Business Intelligence Project

Stephane URENA

université de TOURS

Introduction

Purposes of this final report :

- Provide summaries of different steps of our BI project.
- Give our final results about the Data Mining phase (Uni-Label and Multi-Label)
- Conclusion on the whole project

TABLE OF CONTENTS

Sources

Sources

- EGC's data
  - Two datasets
  - Positions, information, defaults
- Weather, Pollution and Antenna
- Disease
- QGIS

Sources

- EGC's data
- Weather, Pollution and Antenna
  - Measures of temperature, humidity…
  - Measures of different pollutants
  - Distance between trees and Antennas
- Disease
- QGIS

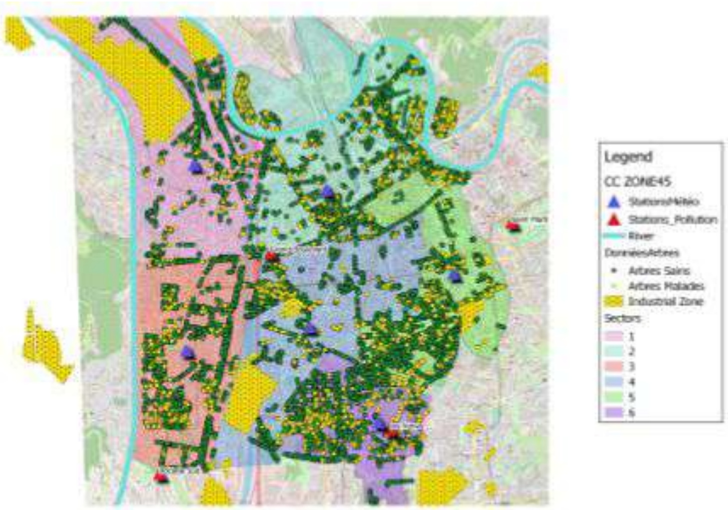
Sources

- EGC's data
- Weather, Pollution and Antenna
- Disease
  - List of disease and parasites
  - Level of weakness, for each species
- QGIS

O	P	Q	R	S	T	U	V	W	
Insectes du	Insectes xyl	Maladies à	Maladies de	Maladies de	Maladies va	Mildious	Mineuses d	Nématodes	Oï
-	-	-	-	-	-	-	-	-	-
0	0	0	0	0	0	0	0	[illegible]	
2	4	4	0	2	0	0	0	1	
0	0	0	0	0	[illegible]	0	0	[illegible]	
0	0	2	0	0	1	0	0	0	[illegible]
0	0	0	0	0	0	0	0	0	[illegible]
0	0	0	0	0	0	0	0	0	[illegible]
1	4	4	0	4	4	0	2	0	[illegible]
0	0	0	0	0	0	0	0	0	4
0	0	0	0	0	[illegible]	1	1	0	1
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1
1	2	[illegible]	0	[illegible]	[illegible]	0	4	0	[illegible]
0	0	0	0	0	0	0	0	1	
0	0	0	0	0	0	0	0	0	2
0	3	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	[illegible]	0	0	0	
0	0	0	0	0	0	0	0	0	[illegible]
0	0	[illegible]	0	0	[illegible]	0	0	1	
0	2	0	0	0	0	0	0	0	[illegible]

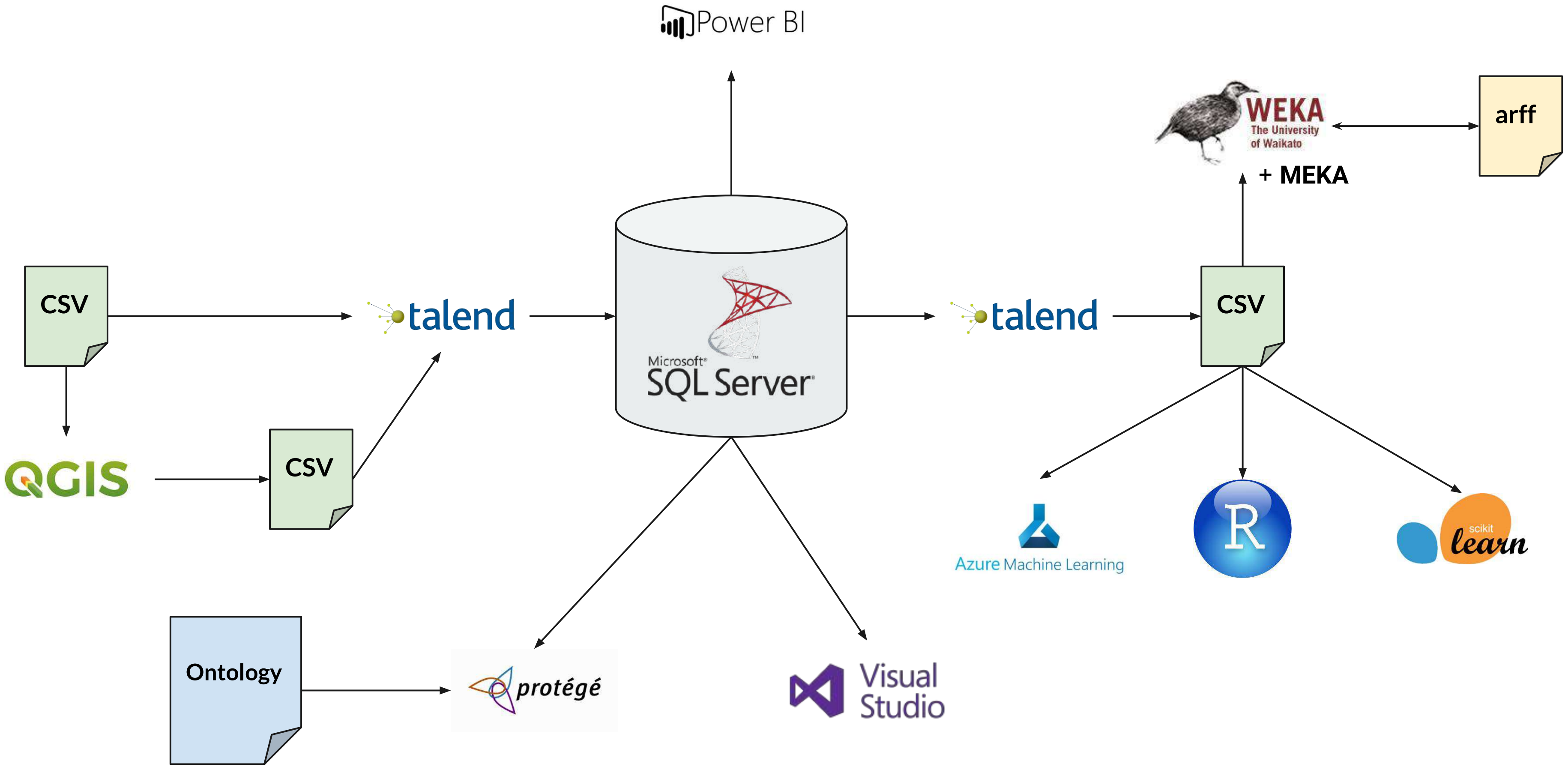
Sources

- EGC’s data
- Weather, Pollution and Antenna
- Disease
- QGIS
  - Distance river
  - Industrial zone
  - Redefine sectors



Technical choices

Technical choices



Power BI
WEKA
The University of Waikato
arff
+ MEKA
CSV
talend
Microsoft
SQL Server
talend
CSV
QGIS
CSV
Azure Machine Learning
R
scikit
learn
Ontology
protégé
Visual Studio

Data Warehouse

Workload and Modeling

Example of query from WL:

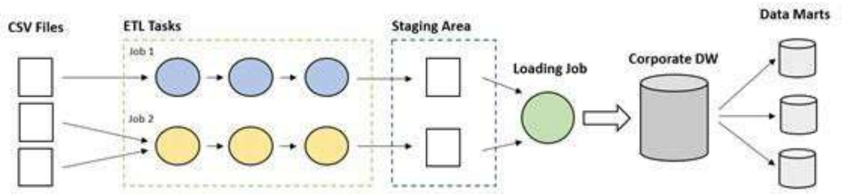
- In natural language query:
  Number of trees according to their sector, genus and development stage
- Formal language query :
  treeCaracteristic[genus,sector,development_stage].number_of_trees

Modelling approach:

- DF Model
- Star Schemas
- Implementation

Extract-Transform-Load

- Talend
- Creation of new tables and constraints
- Used conversion and normalisation function



Databases and Cubes

3 schemas :

- Tree Characteristics (24 Tables)
- Diagnosis (3 Tables)
- Environment (5 Tables)

→ 5 versions (the last one was for the Analysis Phase)

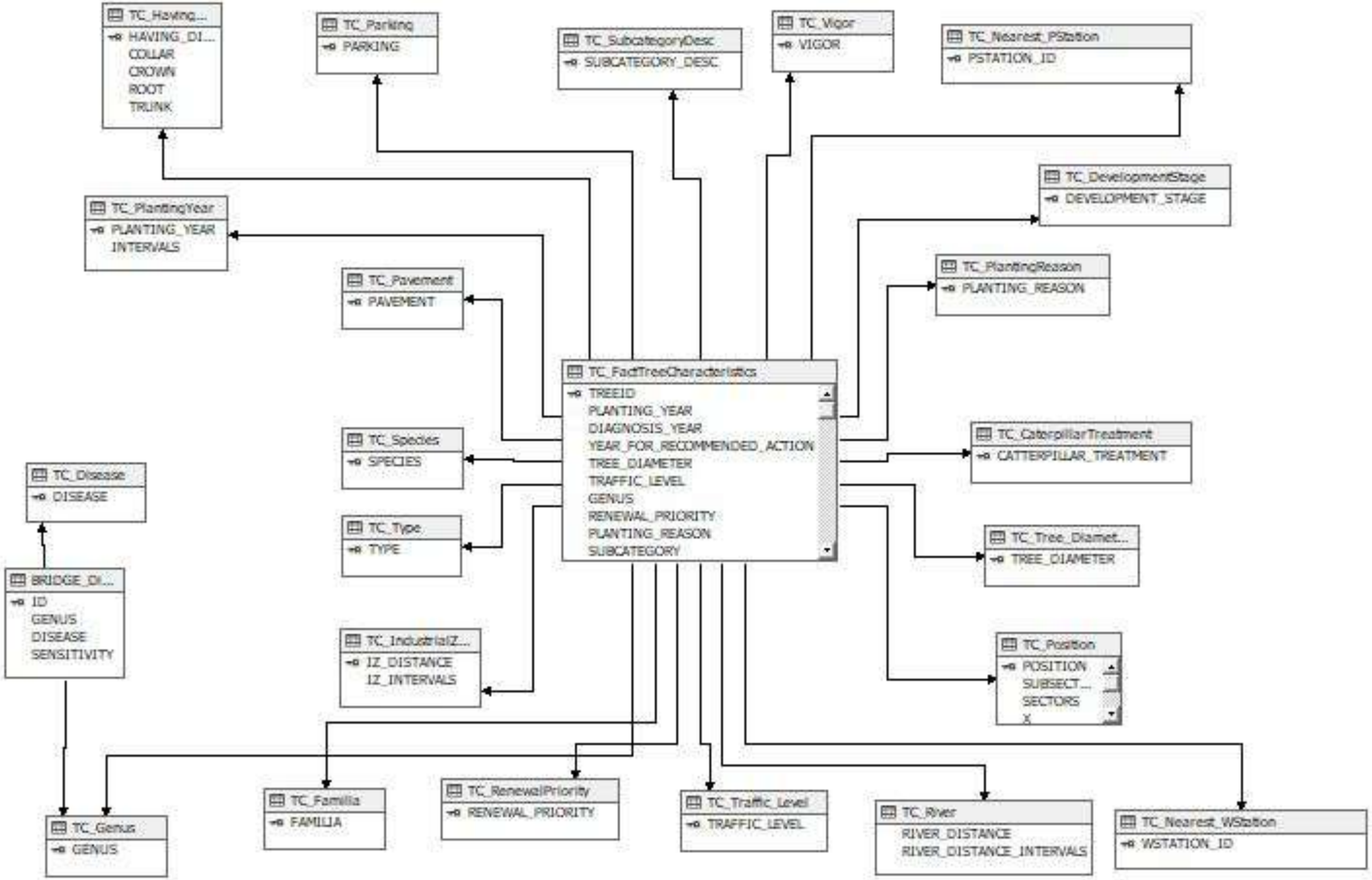
Final :

32 Tables

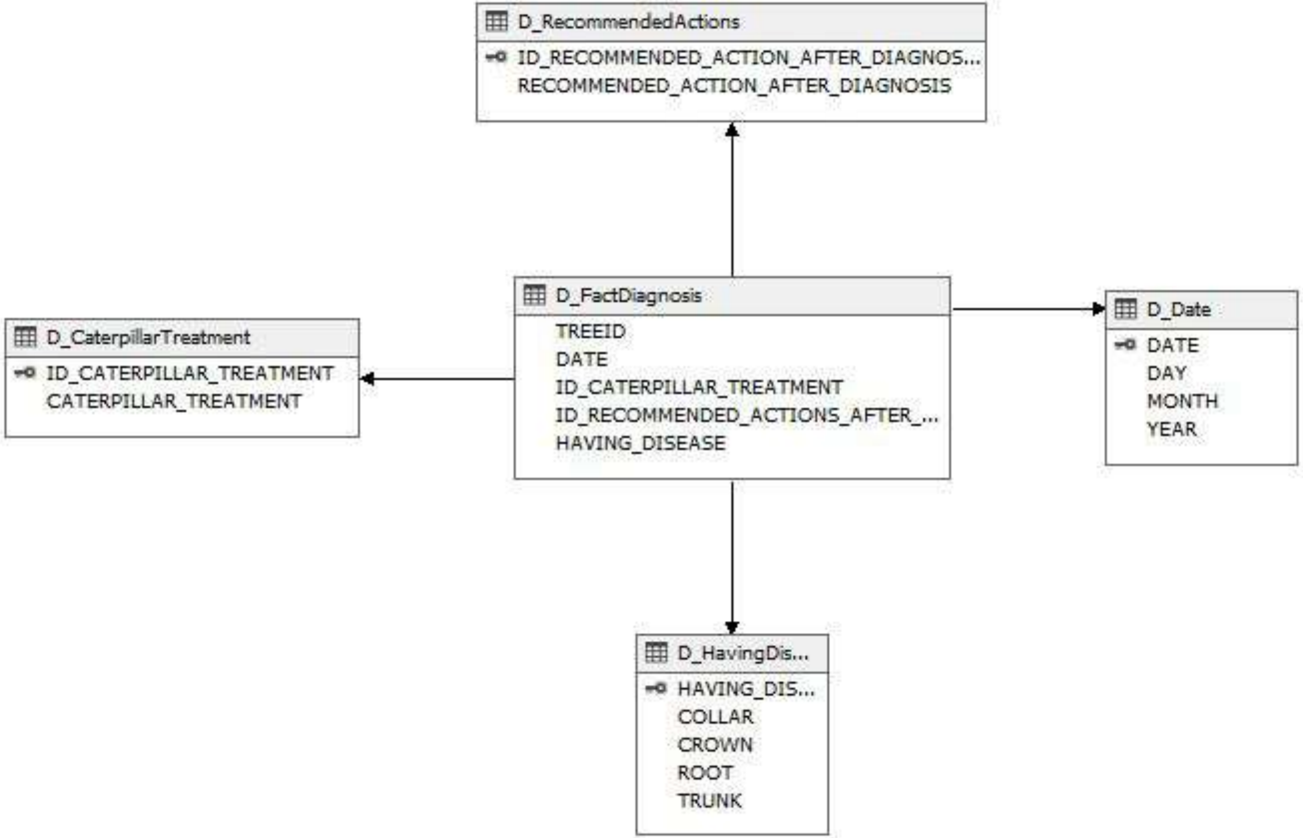
7.9 GB of Data

Cubes : OLAP, MOLAP, Hybrid

Tree Characteristics

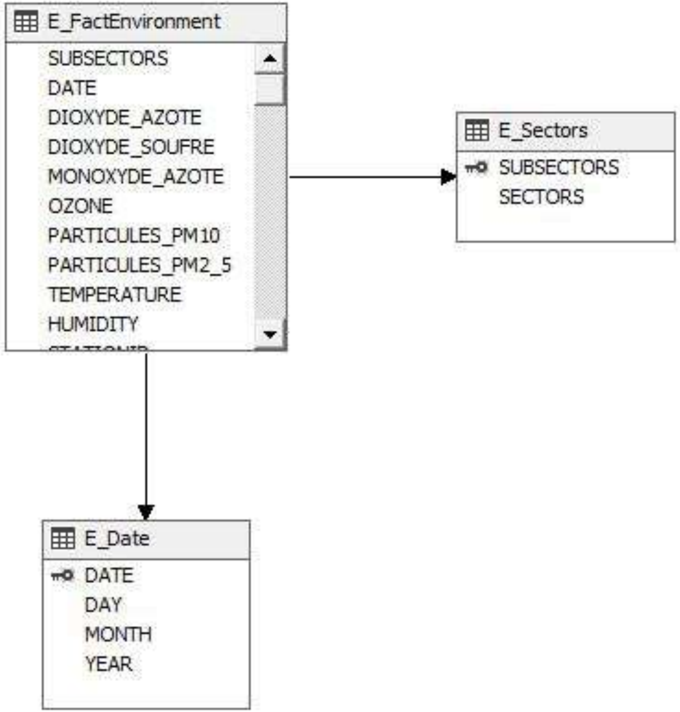


Diagnosis



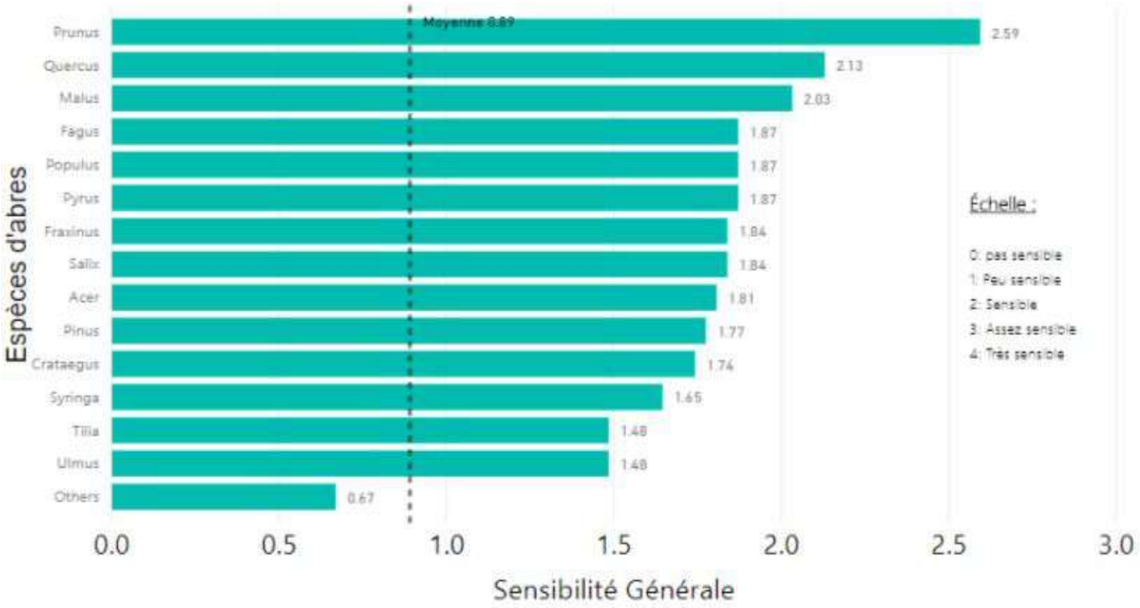
D_RecommendedActions
ID_RECOMMENDED_ACTION_AFTER_DIAGNOS...
RECOMMENDED_ACTION_AFTER_DIAGNOSIS
D_FactDiagnosis
TREEID
DATE
ID_CATERPILLAR_TREATMENT
ID_RECOMMENDED_ACTIONS_AFTER_...
HAVING_DISEASE
D_CaterpillarTreatment
ID_CATERPILLAR_TREATMENT
CATERPILLAR_TREATMENT
D_Date
DATE
DAY
MONTH
YEAR
D_HavingDis...
HAVING_DIS...
COLLAR
CROWN
ROOT
TRUNK

Environment

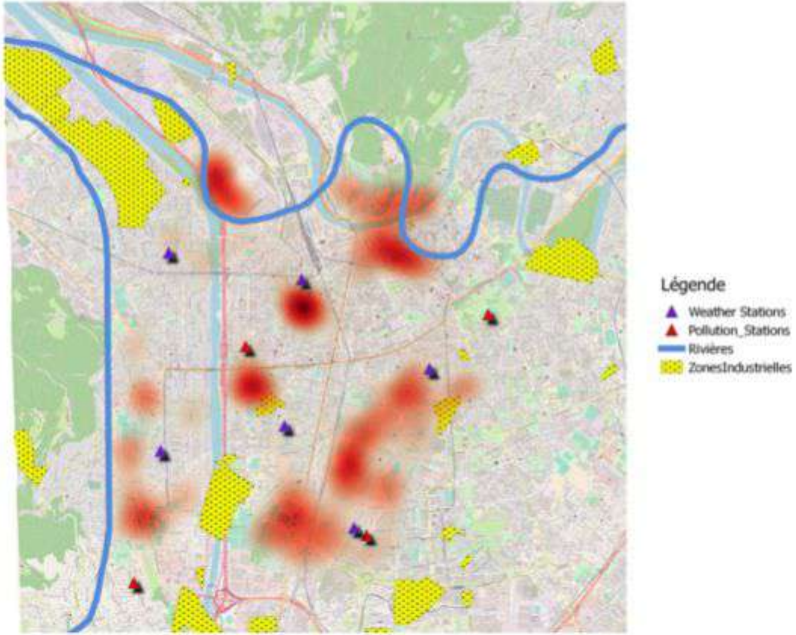


Data Visualisation

The most sensitive species



The distribution of diseased trees in Grenoble



Typical profiles of sick trees

- Prunus trees, Quercus, Malus, etc…
- Diameter > 70 centimeters.
- Age > 6 years
- Trees near rivers.
- Non-vigorous trees.
- Mostly : crown and/or trunk infected
- On red zones of the HeatMap.

Ontology

Taxonomy

- Stanford Protege 5.0
- Built an ontology about the taxonomy of trees
- Used data from DBpedia
- Linked species and genus to existing data on the Web

TABLE 1 Linnaean Hierarchical System

Kingdom	Plantae
Phylum	Anthophyta
Class	Dicotyledonae
Order	Asterales
Family	Asteraceae
Genus	*Aster*
Species	*spectabilis*
common name	showy aster
scientific name	*Aster spectabilis*

On top: Connection to the DB, Mappings and SPARQL

ontop query editor:

Query Editor

```
PREFIX ns5: <http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#>
PREFIX dbpedia-pl: <http://pl.dbpedia.org/resource/>
PREFIX dbc: <http://dbpedia.org/resource/Category:>
PREFIX ns7: <http://mappings.dbpedia.org/index.php/OntologyClass:>
PREFIX dbpedia-nl: <http://nl.dbpedia.org/resource/>
PREFIX yago: <http://dbpedia.org/class/yago/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX obda: <https://w3id.org/obda/vocabulary#>
PREFIX wikidata: <http://www.wikidata.org/entity/>
PREFIX dbpedia-eu: <http://eu.dbpedia.org/resource/>
PREFIX dbp: <http://dbpedia.org/property/>
PREFIX umbel-rc: <http://umbel.org/umbel/rc/>
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX dbpedia-es: <http://es.dbpedia.org/resource/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX yago-res: <http://yago-knowledge.org/resource/>
PREFIX dbpedia-wikidata: <http://wikidata.dbpedia.org/resource/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ns23: <http://purl.org/linguistics/gold/>
PREFIX xml: <http://www.w3.org/XML/1998/namespace>
PREFIX ns0: <http://open.vocab.org/terms/>
PREFIX ns2: <http://open.vocab.org/terms/>
SELECT ?x ?d ?v ?p ?t
WHERE {?x a lgdo:Tree ; :diameter ?d ; :vigor ?v ; :PlantingYear ?p ; :TrafficLevel ?t .}
```

Execution time: 0.187 sec - Number of rows retrieved: 100

Show: 100 All Short IRI Attach Prefixes Execute Save Chan

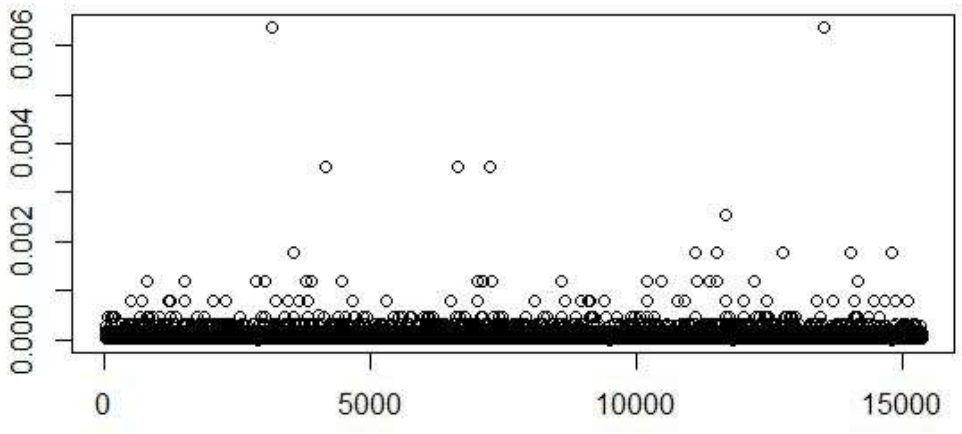
x	d	v	p	t
ESP10000	"40-50"^^string	"vigoureux"^^string	2004	"passages fréquents et arrêts fréquents"^^string
ESP10002	"20-30"^^string	"vigoureux"^^string	2004	"passages fréquents et arrêts fréquents"^^string
ESP10001	"10-20"^^string	"vigoureux"^^string	2004	"passages fréquents et arrêts fréquents"^^string
ESP10004	"10-20"^^string	"vigoureux"^^string	2004	"passages fréquents et arrêts fréquents"^^string
ESP10008	"20-30"^^string	"vigoureux"^^string	2004	"passages fréquents et arrêts fréquents"^^string
ESP10003	"30-40"^^string	"vigoureux"^^string	2004	"passages fréquents et arrêts fréquents"^^string
ESP10012	"30-40"^^string	"vigoureux"^^string	2004	"passages fréquents et arrêts fréquents"^^string
ESP10005	"50-60"^^string	"vigoureux"^^string	2004	"passages fréquents et arrêts fréquents"^^string
ESP10015	"20-30"^^string	"vigoureux"^^string	2004	"passages fréquents et arrêts fréquents"^^string
ESP10014	"30-40"^^string	"vigoureux"^^string	2004	"passages fréquents et arrêts fréquents"^^string
ESP10018	"50-60"^^string	"vigoureux"^^string	2004	"passages fréquents et arrêts fréquents"^^string
ESP10010	"20-30"^^string	"vigoureux"^^string	2004	"passages fréquents et arrêts fréquents"^^string
ESP10033	"10-20"^^string	"vigoureux"^^string	2004	"passages fréquents et arrêts fréquents"^^string
ESP10026	"10-20"^^string	"vigoureux"^^string	2004	"passages fréquents et arrêts fréquents"^^string

Data Analysis

Data Processing

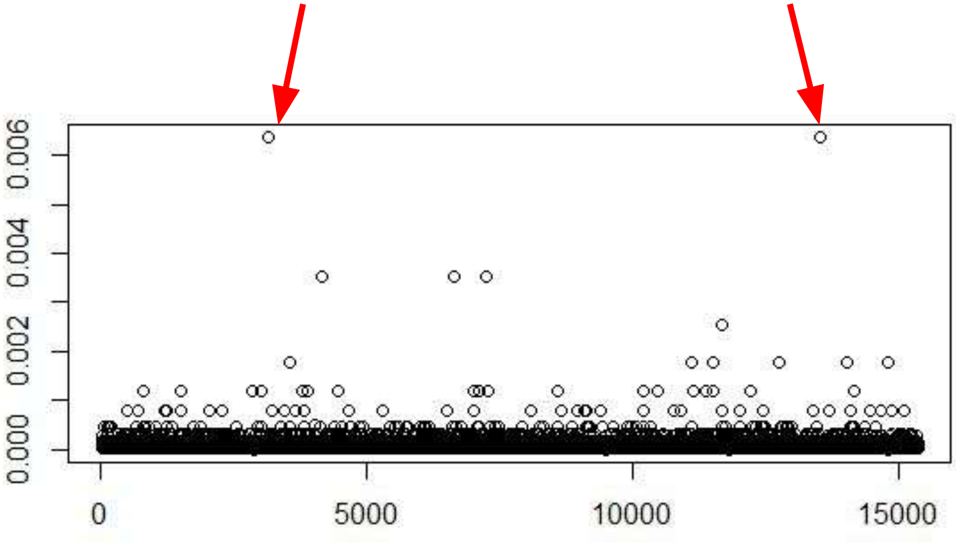
- Outliers
- Treatment of missing values
- Discretization

Outliers



Cook Distance between TREE_DIAMETER et DEFAULT_OR_NOT (with R Studio)

Outliers



Cook Distance between TREE_DIAMETER et DEFAULT_OR_NOT (with R Studio)

Missing Values

Attributes	Before Treatment	After Treatment
CATERPILLAR_TREATMENT	14 287 (93%)	0
DEVELOPMENT_STAGE	51 (0%)	51 (0%)
DEVELOPMENT_STAGE_AT_DIAG	13 (0%)	13 (0%)
DIAGNOSIS YEAR	8 (0%)	8 (0%)
TREE_DIAMETER	67 (0%)	0
FAMILIA	84 (1%)	0
NOTES	9 381 (61%)	0
PLANTING_REASON	15 145 (99%)	15 145 (99%)
RECOMMENDED_ACTIONS_AFTER_DIAG	4 525 (29%)	4 525 (29%)
RENEWAL_PRIORITY	127 (1%)	127 (1%)
SPECIES	1 018 (7%)	1 018 (7%)

Attributes	Before Treatment	After Treatment
TRAFFIC_LEVEL	1 (0%)	0
TYPE	84 (1%)	1 (0%)
VARIETY	13 212 (86%)	13 212 (86%)
VIGOR	11 (0%)	11 (0%)
YEAR_FOR_RECOMMENDED_ACTIONS	4 511 (29%)	4 511 (29%)
Toutes les maladies	236 (2%)	138 (1%)

Discretization

Discretization of some attributes :

- River Distance
- IZ Distance
- Planting Year
- ...

Tests : arbitrary choices and jenks method

Understanding of the data

- Nature of attributes
- Distribution analysis
- Univariate analysis
- Bivariate analysis

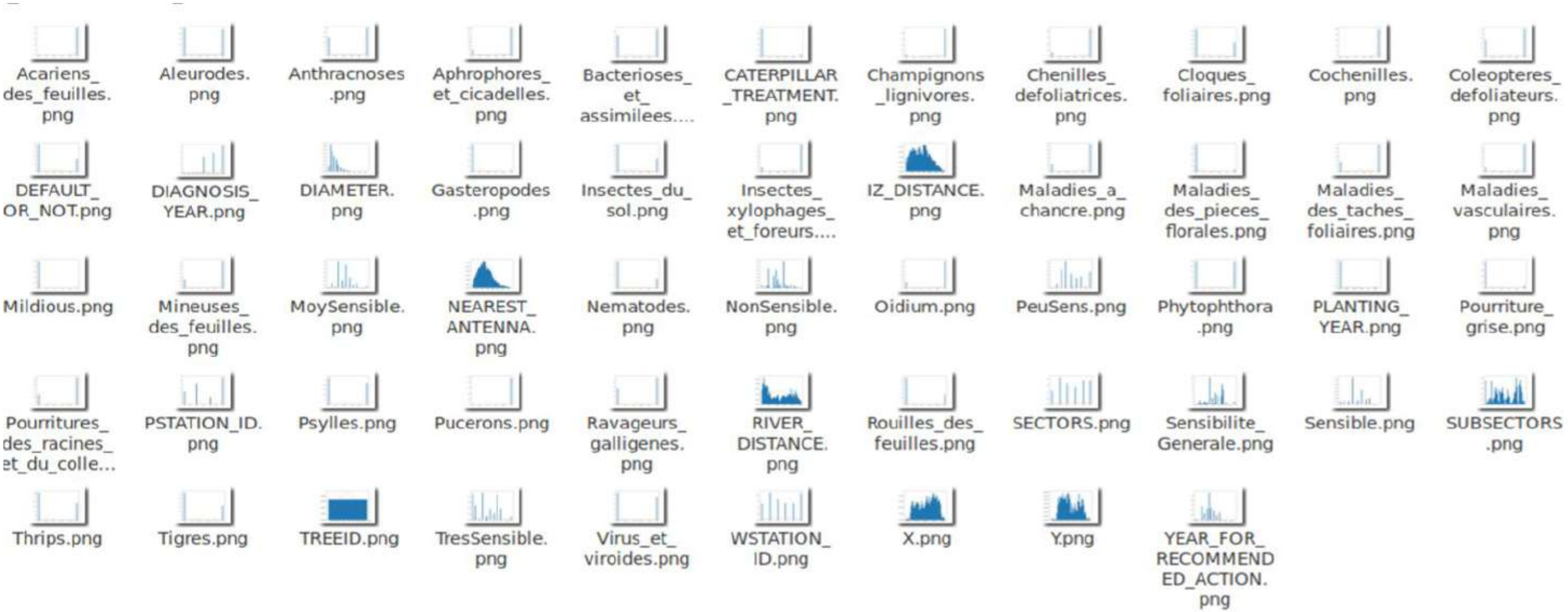
Attributes

Nature of Attributes

- Class attributes
  - Unilabel : Default or not
  - Multilabel : crown, collar, trunk, root
- Descriptive attributes
  - 70 attributes
  - 53 numeric
- Basically, we try to explain the value of the the class attribute with the descriptive attributes

Distribution Analysis

- It is important to determine the distribution law for an attribute
  - Normal, geometric...
- To avoid to having false results with wrong methods



Distribution analysis

Univariate analysis

Purpose :

Study each attributes one by one

Must be able to :

- the possible value field
- the (relative) strength
- identify the null or outliers

Symbolic attributes: numbers, missing values and their meanings

Numeric attributes: mean, the standard deviation and the law that follow the values

Result :

Null values and their interpretation

A majority of features do not follow a normal law

Bivariate analysis

Purpose : Study pairs of attributes and see if there is a correlation between them, study an attribute against the class.
Goal : Remove redundant values and eliminate values that have a low gain

Two type of data :

- Numerical
- Symbolic

Three analysis :

- Correlation
- Chi-squared
- ANOVA/Kruskal-Wallis

Results :

- Matrix of correlations between attributes
- Matrix of p-value resulting from Chi-squared test
- Ranking of attributes according to ANOVA and Kruskal Wallis test

Data Mining

Unilabel Classification

- State of the art
- Tests
- Features selections
- Results

States of the art

The main unilabel algorithms:

- Neural networks
- Two-class averaged perceptron
- Boosted decision trees
- Random Forest
- Bayesian classification
- Locally-deep SVM
- Logistic regression
- SVM
- Decision jungle

Decision jungles (did not see in class)

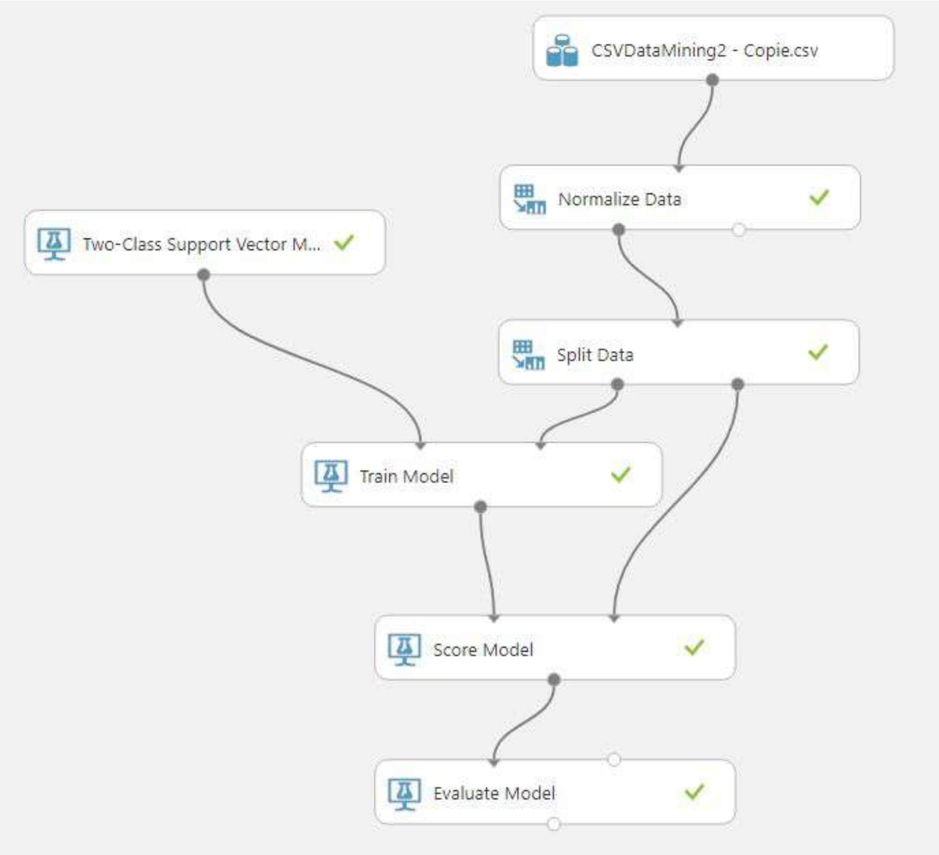
Extension to decision forests

Set of decision directed acyclic graphs :

- lower memory footprint and better generalization performance
- non-parametric models
- represent non-linear decision boundaries
- resilient in the presence of noisy features

Tests

- **Technologies**
    - **R Studio**
    - **Weka**
    - **Python**
    - **Microsoft Azure ML Studio**
- **Predicted attribute :** Default or not
- **Apply the different algorithms**
- **Evaluate**

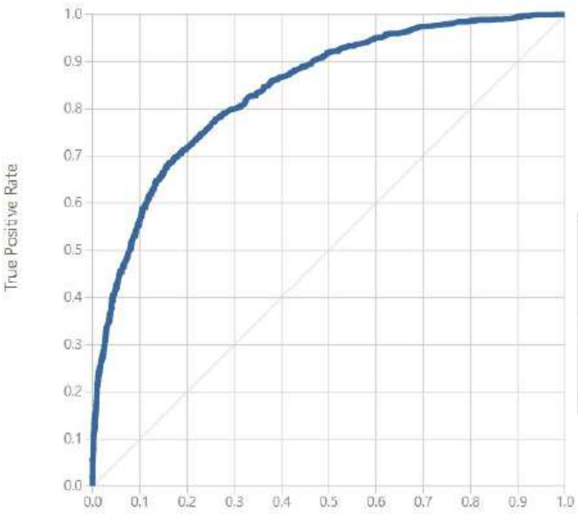


Main Steps

- Transform nominal attributes into numeric attributes
- Split Data
- Create Train and Test datasets
- Apply the algorithm on the training set
- Test the model
- Evaluation

Example of one experiment with Azure Machine learning studio

ROC PRECISION/RECALL LIFT




```
Accuracy ::  0.8275684047496128
Recall :: 0.6901004304160688
Precision :: 0.8030050083472454
```

```
Thresh=0.012, n=28, Accuracy: 81.98%
Thresh=0.013, n=27, Accuracy: 81.93%
Thresh=0.014, n=26, Accuracy: 83.22%
Thresh=0.014, n=25, Accuracy: 82.71%
Thresh=0.014, n=24, Accuracy: 82.65%
Thresh=0.020, n=23, Accuracy: 81.83%
```

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
873	623	0.794	0.728	0.5	0.840
False Positive	**True Negative**	**Recall**	**F1 Score**		
326	2790	0.584	0.648		

Definition

Accuracy :

$$\frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Recall :

$$\frac{\text{Number of True Positive}}{\text{Number of True Positive} + \text{Number of False Positive}}$$

Precision :

$$\frac{\text{Number of True Positive}}{\text{Number of True Positive} + \text{Number of False Negative}}$$

Evaluation

Features Selection

- Weka

Methods used :

- **Information Gain** on DEFAULT_OR_NOT

Evaluates the worth of an attribute by measuring the information gain with respect to the class.

InfoGain(Class,Attribute) = H(Class) - H(Class | Attribute).

Output : 20 Attributes

Top 5 :
RENEWAL_PRIORITY,
VIGOR,
DEVELOPMENT_STAGE_AT_DIAG,
DEVELOPMENT_STAGE,
PLANTING_YEAR_INTERVALS

Results (1)

- Python & ML studio

	Accuracy	Precision	Recall
Random Forest	0.8404	0.81	0.71
Averaged perceptron	0.814	0.751	0.638
Boosted Decision Trees	0.842	0.833	0.642
Bayes Point	0.802	0.744	0.596
Decision Jungle	0.789	0.746	0.527
Locally-Deep SVM	0.825	0.777	0.646
Logistic regression	0.812	0.747	0.634
SVM	0.794	0.724	0.591

Results (2)

- Weka

	Accuracy	Precision	Recall
K-Nearest Neighbors (k=5)	0.797	0.71	0.797
J48	0.815	0.811	0.815
Random Forest	0.823	0.814	0.798

Multilabel Classification

- State of the art
- Tests
- Features selections
- Results

States of the art

Transform the problem by simplifying it

Develop methods that adapt the uni-label algorithms

States of the art

Transform the problem by simplifying it

"One VS all" algorithm

X	Y_1
$x^{(1)}$	0
$x^{(2)}$	1
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	0

X	Y_1	Y_2
$x^{(1)}$	0	1
$x^{(2)}$	1	0
$x^{(3)}$	0	1
$x^{(4)}$	1	0
$x^{(5)}$	0	0

X	Y_1	Y_2	Y_3
$x^{(1)}$	0	1	1
$x^{(2)}$	1	0	0
$x^{(3)}$	0	1	0
$x^{(4)}$	1	0	0
$x^{(5)}$	0	0	0

X	Y_1	Y_3	Y_3	Y_4
$x^{(1)}$	0	1	1	0
$x^{(2)}$	1	0	0	0
$x^{(3)}$	0	1	0	0
$x^{(4)}$	1	0	0	1
$x^{(5)}$	0	0	0	1

Label PowerSet Methods

X	$Y \in 2^L$
$x^{(1)}$	0110
$x^{(2)}$	1000
$x^{(3)}$	0110
$x^{(4)}$	1001
$x^{(5)}$	0001

States of the art

Develop methods that adapt the uni-label algorithms

- ML-kNearest Neighbors
- Multi-label Decision Trees
- Rank SVM
- Neural networks

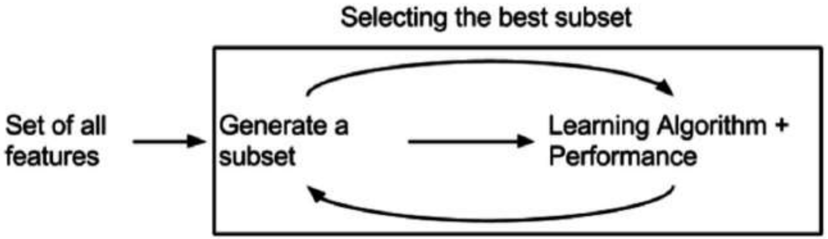
Tests

- Technologies
  - Meka
  - Python
- Predicted attributes
  - Collar, Crown, Root, Trunk
  - Default or not
- Same Steps
- Evaluation

Features Selection

Same goal as unilabel feature selection

- Focuses on each label
- Computed the best subset of attributes



Methods used :

- Feature Importance of Tree-based methods
- Univariate Feature Selection
- Variance Threshold

Output :
27 attributes

Top 5 :
**PLANTING_YEAR,
PLANTING_YEAR_INTERVALS,
DIAGNOSIS_YEAR,
TREE_DIAMETER,
TRAFFIC_LEVEL.**

Results

- Python

Random Forest : Accuracy : 0.731

Python	Micro	Macro
Precision	0.7403 *(0.70)*	0.673 *(0,64)*
Recall	0.502 *(0,47)*	0.381 *(0,37)*

Kneighbors Classifier: Accuracy : 0.69

Python	Micro	Macro
Precision	0.61 *(0.70)*	0.58 *(0,64)*
Recall	0.41 *(0,47)*	0.311 *(0,37)*

Results

- **Meka**

BR method : Accuracy : 0.741

Meka	Micro	Macro
Precision	**0.733** *(0.70)*	0.565 *(0,64)*
Recall	0.463 *(0,47)*	0.312 *(0,37)*

Label Powerset Method : Accuracy : 0.763

Meka	Micro	Macro
Precision	**0.712** *(0.70)*	0.602 *(0.64)*
Recall	0.422 *(0.47)*	0.33 *(0.37)*

Results

Chained Classifier in a trellis structure : Accuracy : 0.764

Meka	Micro	Macro
Precision	**0.712** *(0.70)*	**0.651** *(0.64)*
Recall	**0.481** *(0.47)*	0.353 *(0.37)*

Conclusion

Conclusion

Results :

- We reached almost all baselines given by the EGC Challenge.
- Models and methods seem to be quite good.

General Conclusion :

- Opportunity to work on a real dataset, a real challenge
- A way to put into practice every subjects learned during this school year
- A good experience with some good and bad points.

Good points :

- A good relation between members
- Kept in touch everyday, even if it was not about the project (we went out together ,etc…
- Good technological choices (at least one member used to work with them)
- Weekly meeting (from 3 to 6 hours)
- A good division of work
- Motivation
- No delay in rendering

Bad points :

- Many changes in the group (3 in only 3 months)
- Settings of the VM at the beginning (organisation : shared folders, disk…)
- Settings of some softwares (workspace, users,...)
- Many changes in the data warehouse and some updates were a bit long
- A better DW schema
- Keep all the attributes of the EGC Challenge at the beginning

Thanks for your attention !