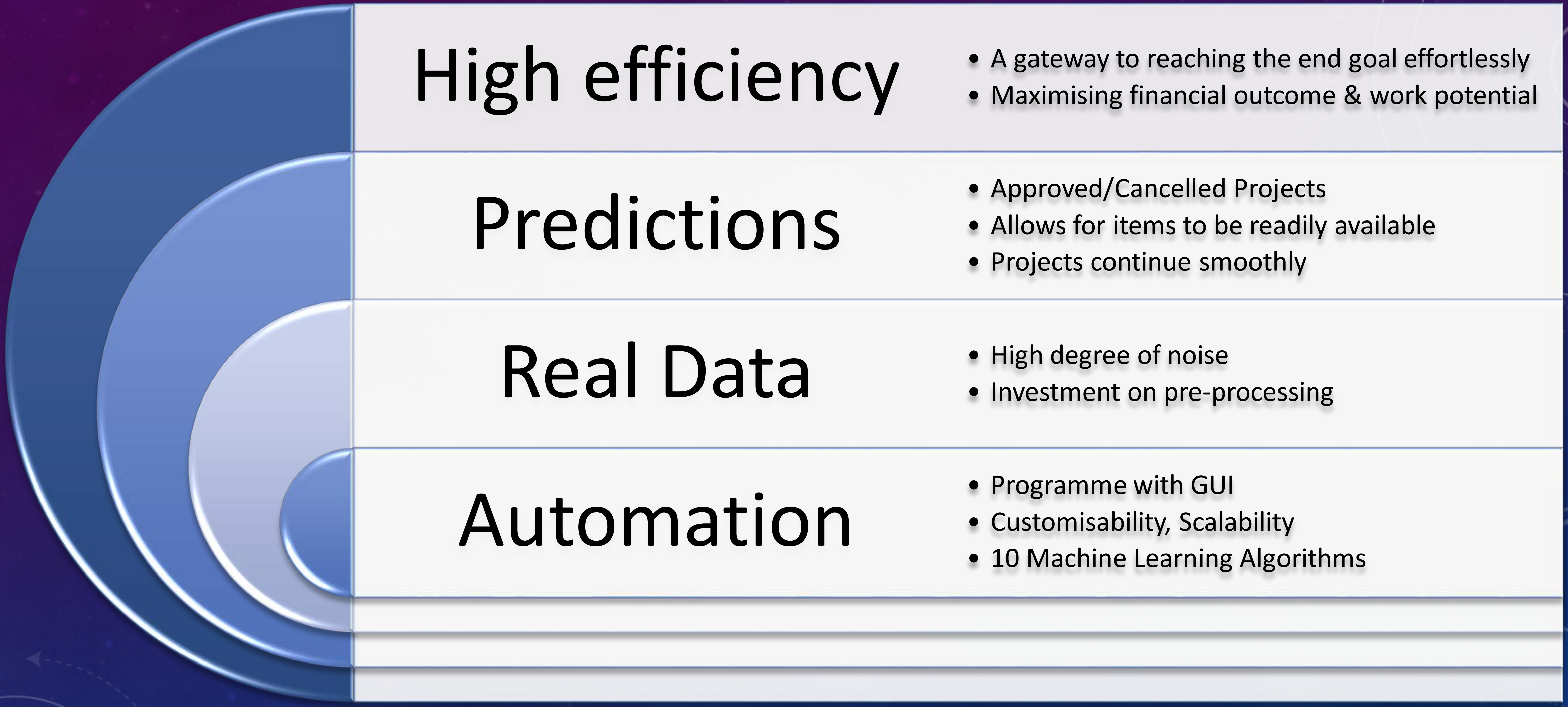
Ioannis Mamalikidis, UID: 633

MODEL EVALUATION

Model Name	Logistic Regression	Decision Trees	Naive Bayes	Random Forest	Stochastic Gradient Boosting	Stochastic Dual Coordinate Ascent	Boosted Decision Trees	Ensemble of Decision Trees	Neural Networks	Logistic Regression
Algorithm Name	rxLogit	rxDTree	rxNaiveBayes	rxDForest	rxBTrees	rxFastLinear	rxFastTrees	rxFastForest	rxNeuralNet	rxLogisticRegression
Correctly Classified	80.878%	82.635%	77.648%	81.098%	82.542%	78.072%	79.639%	80.305%	82.565%	80.932%
Incorrectly	19.122%	17.365%	22.352%	18.902%	17.458%	21.928%	20.361%	19.695%	17.435%	19.068%
AUC	0.756	0.778	0.730	0.784	0.796	0.738	0.807	0.731	0.791	0.756
F1	0.885	0.895	0.868	0.889	0.891	0.860	0.866	0.885	0.896	0.886
G	0.888	0.897	0.872	0.893	0.892	0.860	0.866	0.890	0.899	0.889
PhiMCC	0.369	0.444	0.213	0.368	0.463	0.353	0.445	0.329	0.435	0.370
CohensK	0.329	0.413	0.175	0.286	0.453	0.352	0.444	0.241		
YoudensJ	0.265	0.345	0.134	0.214	0.408	0.336	0.458	0.176		
Accuracy	0.809	0.826	0.776	0.811	0.825	0.781	0.796	0.803		
BalancedAccuracy	0.632	0.673	0.567	0.607	0.704	0.668	0.729	0.588		
DetectionRate	0.738	0.737	0.735	0.758	0.715	0.675	0.657	0.759		
MisclassRate	0.191	0.174	0.224	0.189	0.175	0.219	0.204	0.197		
SensitRecallTPR	0.960	0.958	0.956	0.985	0.929	0.877	0.854	0.987		
FPR	0.695	0.613	0.822	0.771	0.521	0.541	0.395	0.811		
SpecificityTNR	0.305	0.387	0.178	0.229	0.479	0.459	0.605	0.189		
FNR	0.040	0.042	0.044	0.015	0.071	0.123	0.146	0.013		
PrecisionPPV1	0.822	0.839	0.795	0.810	0.856	0.844	0.878	0.803		
PPV2	1.070	1.075	1.062	1.049	1.086	1.108	1.100	1.044		
NPV1	0.693	0.733	0.545	0.824	0.670	0.528	0.553	0.812		
NPV2	0.460	0.560	0.246	0.516	0.572	0.483	0.582	0.450		
FDR	0.178	0.161	0.205	0.190	0.144	0.156	0.122	0.197		



CONCLUSIONS



High efficiency

- A gateway to reaching the end goal effortlessly
- Maximising financial outcome & work potential

Predictions

- Approved/Cancelled Projects
- Allows for items to be readily available
- Projects continue smoothly

Real Data

- High degree of noise
- Investment on pre-processing

Automation

- Programme with GUI
- Customisability, Scalability
- 10 Machine Learning Algorithms

Machine learning methods for the analysis of data of an Electricity Distribution Network Operator

MASTER'S THESIS

Aristotle University of Thessaloniki, Faculty of Sciences, Department of Informatics
Supervisor: Dr. Eleftherios Angelis; Thesis Committee: Grigorios Tsoumakas, Ioannis Vlahavas

Thank You !