REPORT

TOPIC : PROJECT PRODUCTION AND BILLING TIMESHEET TRACKING

Elaborated by :

Amira DOGHRI

Eya AKRIMI

Ines ZHIOUA

Raoua LAHOUAGUE

Maroua ABBOUR

Nour TRABELSI

4 ERP-BI5

2020-2021

THE HOUSE OF DIGITAL EXPERTS

List of contents

List of figures

List of tables

General Introduction

Business intelligence is a subject in full Evolution, addressing the general direction as the trades. it is the decision support tool that enables a whole different business activities and its view .This view environment requires knowledge of the different Trades of the company and implies some organizational specifications . The implementation of projects BI can not be done without having defined a comprehensive BI strategy. So we can say that the Business Intelligence is the process of distill

So we can say that the Business Intelligence is the process of distilling information and knowledge from data. With this, companies try to find business insights that can help in running or improving the company. Today, the BI market offers relatively comprehensive solutions through analysis concerning aspects of reporting and data consolidation. The key to successful analytics is gaining timely insights from data that leads to better decision making.

g.It resolve their business challenges by turning high volumes of data into actionable insights that can enable them . to advance their business objectives, achieve their revenue goals, and sustain a competitive advantage. The business insights can be disclosed via different data presentations like reports, dashboards and visualizations, or via analysis tools like OLAP or advanced statistical learning tools.

Usually there are three different ways to incorporate the business insights in the business:

- Managed reports that are periodically refreshed.
- Self-service analytics.
- Input for operational systems .

Business intelligence also includes all activities to gather, prepare or decide on the necessary data and data definitions. Examples of these activities are:

- Data warehouse modeling.
- Data integration and ETL (extraction, transformation, load).
- Data governance for data definitions and data ownership.
- Validating data quality.

Chapter 1: Project's Context

1. Introduction

In this chapter, we are going to present the context of our project, starting with study of existing, then specifying the objectives of our project and finally presenting the different solutions.

2. The host organization

TALYS is a group of technological companies specializing in Organization, **Information Systems** and **Digital Transformation**.

Since 2006, its two founders **Hatem MSADAA** and **Elyssa AOUNALLAH** have had the desire to **create added value for customers** by supporting them in **the realization and success of their** innovative **projects** and the transformation of their business.

Today, TALYS is positioned as a major player in the financial **sector** in Tunisia and has succeeded in extending its **scope of action on 3 continents (Africa, Europe and the Middle East).** This is thanks to a team of **functional and technical experts** who combine several business expertise: **Banking**, insurance, **microfinance** and **leasing** .

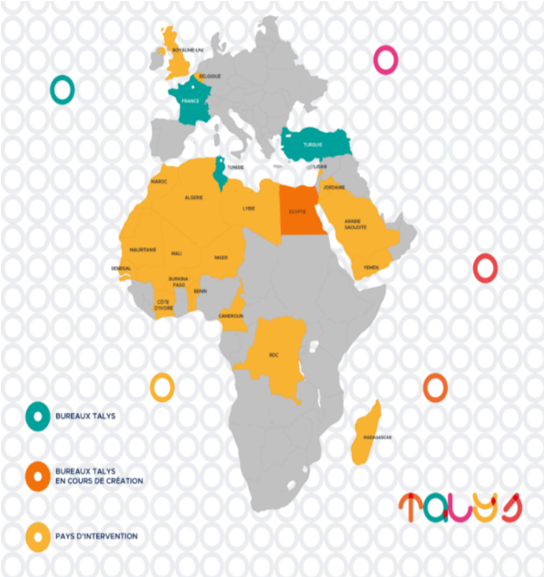


Figure 1 *:* position of TALYS

Figure 2: TALYS in numbers

3. The used application

In a decidedly digital and connected world, TALYS manages their projects through the project management tool ProjeQtOr.

Quality based Open-Source Project Organizer.

Complete, Collaborative, Quality based Open-Source Project Organizer.

ProjeQtOr (formerly Project'Or RIA) is a collaborative project management software. It is a tool designed to be a Project Organizer as a Rich Internet Application. Web based, it is very easy to use and targets to include every feature needed to the management of your projects.

Figure *3:* Features of PROJEQTOR

4. Problematic

The application limits

This project management tool is limited in the reporting and visualization area:

TALYS employees manage their **consulting** projects through ProjeQtor, in order to give a better-quality service, they need more visualizations on:

- Projects' deadlines: phases, delays, and timing
- Projects' and tasks assignment
- Teams/Consultants assigning to projects.
- Time/charge tracking and management.
- Billing progress
- Projects' activity sector

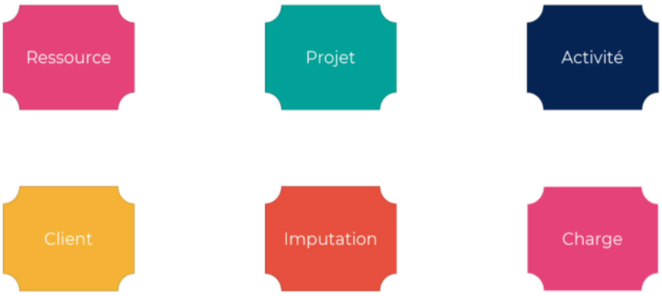


Figure 4: project keyword

Challenges

There could be scenarios where the information collection frequency and the quality of the collected data are not aligned with the Business Intelligence project requirements. This generally results in a lack of trust of the decision makers with the information obtained from the reports.

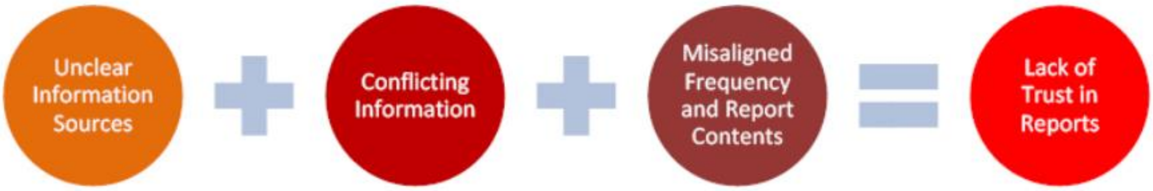

Unclear Information Sources
Conflicting Information
Misaligned Frequency and Report Contents
Lack of Trust in Reports

Figure 5: Challenges

In order to ensure the project's reliability, there are some crucial steps to do:

- Sources of information are identified and verified.
- Reporting mechanism is transparent, including the clear process through which the reports are created.
- Source used is accurate and relevant to the information required.
- Information is sufficiently specific and updated.

Business Objectives & Functional requirements

5. Business Objectives

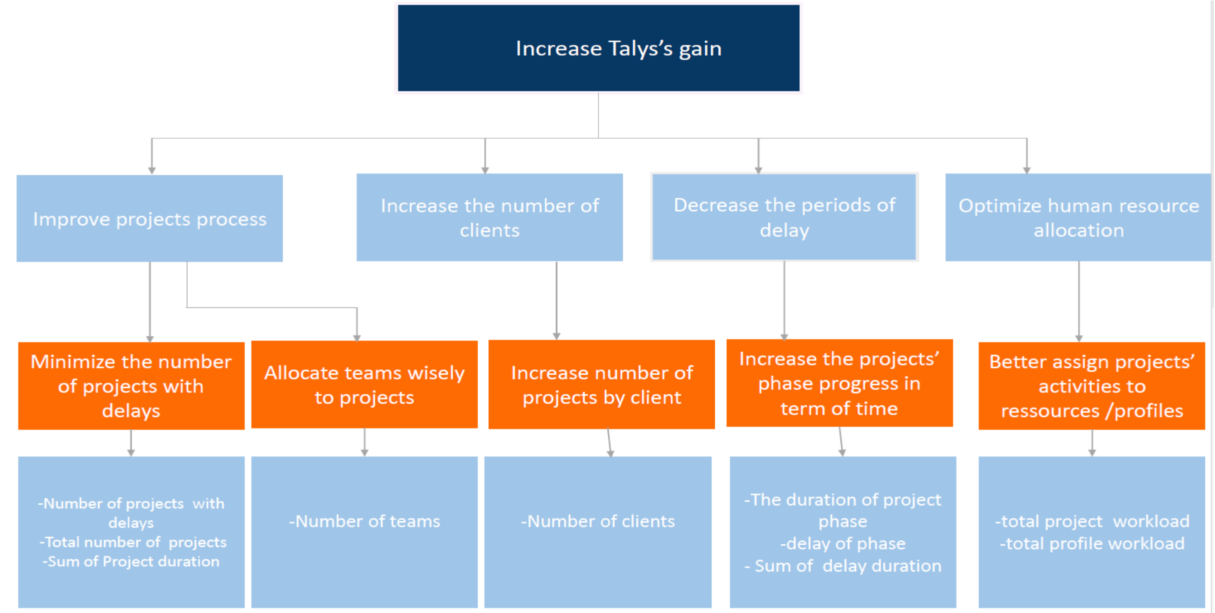
Business goals and objectives are part of the planning process. They describe what a company expects to accomplish throughout the year.

These goals and objectives might pertain to the company, departments, employees, customers and even marketing efforts.

For each business objective, functional requirements are derived in order to be deployed in the project.

The functional requirements come from the specifications of the project. These are the required needs by the end user. They are the features and actions that the dashboard must obligatorily carry out. The final dashboard must meet the following functional requirements:

The project mainly aims to help the company **improve its performance and lead a better strategy.**



Increase Talys's gain

Improve projects process

Increase the number of clients

Decrease the periods of delay

Optimize human resource allocation

Minimize the number of projects with delays

Allocate teams wisely to projects

Increase number of projects by client

Increase the projects' phase progress in term of time

Better assign projects' activities to ressources /profiles

-Number of projects with delays
-Total number of projects
-Sum of Project duration

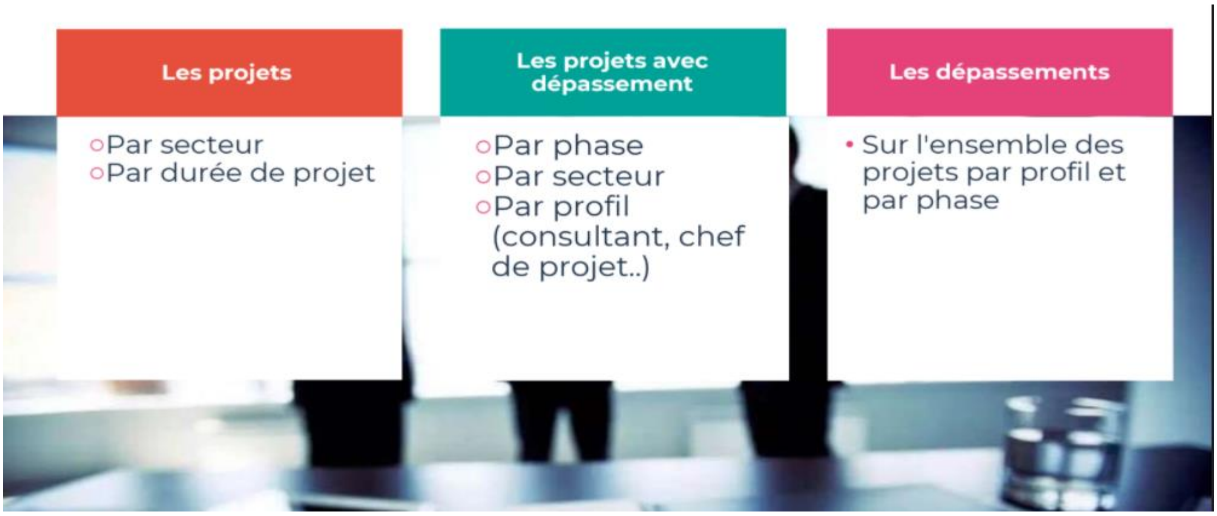
-Number of teams

-Number of clients

-The duration of project phase
-delay of phase
- Sum of delay duration

-total project workload
-total profile workload

Figure 6:Objective analysis tree



Les projets

- Par secteur
- Par durée de projet

Les projets avec dépassement

- Par phase
- Par secteur
- Par profil (consultant, chef de projet..)

Les dépassements

- Sur l'ensemble des projets par profil et par phase

Figure 7:the project objectives

6. functional requirements

1:Business Objectives

Business Objective	*Functional requirements*
Reduce project delays	View projects with delays (by phase / sector) View projects without delays

	View delay rate by phase View projects by sector View projects by duration and phase View phases that caused delays
Optimize resource allocation	View the delay rate by profile. View project with delays by profile
Ameliorate the billing progress	View billed projects. View billing project rate by sector View billing project rate by profile

7. Conclusion

In this first chapter, we presented the context of our project in order to make things clear, and to help you more understand the main idea.

Chapter 2: Methodology of work

1. Introduction

The main objective of this chapter is to present the methodology that we will work on, in order to manage our project well

2. Methodology

What is CRISP-DM?

The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM) is a process model with six phases:

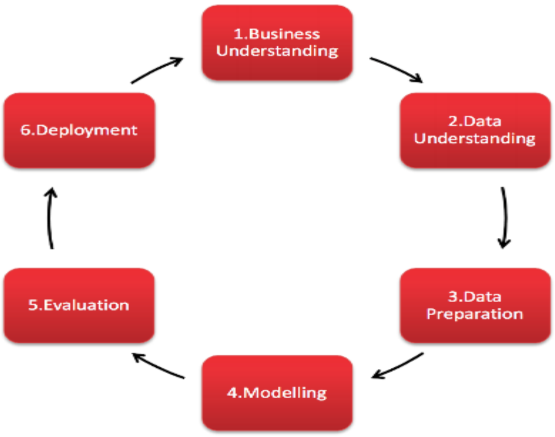


Figure 8 :Methodology

1) **Business understanding** – What does the business need?
2) **Data understanding** – What data do we have / need? Is it clean?
3) **Data preparation** – How do we organize the data for modeling?
4) **Modeling** – What modeling techniques should we apply?
5) **Evaluation** – Which model best meets the business objectives?
6) **Deployment** – How do stakeholders access the results?

Why CRISP-DM?

➔ **Flexibility**: This methodology makes it possible for models and processes to be imperfect at the very beginning.

➔ **Long-term Strategy**: CRISP-DM methodology allows to create a long-term strategy based on short iterations at the beginning of project development. During first iterations, a team can create a basic and simple model cycle that can easily be improved in further iterations.

➔ **Functional Templates**: Using a CRISP-DM approach makes it possible to develop functional templates for project management processes.

Steps

Business Understanding: Give context to the goals and to the data so that the developer/engineer gets a notion of the relevance of data in a particular business model.

Data Understanding: Understanding what can be expected and achieved from the data. This process goes through checking the quality of the data, in several terms, such as data completeness, values distributions and data governance compliance. This is a crucial part of the project because it defines how viable and trustworthy can be the final results.

Data Preparation: Involves the ETL process that turns the pieces of data into something useful using algorithms and process.

Modeling: This is the core of any machine learning project. This step is responsible for the results that should satisfy the project goals, some algorithms such as k-means, hierarchical clustering, time series, linear regression, k-nearest neighbors, and several others, are the core code lines of this step in the methodology.

Evaluation: Verify that the results are valid and correct. We can expect two results - Wrong results: In such a case, we have to go back to the first step, in order to understand why the results are mistaken. - Accurate results: If we reach a satisfying test accuracy results, we can move on to the next step.

Deployment: Presenting the results in a useful and understandable manner, and by achieving this, the project should achieve its goals. It is the only step not belonging to a cycle.

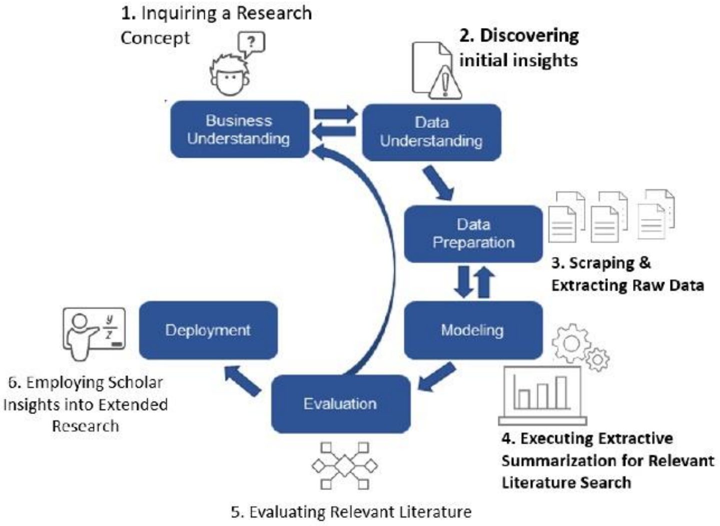

1. Inquiring a Research Concept
?
2. Discovering initial insights
Business Understanding
Data Understanding
Data Preparation
3. Scraping & Extracting Raw Data
Deployment
Modeling
6. Employing Scholar Insights into Extended Research
Evaluation
4. Executing Extractive Summarization for Relevant Literature Search
5. Evaluating Relevant Literature

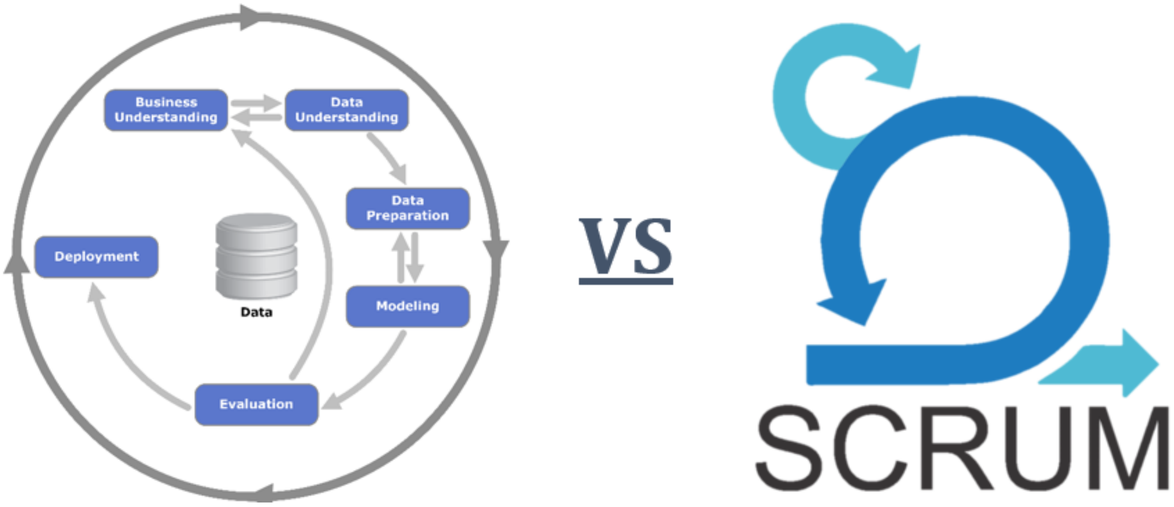
Figure 10:Project cycle

CRISP-DM VS Scrum

One key difference between the two approaches is what is being delivered. The deliverable for a Crisp-DM (Analytics) project is a piece of insight embedded into a decision space. The deliverable for an Agile (Software) project is a working piece of software or a product.
Another difference is that for Software projects developed using Agile methods focus on the needs of the end user. Gathering requirements using user-stories is a very powerful tool and works really well for software projects. Focusing on the end user may not always work well in an analytics context. Usually, focusing on the business objective works much better, because in most cases, the end user will never know of the existence of the analytics output.

Actually, although Scrum is the most used methodology but it’s used for developing software projects that’s why it’s not a perfect match for analytics project.

Moreover, Crisp-Dm is better methodology for modern analytics project.



3. Conclusion

After setting the objectives and understanding the methodology, in the next chapter we will present the data and explain each field in the database.

Chapter 3: Data source identification and Modeling

1. Introduction

The ability to conduct effective research and analysis depends on data collection and a good understanding of a database.

2. Data tables identification

2:bill

Database	Tables	Columns	Description
Projeqtor	This table contains informations about bills	id	Bill id
		Name	Bill’s name
		idProject	Id of project
		idCilent	Id of client
		Done	Payment status 1 done 0 not done
		paymentDate	Real date of payment
		PaymentDueDate	Date of payment

3:Project

Database	Tables	Columns	Description
Projeqtor	This table contains informations about projects	id	Id of project
		Name	Name of project
		idClient	Id of client
		projectCode	Code of project
		Done	Project done or under progress 1 done 0 not yet
		DoneDate	End project date
		paymentDelay	Number of days after payment date
		idStatus	Id of project status
		ClientCode	Code of client
		creationDate	Date of creation
		idResource	Id of employees

4:Client

Database	Tables	Columns	Description
Projeqtor	This table contains informations about clients of Talys	Id	Id of client
		name	Name of client
		clientCode	Code of client
		PaymentDelay	Delay of payment

5:Resource

Database	Table	Columns	Description
Projeqtor	This table contains informations about employees of Talys	Id	Id of employee
		Name	Name of employee
		idProfile	Id of profile
		isResource	0 free 1 not free
		idTeam	Id of team
		idClient	Id of client

6:Profile

Database	Table	Columns	Description
Projeqtor	This table contains description of profile	Id	Id of profile
		Name	Name of profile
		ProfileCode	Code of profile

7:Role

Database	Table	Columns	Description
Projeqtor	This table contains description of role	Id	Id of role
		Name	Name of profile

8:Status

Database	Table	Columns	Description
Projeqtor	This table contains description of status of project	Id	Id of status
		Name	Name of status
		SetDoneStatus	1 done 0 in progress

9:Team

Database	Table	Columns	Description
Projeqtor	This table contains description of team	Id	Id of team
		Name	Name of team

10:Activity

Database	Table	Columns	Description
Projeqtor	This table contains description of activity	Id	Id of activity
		IdProjet	Id of project
		Name	Name of activity
		CreationDate	Date of creation
		IdStatus	Id of status
		IdRessource	Id of resource
		DoneDate	End date of activity

11 :Assignment

Database	Table	Columns	Description
Projeqtor	This table contains description of assignment	Id	Id of assignment
		IdRessource	Id of resource
		IdProject	Id of project
		realStartDate	Real start date of assignment
		realEndDate	Real end date of assignment
		plannedStartDate	Planned start date of assignment
		plannedEndDate	Planned end date of assignment
		idRole	Id of role

3. the Inmon approach

Bill Inmon, the father of data warehousing, came up with the concept to develop a data warehouse that starts with designing the corporate data warehouse data model, which identifies the main subject areas and entities the enterprise works with, such as customer, product, vendor, and so on.

Bill Inmon's definition of a data warehouse is that it is a "subject-oriented, nonvolatile, integrated, time-variant collection of data in support of management's decisions".

4. Inmon vs Kimball

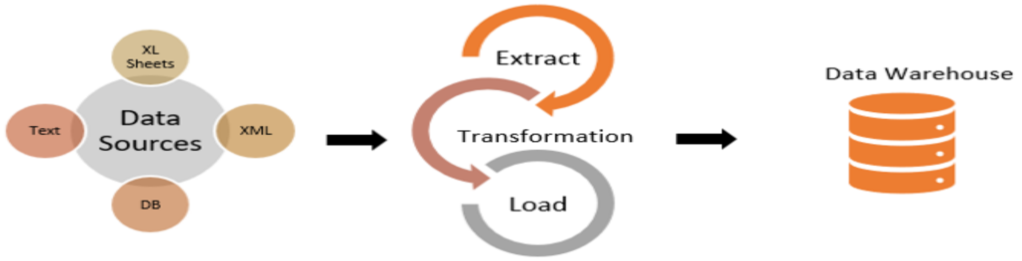


Figure 11:the Inmon approach

The Inmon approach is referred to as the top-down or data-driven approach, whereby we start from the data warehouse and break it down into data marts, specialized as needed to meet the needs of different departments within the organization, such as finance, accounting, HR.

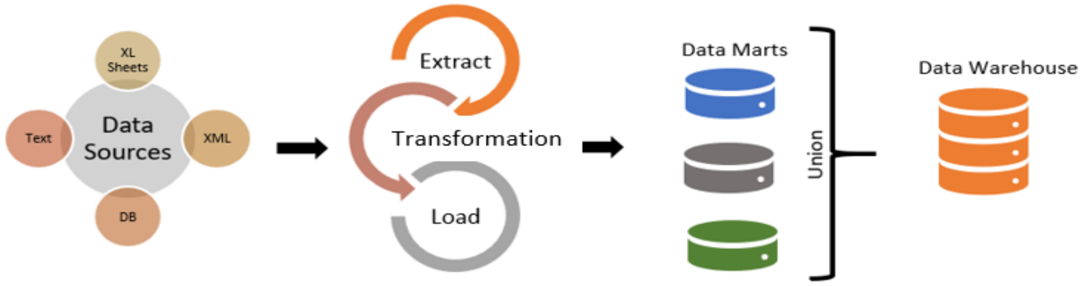


Figure 12:the Kimball approach

The Kimball approach is referred to as bottom-up or user-driven, because we start from the user-specific data marts, and these form the very building blocks of our conceptual data warehouse. It’s important to know from the outset which model best suits your needs so that it can be built into the data warehouse schema.

12:table Inmon vs Kimball

	Kimball	Inmon
Introduced by	Introduced by Ralph Kimball.	Introduced by Bill Inmon.
Approach	It has Bottom-Up Approach for implementation.	It has Top-Down Approach for implementation.
Data Integration	It focuses Individual business areas.	It focuses Enterprise-wide areas.
Building Time	It is efficient and takes less time.	It is complex and consumes a lot of time.
Cost	It has iterative steps and is cost effective.	Initial cost is huge and development cost is low.
Skills Required	It does not need such skills but a generic team will do job.	It needs specialized skills to make work.
Maintenance	Maintenance is difficult.	Here maintenance is easy.
Data Model	It prefers data to be in De-normalized model.	It prefers data to be in normalized model.
Data Store Systems	In this, source systems are highly stable.	In this, source systems have high rate of change.

5. Model star schema

The star data model owes its name to its shape. This design model favours the user approach, the business orientation.

The reference table contains the facts. The facts or measures are the figures (such as results by sector). The satellite tables correspond to the dimensions. These are the user analysis axes.

Star schema is the fundamental schema among the data mart schema and it is simplest. This schema is widely used to develop or build a data warehouse and dimensional data marts. It includes one or more fact tables indexing any number of dimensional tables.

Why the model star schema in data warehouse:

- **Simpler Queries:**
  Join logic of star schema is quite cinch in compare to other join logic which are needed to fetch data from a transactional schema that is highly normalized.
- **Simplified Business Reporting Logic:**
  In compared to a transactional schema that is highly normalized, the star schema makes simpler common business reporting logic, such as as-of reporting and period-over-period.
- **Loading Cubes:**
  Star schema is widely used by all OLAP systems to design OLAP cubes efficiently. In fact, major OLAP systems deliver a ROLAP mode of operation which can use a star schema as a source without designing a cube structure.

6. Model explication

In our case we have one fact table that we judge necessary, which is joined to specific dimensions.

DimDate: a dimension that handles the time spans related to the fact table. It has many columns that are able to precisely describe the notion of time, with for example the name of the day, the name of the quarter etc…

DimSector: a dimension that contains different domains and category exploied in the project . its columns are Category, Name, Status, and done.

DimPhase: a dimension that indicates a fixed length events of one month or less to create consistency. it has many columns that describe this period of time such as the name, status,result, creationDate, handlded, handledDate, Done, DoneDate.

DimDelay: a dimension that specialize in the time by which something is late or postponed, its colmuns are RealStartDate, RealEndDate, PlannedStartDate, PlannedEndDate which are related to timing.

DimResource: a dimension that contains the different human resources who works for the company and it has many colmuns that define these resources such as the name, profile; role, capacity, team etc ...

DimClient: a dimension that describes the company's clients with just the columns Name and paymentdelay.

DimBill: a dimension that deals with the billing system and information about the payment date and type…. the columns of this dimension are billingType, paymentDate, paymentDone, paymentDelay …

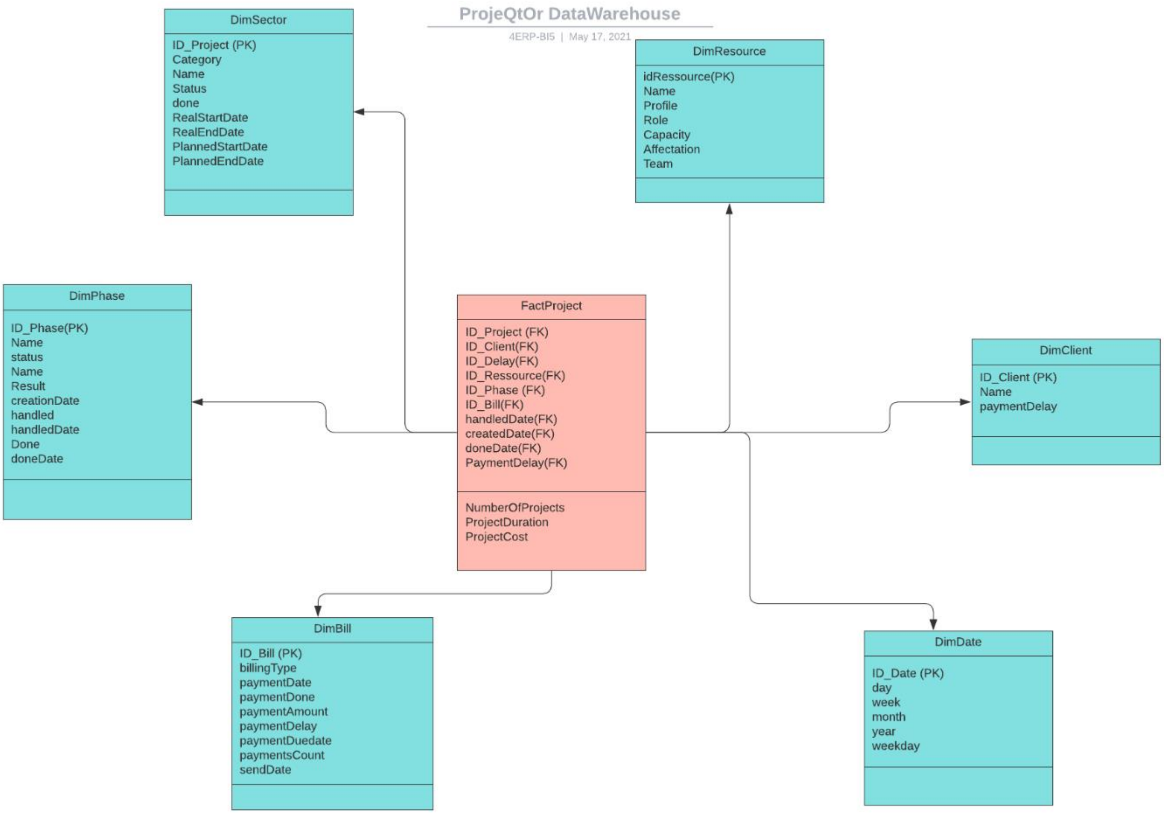


Figure 13:Modeling of data Warehouse

7. Data integration with Talend

- Loading the Phase dimension from the source data table to the Datawarehouse

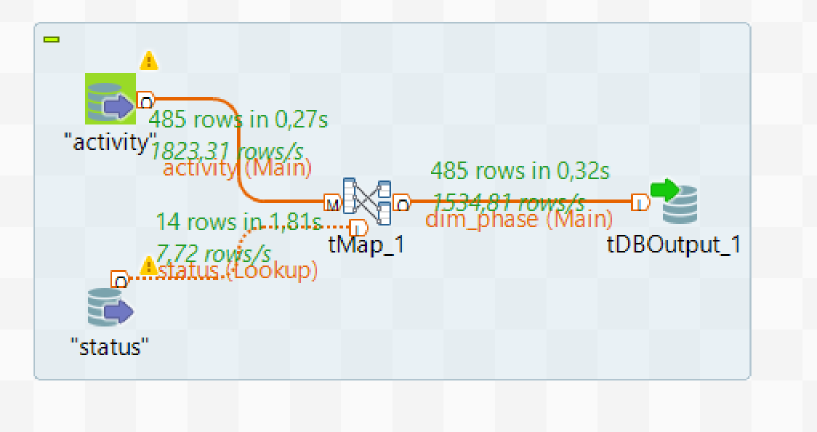


Figure *14:dim_phase*

- Loading the Bill dimension from the source data table to the Datawarehouse

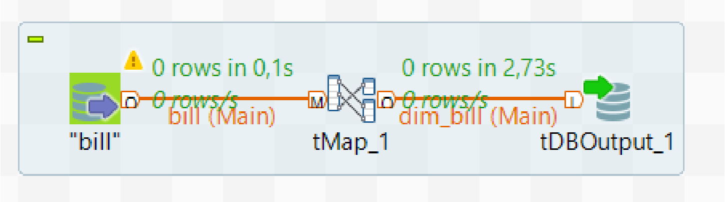


Figure 15:dim_bill

- Loading the Date dimension generated through the tRowGenerator

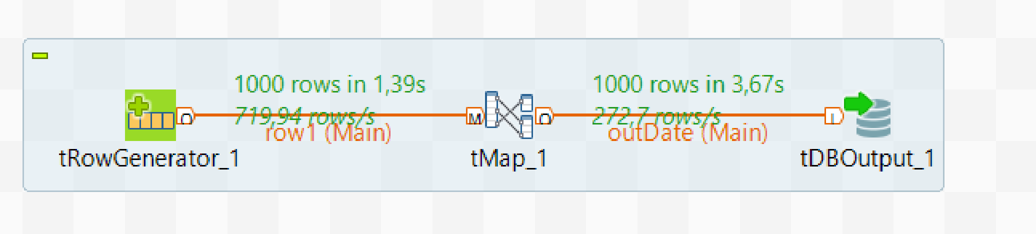


Figure 16:dim_date

- Loading the Delay dimension from the source data table to the Datawarehouse

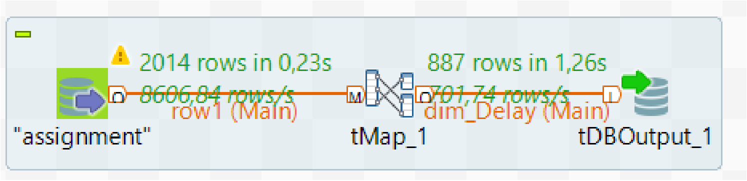
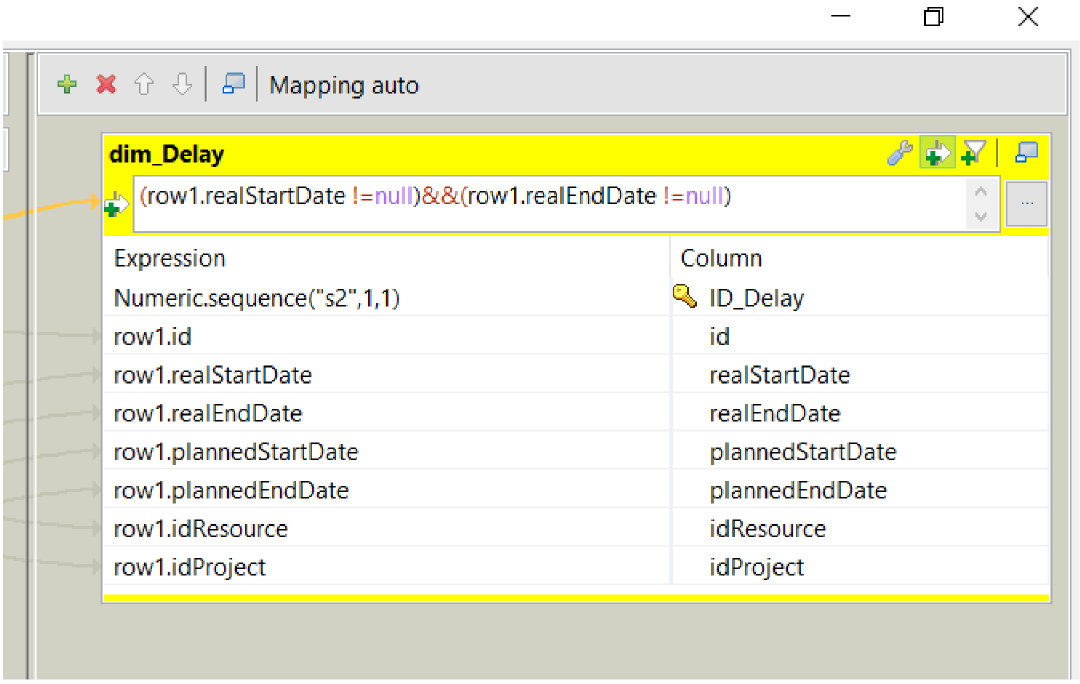


Figure 17:dim_delay

- Filtered the delay rows , both the start date and the end date must contain data in order to calculate the project duration correctly



```
(row1.realStartDate !=null)&&(row1.realEndDate !=null)
```

Expression	Column
Numeric.sequence("s2",1,1)	ID_Delay
row1.id	id
row1.realStartDate	realStartDate
row1.realEndDate	realEndDate
row1.plannedStartDate	plannedStartDate
row1.plannedEndDate	plannedEndDate
row1.idResource	idResource
row1.idProject	idProject

Figure 18:filter dim_delay

- Loading the Sector dimension from the source data table to the Datawarehouse

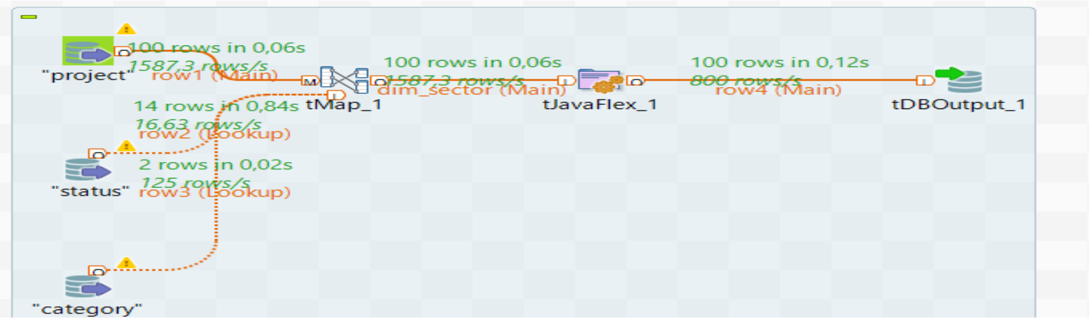


Figure 19:dim_sector

- Loading the Client dimension from the source data table to the Datawarehouse

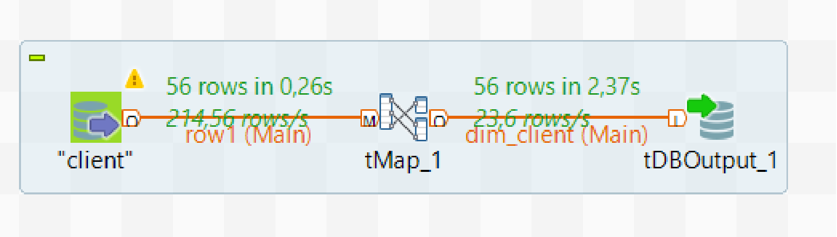


Figure 20:dim_client

- Loading the Resource dimension from the source data table to the Datawarehouse

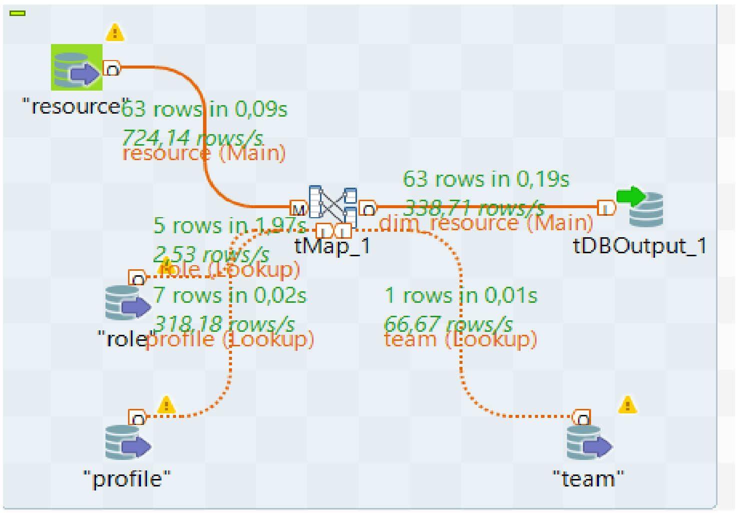


Figure 21:dim_resource

- Loading the Fact table with dim Sector as the main data source and the others as lookup dimensions

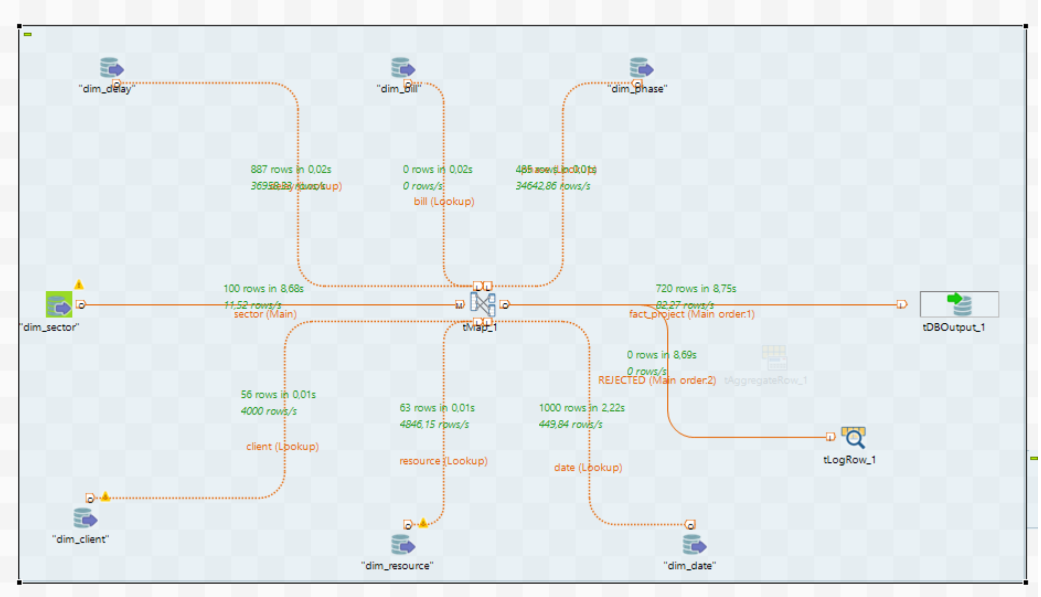
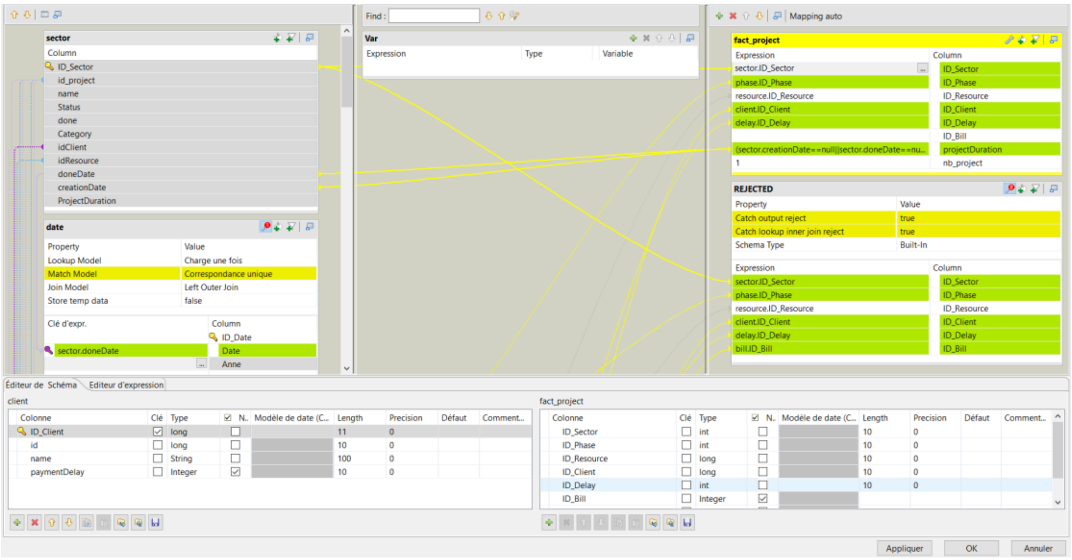


Figure 22:Fact_project

- The tMap Component structure with the join between the Datawarehouse tables



sector

Column
ID_Sector
id_project
name
Status
done
Category
idClient
idResource
doneDate
creationDate
ProjectDuration

date

Property	Value
Lookup Model	Charge une fois
Match Model	Correspondance unique
Join Model	Left Outer Join
Store temp data	false

Clé d'expr.	Column
	ID_Date
sector.doneDate	Date
...	Anne

fact_project

Expression	Column
sector.ID_Sector	ID_Sector
phase.ID_Phase	ID_Phase
resource.ID_Resource	ID_Resource
client.ID_Client	ID_Client
delay.ID_Delay	ID_Delay
	ID_Bill
(sector.creationDate==null||sector.doneDate==nu...	projectDuration
1	nb_project

REJECTED

Property	Value
Catch output reject	true
Catch lookup inner join reject	true
Schema Type	Built-In

Expression	Column
sector.ID_Sector	ID_Sector
phase.ID_Phase	ID_Phase
resource.ID_Resource	ID_Resource
client.ID_Client	ID_Client
delay.ID_Delay	ID_Delay
bill.ID_Bill	ID_Bill

Éditeur de Schéma | Editeur d'expression

client

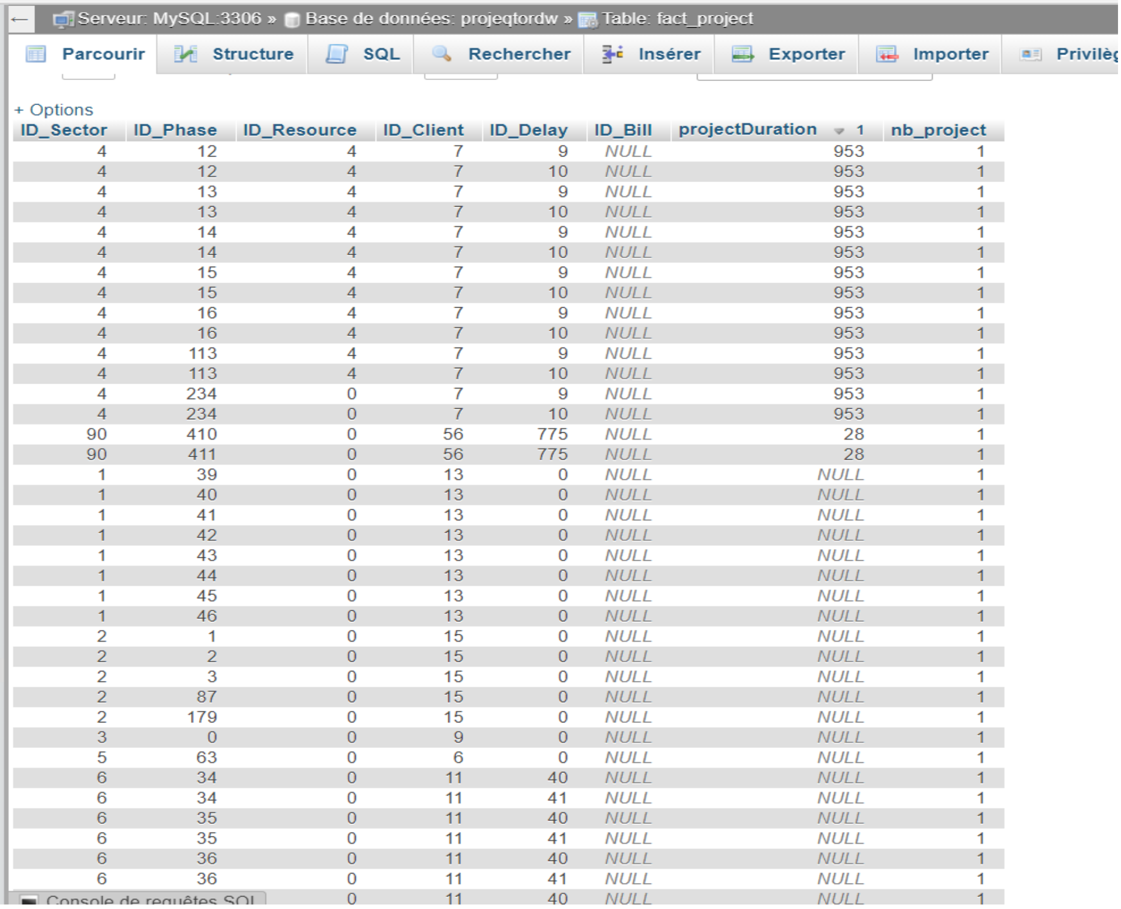
Colonne	Clé	Type	N..	Modèle de date (C...	Length	Precision	Défaut	Comment...
ID_Client	☑	long	☐		11	0		
id	☐	long	☐		10	0		
name	☐	String	☐		100	0		
paymentDelay	☐	Integer	☑		10	0		

fact_project

Colonne	Clé	Type	N..	Modèle de date (C...	Length	Precision	Défaut	Comment...
ID_Sector	☐	int	☐		10	0		
ID_Phase	☐	int	☐		10	0		
ID_Resource	☐	long	☐		10	0		
ID_Client	☐	long	☐		10	0		
ID_Delay	☐	int	☐		10	0		
ID_Bill	☐	Integer	☑					

Figure 23:tMap Fact_project

Result



Serveur: MySQL:3306 » Base de données: projeqtordw » Table: fact_project
Parcourir
Structure
SQL
Rechercher
Insérer
Exporter
Importer
Privilèg
+ Options

ID_Sector	ID_Phase	ID_Resource	ID_Client	ID_Delay	ID_Bill	projectDuration ▼ 1	nb_project
4	12	4	7	9	NULL	953	1
4	12	4	7	10	NULL	953	1
4	13	4	7	9	NULL	953	1
4	13	4	7	10	NULL	953	1
4	14	4	7	9	NULL	953	1
4	14	4	7	10	NULL	953	1
4	15	4	7	9	NULL	953	1
4	15	4	7	10	NULL	953	1
4	16	4	7	9	NULL	953	1
4	16	4	7	10	NULL	953	1
4	113	4	7	9	NULL	953	1
4	113	4	7	10	NULL	953	1
4	234	0	7	9	NULL	953	1
4	234	0	7	10	NULL	953	1
90	410	0	56	775	NULL	28	1
90	411	0	56	775	NULL	28	1
1	39	0	13	0	NULL	NULL	1
1	40	0	13	0	NULL	NULL	1
1	41	0	13	0	NULL	NULL	1
1	42	0	13	0	NULL	NULL	1
1	43	0	13	0	NULL	NULL	1
1	44	0	13	0	NULL	NULL	1
1	45	0	13	0	NULL	NULL	1
1	46	0	13	0	NULL	NULL	1
2	1	0	15	0	NULL	NULL	1
2	2	0	15	0	NULL	NULL	1
2	3	0	15	0	NULL	NULL	1
2	87	0	15	0	NULL	NULL	1
2	179	0	15	0	NULL	NULL	1
3	0	0	9	0	NULL	NULL	1
5	63	0	6	0	NULL	NULL	1
6	34	0	11	40	NULL	NULL	1
6	34	0	11	41	NULL	NULL	1
6	35	0	11	40	NULL	NULL	1
6	35	0	11	41	NULL	NULL	1
6	36	0	11	40	NULL	NULL	1
6	36	0	11	41	NULL	NULL	1
		0	11	40	NULL	NULL	1

Figure 24:table Fact_project

8. Conclusion

After identifying the source of the data and clarifying it, now we are ready to set up the architecture of our solution that we will embrace throughout the project

Chapter 4: Data Visualization

1. Introduction

2. Tools and technologies

Microsoft Power BI is a business intelligence platform that provides nontechnical business users with tools for aggregating, analyzing, visualizing and sharing data. Power BI’s user interface is intuitive for users familiar with Excel and its deep integration with other Microsoft products makes it a very versatile self-service tool that requires little upfront training.

Figure 25:PowerBi logo

3. Home Dashboard

The main dashboard contains analysis based on 3 axes :

a. Number of projects

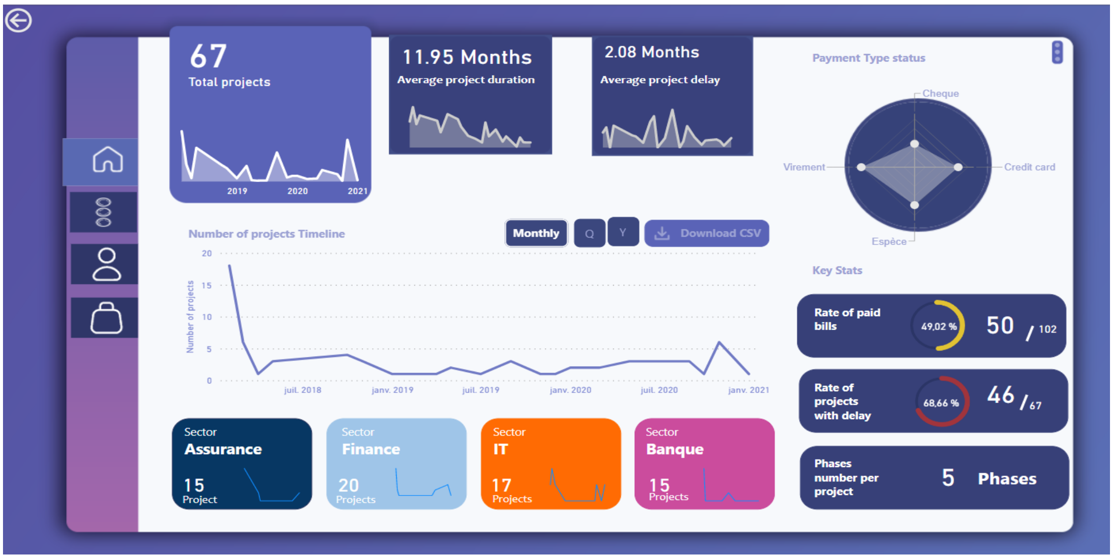
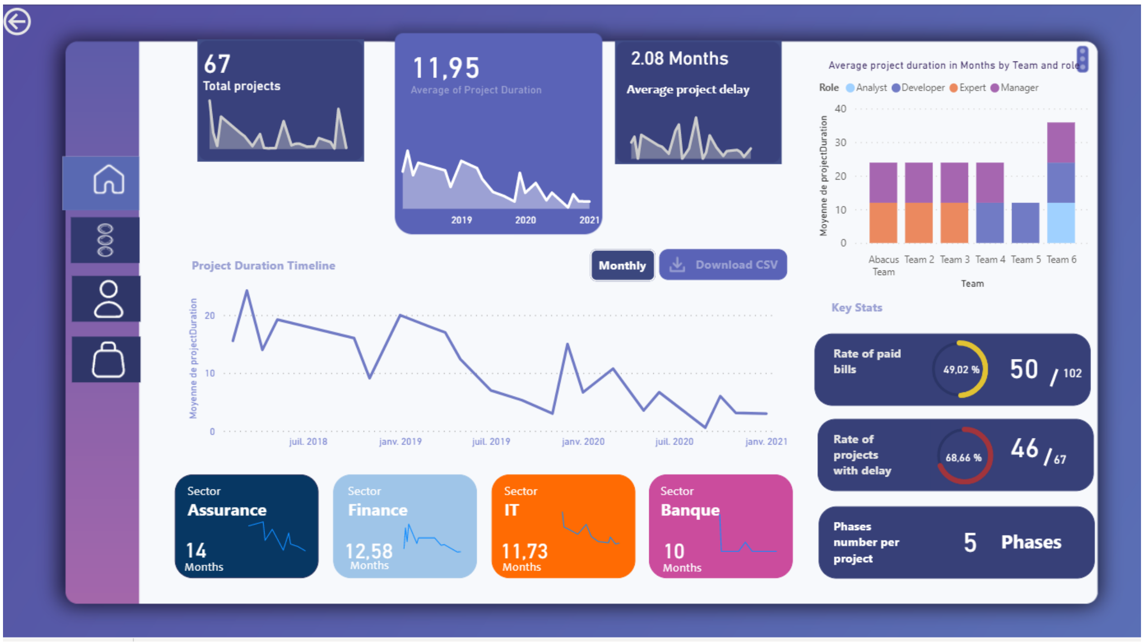


Figure 26:Main Dashboard

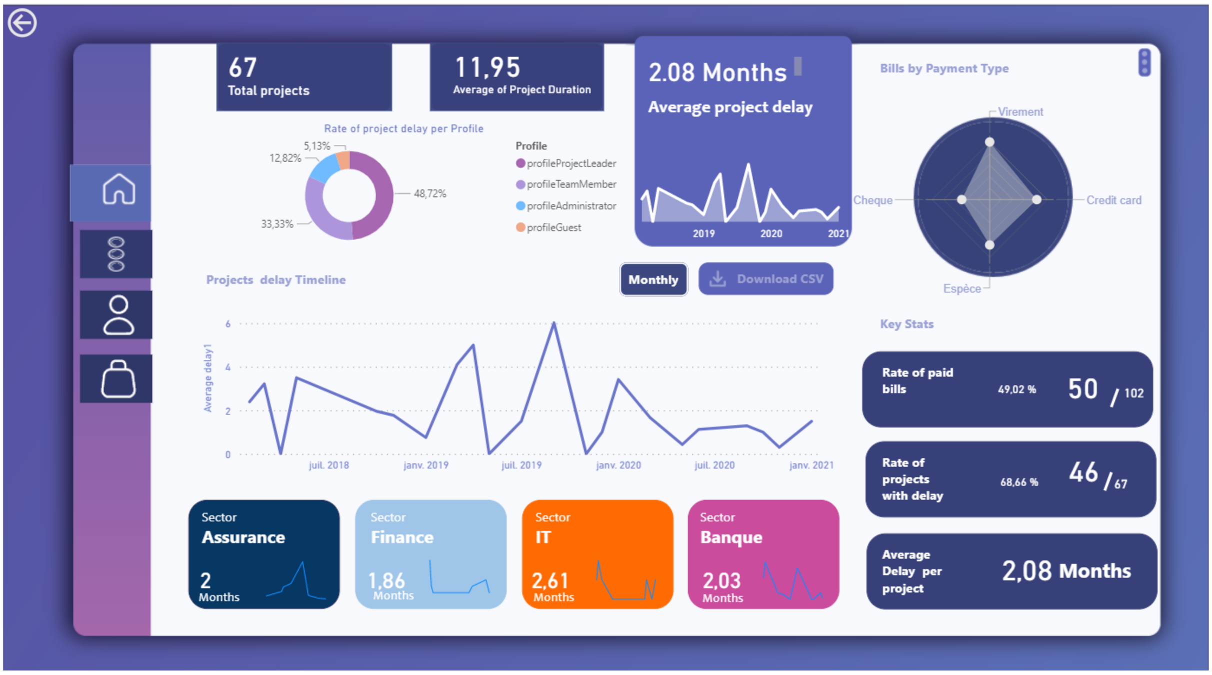
- Number of projects timeline (Monthly /Yearly / Quarterly)
- Number of projects per sector / timeline

b. Project Duration



- Average project duration timeline (Monthly)
- Average project duration per sector / timeline
- Average project duration by team and role

c. Project Delay



- Average project delay /timeline
- Rate of project delay per profile

4. Project Analysis details Dashboard

Project’s analyisis dashboard contains:

- Total of projects per year per sector
- Rate of projects per sector
- Rate of projects per team
- Number of projects in each sector

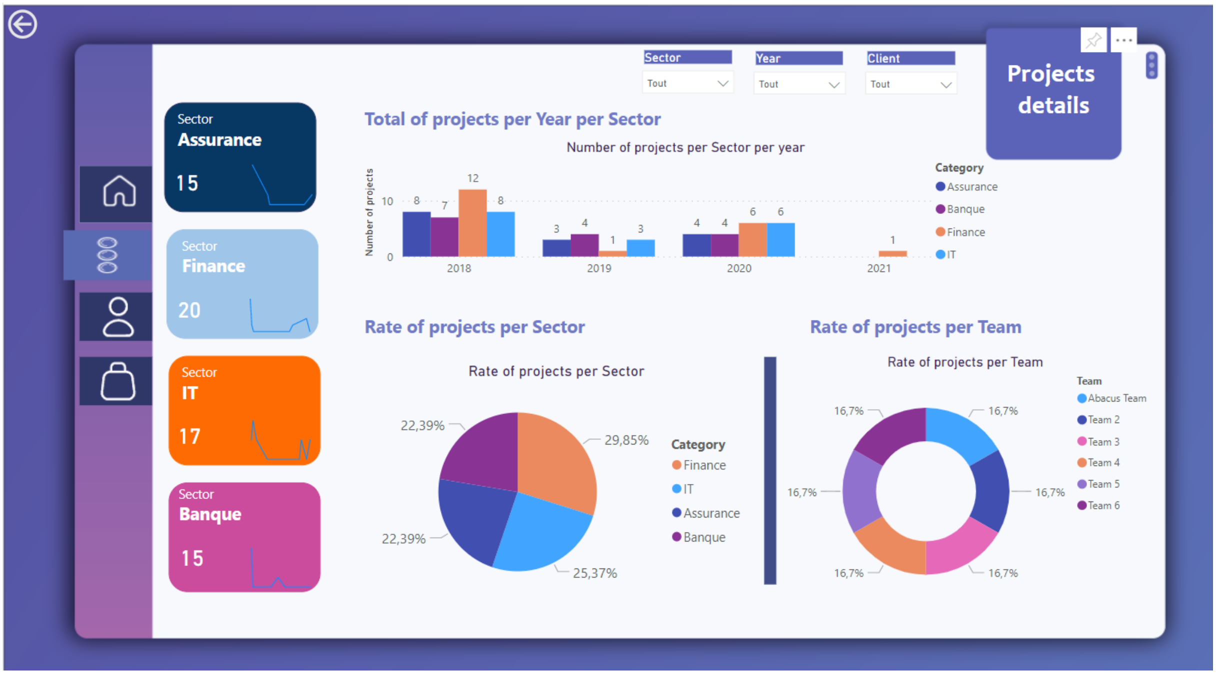
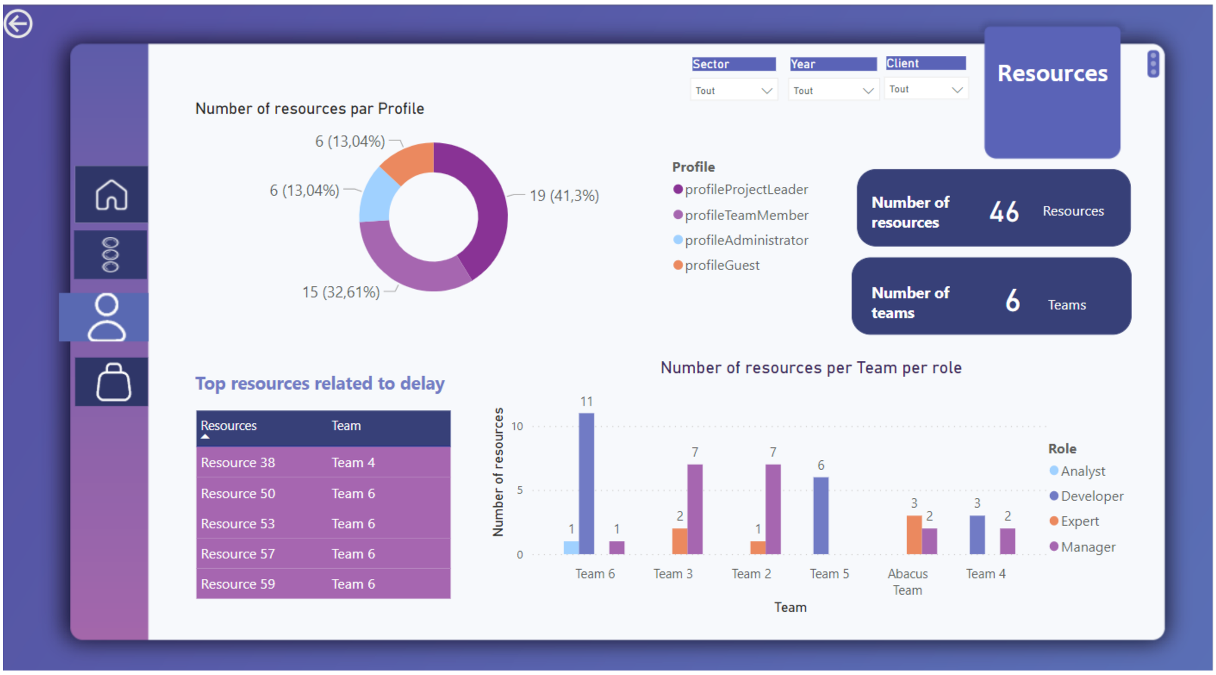


Figure 27 :Project’s duration Dashboards

5. Resources Dashboard

Resource’s dashboard contains:

- Top 5 best resources related to delay
- Number of resources per team per role
- Number of resources per profile
- Number of resources
- Number of teams



Top resources related to delay

Resources	Team
Resource 38	Team 4
Resource 50	Team 6
Resource 53	Team 6
Resource 57	Team 6
Resource 59	Team 6

Figure 28:Resources Dashboard

6. Billing Dashboard

Billing dashboard contains:

- Rate of paid bills
- Rate of projects with delay
- Rate of projects with delay
- Number of clients by payment and delivery status

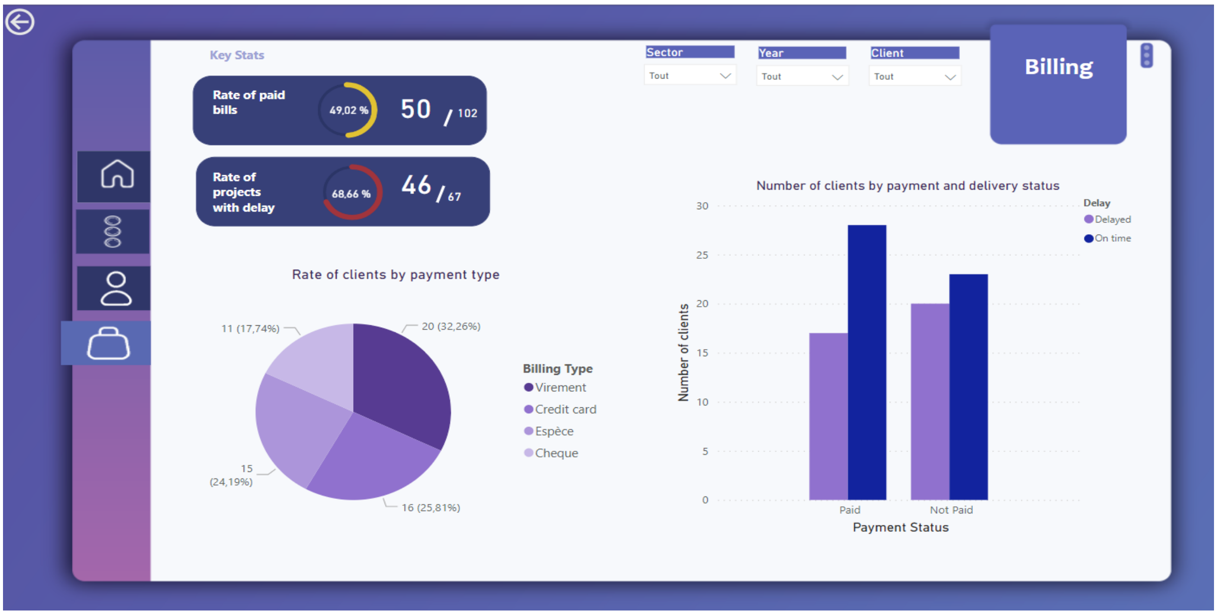


Figure 29:Billing Dashboard

7. Conclusion

After we visualized data dynamically by using dashboards and charts , we will present you our data analysis modeling and evaluation.

Chapter 5: Data Analysis Modeling and Evaluation

1. Introduction

Data modeling is a set of tools and techniques used to understand and analyse how an organisation should collect, update, and store data.

It is a critical skill for the business analyst who is involved with discovering, analysing, and specifying changes to how software systems create and maintain information and data evaluation process determines whether data is usable for calculating risk estimates.

Data that is unusable for calculating the risk estimates still may provide useful information for determining the distribution which describes the probability of any potential value.

2. Tools and technologies

Jupyter

Jupyter is a web application used to program in more than 40 programming languages, including Python, Julia, Ruby, R, or Scala2. Jupyter is an evolution of the IPython project. Jupyter allows you to make notebooks or notebooks, that is to say programs containing both text in markdown and code in Julia, Python, R ... These notebooks are used in data science to explore and analyze Datas.

Figure 30:JupyterLab logo

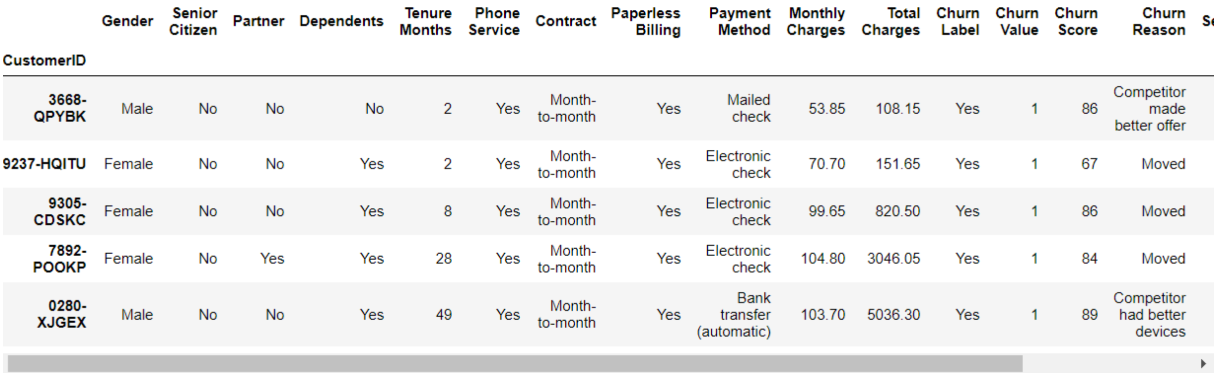
Python

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant 18 white space. It provides constructs that enable clear programming on both small and large scales.

Figure 31:Python logo

3. Descriptive analytics

A. external data



	Gender	Senior Citizen	Partner	Dependents	Tenure Months	Phone Service	Contract	Paperless Billing	Payment Method	Monthly Charges	Total Charges	Churn Label	Churn Value	Churn Score	Churn Reason	S
CustomerID																
3668-QPYBK	Male	No	No	No	2	Yes	Month-to-month	Yes	Mailed check	53.85	108.15	Yes	1	86	Competitor made better offer	
9237-HQITU	Female	No	No	Yes	2	Yes	Month-to-month	Yes	Electronic check	70.70	151.65	Yes	1	67	Moved	
9305-CDSKC	Female	No	No	Yes	8	Yes	Month-to-month	Yes	Electronic check	99.65	820.50	Yes	1	86	Moved	
7892-POOKP	Female	No	Yes	Yes	28	Yes	Month-to-month	Yes	Electronic check	104.80	3046.05	Yes	1	84	Moved	
0280-XJGEX	Male	No	No	Yes	49	Yes	Month-to-month	Yes	Bank transfer (automatic)	103.70	5036.30	Yes	1	89	Competitor had better devices	

Figure 32 :Description of the dataset

The dataset contains 7043 rows and 18 columns.

Explore data

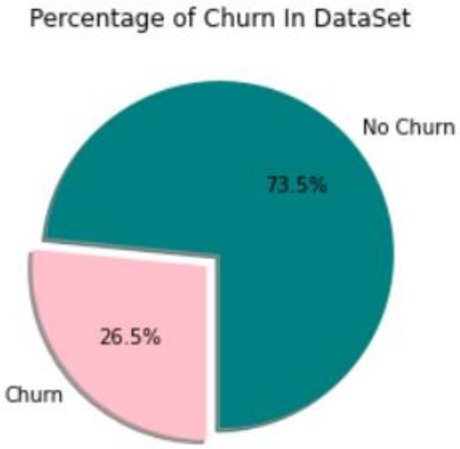


Figure 33:Diagram churn/no churn

In detail we have a look at the target feature, the actual ―Churn‖. Therefore we plot it accordingly and see that 26,5% Of the total amount of customer churn. This is important to know so we have the same proportion of Churned Customers to Non-Churned Customers in our training data.

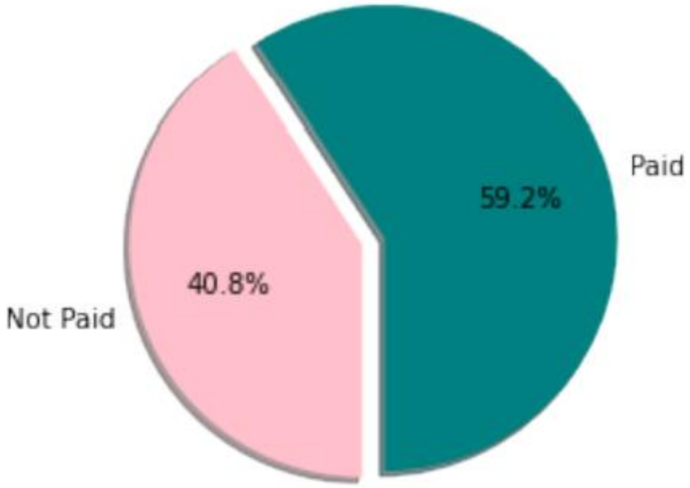


Figure 34:diagram paid / not paid project

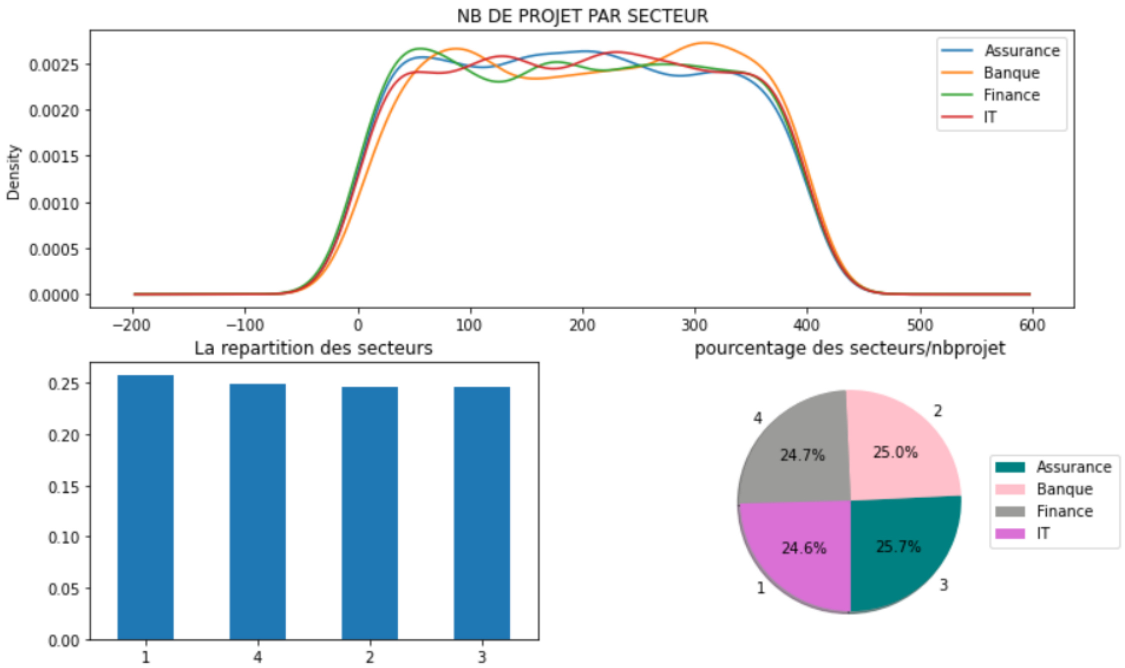
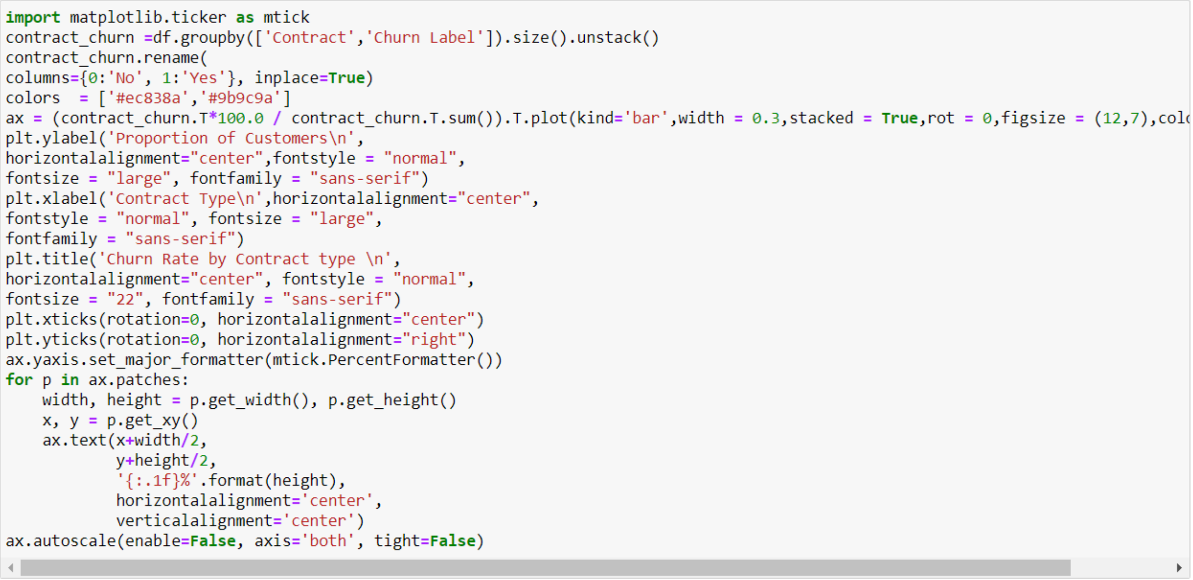


Figure 35:the distribution of sectors on projects



```
import matplotlib.ticker as mtick
contract_churn =df.groupby(['Contract','Churn Label']).size().unstack()
contract_churn.rename(
columns={0:'No', 1:'Yes'}, inplace=True)
colors  = ['#ec838a','#9b9c9a']
ax = (contract_churn.T*100.0 / contract_churn.T.sum()).T.plot(kind='bar',width = 0.3,stacked = True,rot = 0,figsize = (12,7),col
plt.ylabel('Proportion of Customers\n',
horizontalalignment="center",fontstyle = "normal",
fontsize = "large", fontfamily = "sans-serif")
plt.xlabel('Contract Type\n',horizontalalignment="center",
fontstyle = "normal", fontsize = "large",
fontfamily = "sans-serif")
plt.title('Churn Rate by Contract type \n',
horizontalalignment="center", fontstyle = "normal",
fontsize = "22", fontfamily = "sans-serif")
plt.xticks(rotation=0, horizontalalignment="center")
plt.yticks(rotation=0, horizontalalignment="right")
ax.yaxis.set_major_formatter(mtick.PercentFormatter())
for p in ax.patches:
    width, height = p.get_width(), p.get_height()
    x, y = p.get_xy()
    ax.text(x+width/2,
            y+height/2,
            '{:.1f}%'.format(height),
            horizontalalignment='center',
            verticalalignment='center')
ax.autoscale(enable=False, axis='both', tight=False)
```

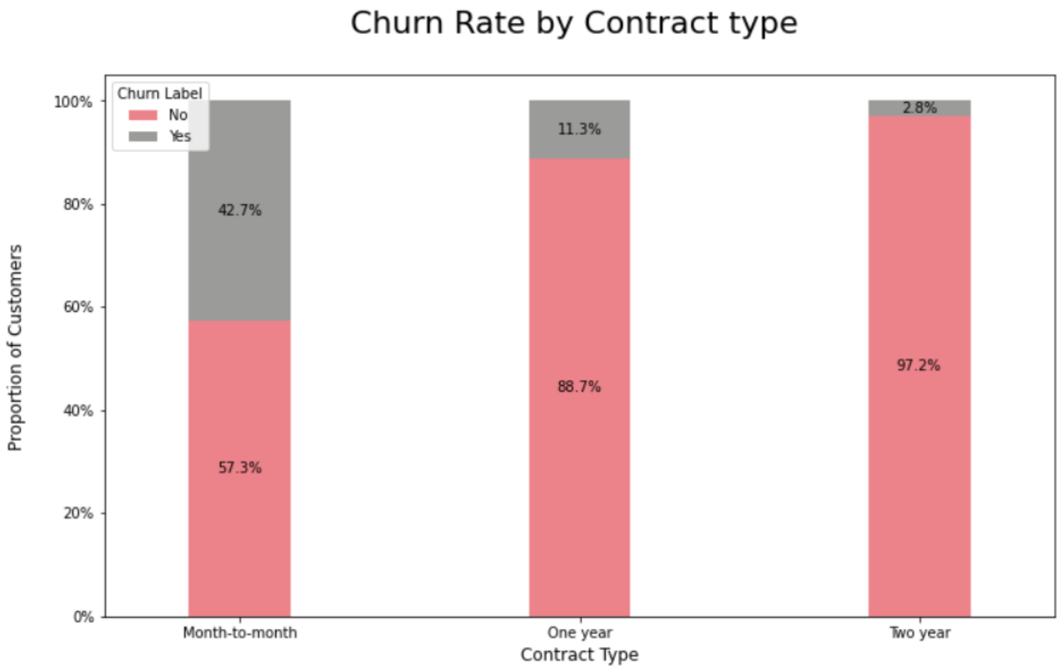
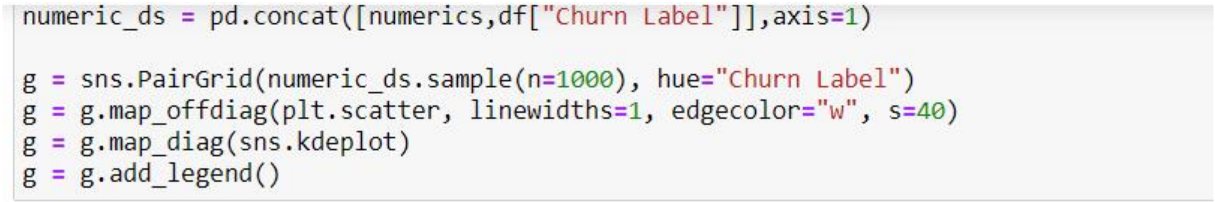


Figure 36:churn rate by contract type

Churn Rate by Contract Type : Customers with a prepaid or rather a month-to-month connection have a very high probability of churn compared to their peers on 1 or 2 years contracts.



```
numeric_ds = pd.concat([numerics,df["Churn Label"]],axis=1)

g = sns.PairGrid(numeric_ds.sample(n=1000), hue="Churn Label")
g = g.map_offdiag(plt.scatter, linewidths=1, edgecolor="w", s=40)
g = g.map_diag(sns.kdeplot)
g = g.add_legend()
```

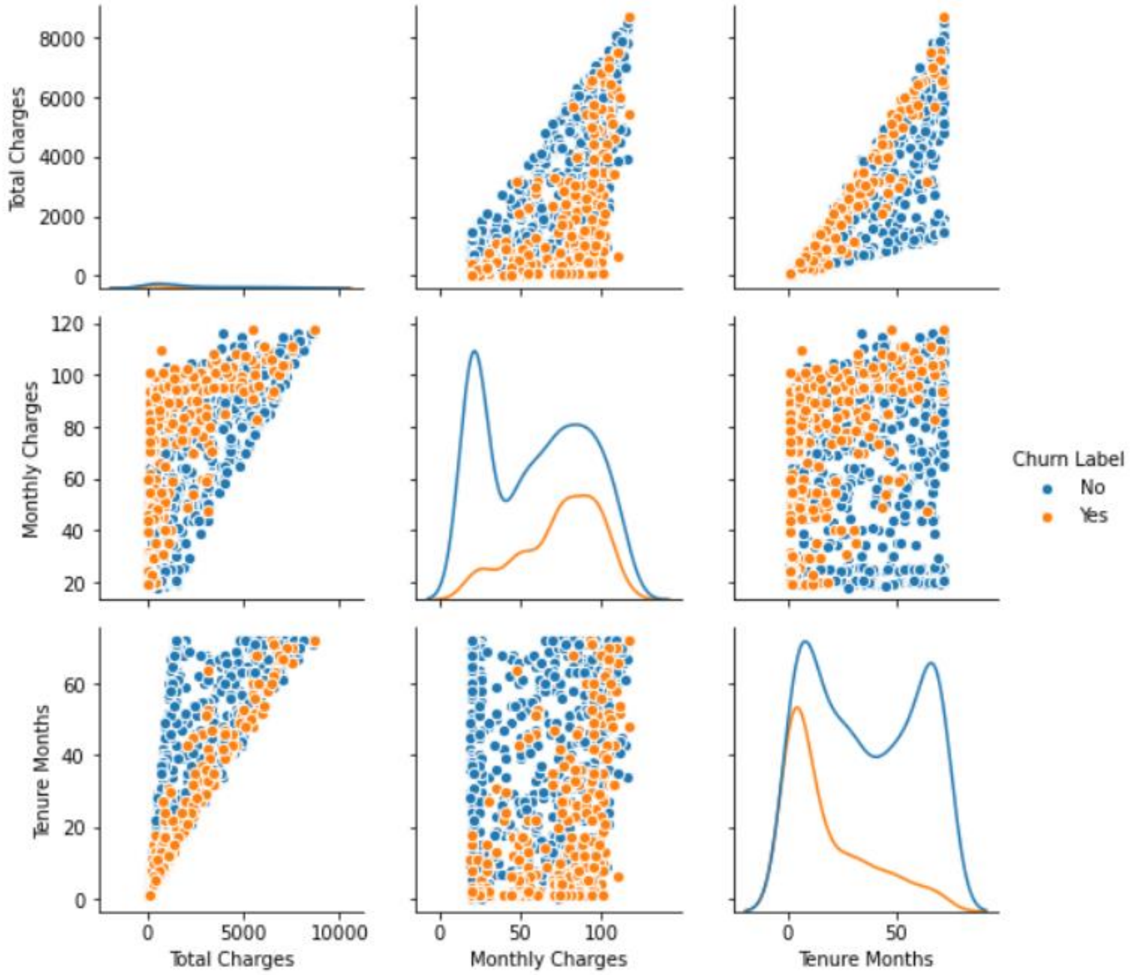
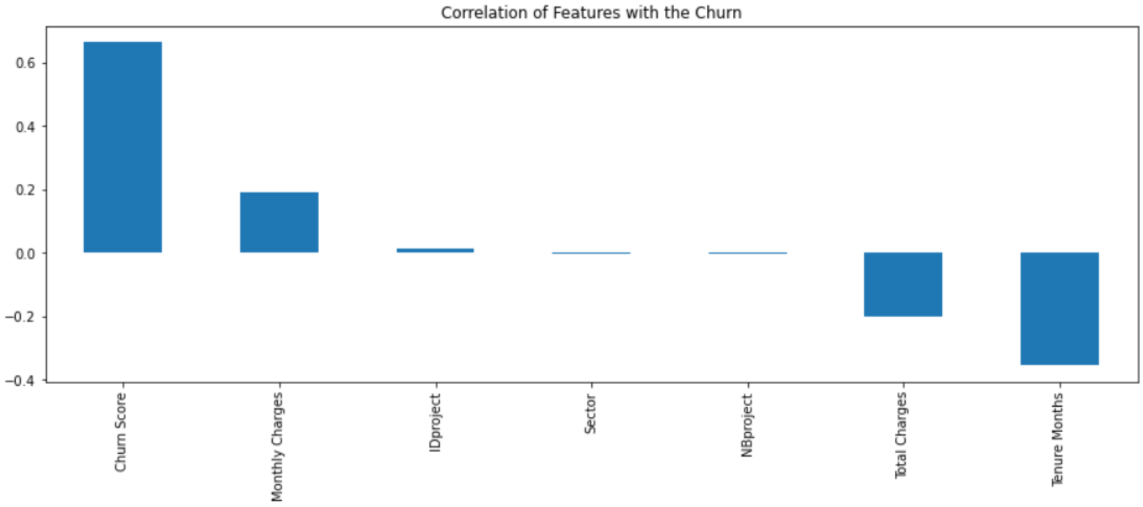


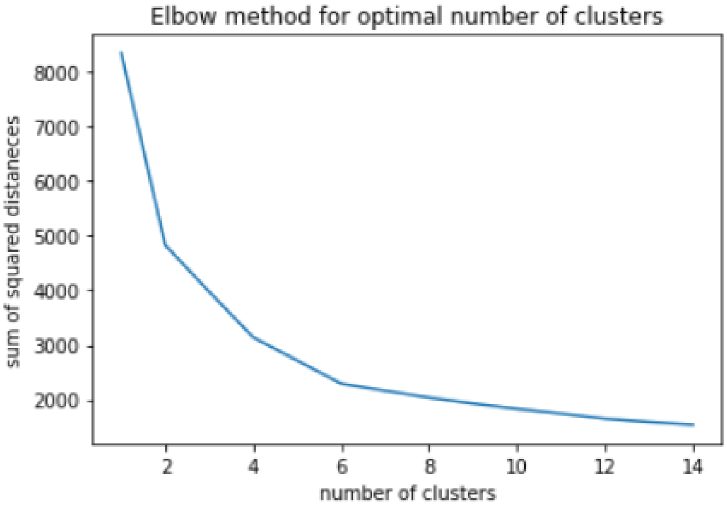
Figure 37:Plots related on tenure months, monthly charges,tenure months

From the plots we can say that both the tenure months and the Monthly Charges are looking like good predictors of the Churn Values variable.

Clustering



Graphical representation of the correlation of each variable with the target variable ««Churn»



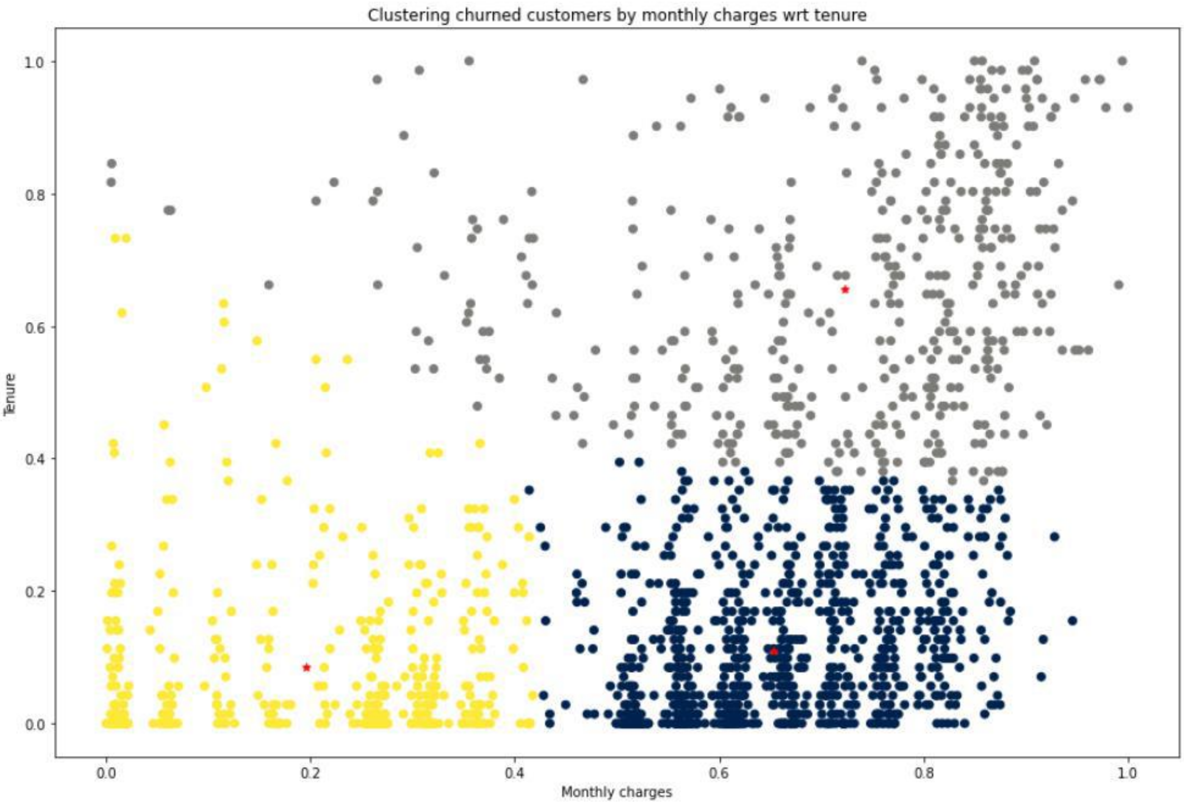
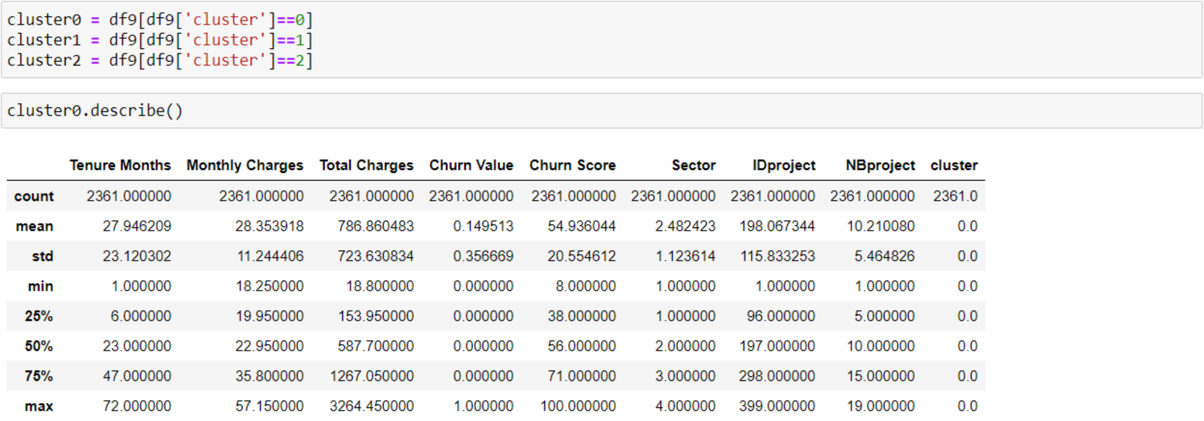


Fig clustering churned customers by monthly charges with tenure

Customers with low monthly charges and low tenure: Could have been a temporary connection for them or people looking for very minimal service who found a service provider offering even lower charges for basic services and churned quickly despite low monthly charges.

Customers with high monthly charges and low tenure: The heaviest concentration of churned users. The most common churned users who were possibly unhappy with the prices and stayed for a little while before quickly leaving the service provider for better , cheaper options.



```
cluster0 = df9[df9['cluster']==0]
cluster1 = df9[df9['cluster']==1]
cluster2 = df9[df9['cluster']==2]
```

```
cluster0.describe()
```

	Tenure Months	Monthly Charges	Total Charges	Churn Value	Churn Score	Sector	IDproject	NBproject	cluster
count	2361.000000	2361.000000	2361.000000	2361.000000	2361.000000	2361.000000	2361.000000	2361.000000	2361.0
mean	27.946209	28.353918	786.860483	0.149513	54.936044	2.482423	198.067344	10.210080	0.0
std	23.120302	11.244406	723.630834	0.356669	20.554612	1.123614	115.833253	5.464826	0.0
min	1.000000	18.250000	18.800000	0.000000	8.000000	1.000000	1.000000	1.000000	0.0
25%	6.000000	19.950000	153.950000	0.000000	38.000000	1.000000	96.000000	5.000000	0.0
50%	23.000000	22.950000	587.700000	0.000000	56.000000	2.000000	197.000000	10.000000	0.0
75%	47.000000	35.800000	1267.050000	0.000000	71.000000	3.000000	298.000000	15.000000	0.0
max	72.000000	57.150000	3264.450000	1.000000	100.000000	4.000000	399.000000	19.000000	0.0

Figure 38:description of the cluster0

4. Predictive analytics

A. internal data

KNN

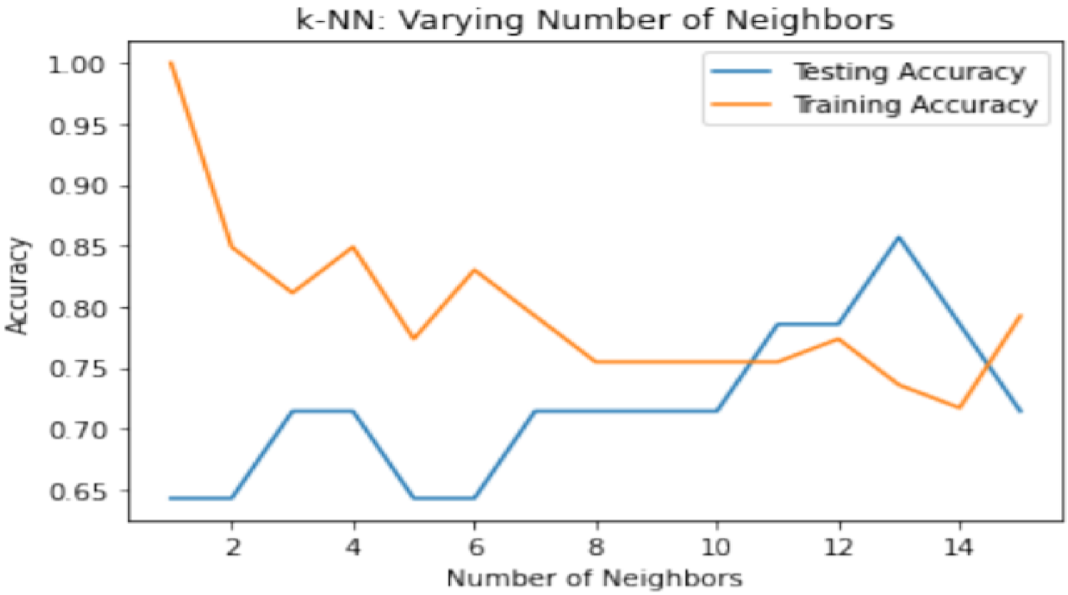
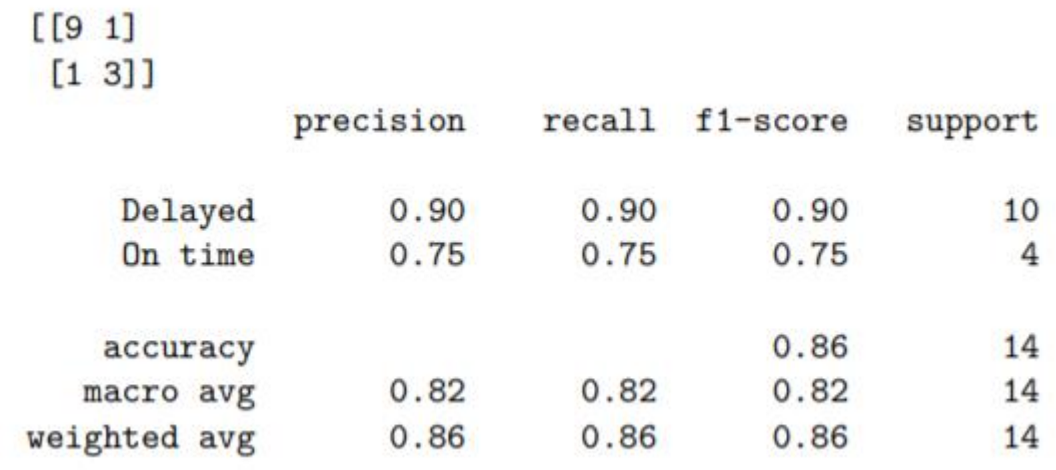


Figure 39:KNN varying number of neighbors

In the beginning, when the n_neighbors were 1, 2, or 3, training accuracy was a lot higher than test accuracy. So, the model was suffering from high overfitting.

After that training and test accuracy became closer. That is the sweet spot. We want that.

But when n_neighbors was going even higher, both training and test set accuracy was going down. We do not need that. From the graph above, the perfect n_neighbors for this particular dataset and model should be 11



```
[[9 1]
 [1 3]]
              precision    recall  f1-score   support

     Delayed       0.90      0.90      0.90        10
     On time       0.75      0.75      0.75         4

    accuracy                           0.86        14
   macro avg       0.82      0.82      0.82        14
weighted avg       0.86      0.86      0.86        14
```

In our prediction, we can see that the accuracy of the KNN model is equal to **0.86.**

Decision Tree

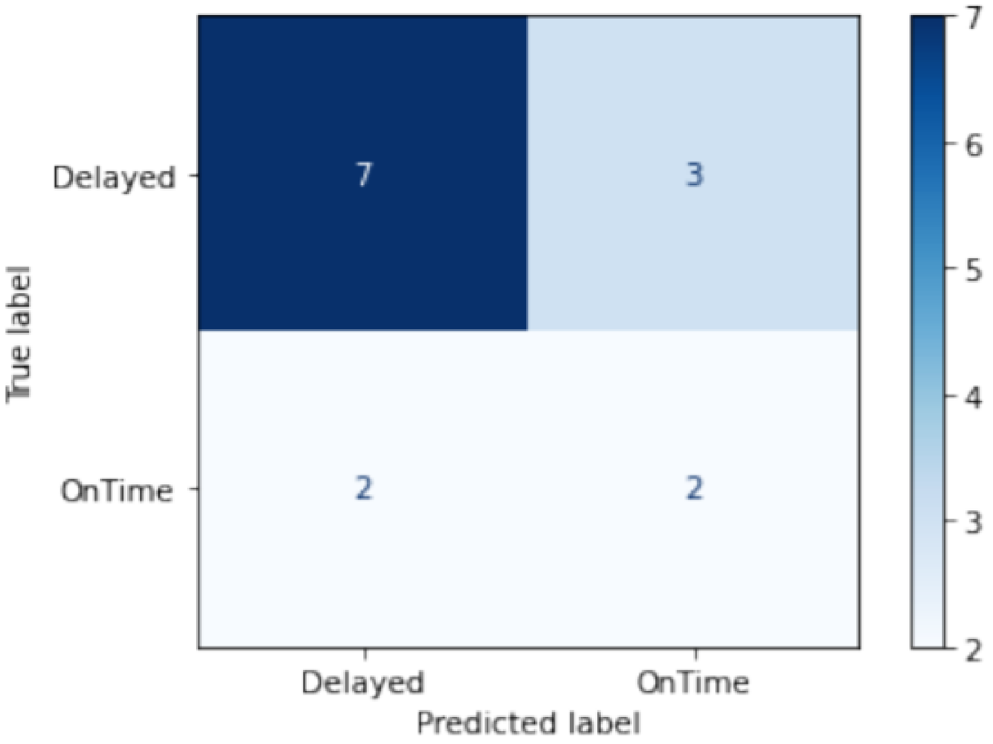
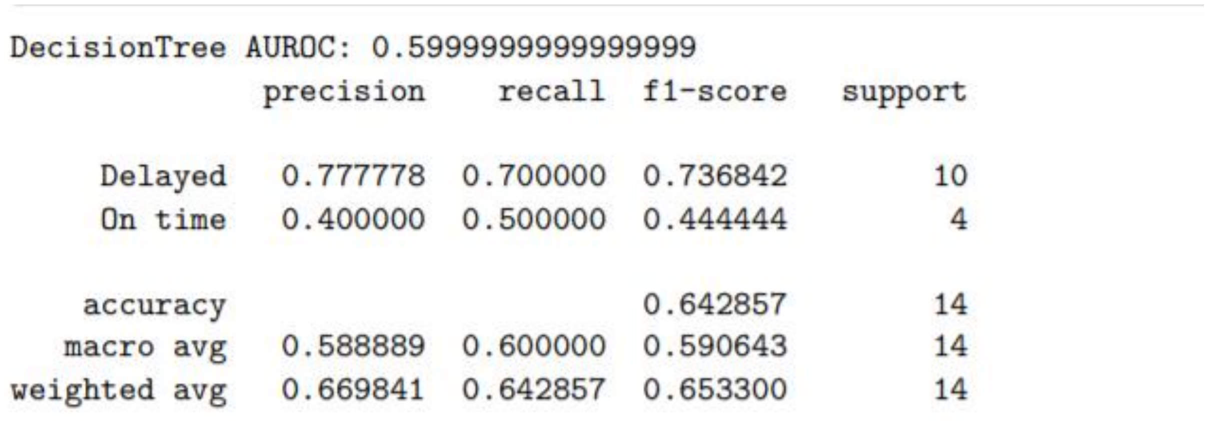


Figure 40:confusion matrix of Decision tree model



```
DecisionTree AUROC: 0.5999999999999999
              precision    recall  f1-score   support

     Delayed   0.777778  0.700000  0.736842        10
     On time   0.400000  0.500000  0.444444         4

    accuracy                       0.642857        14
   macro avg   0.588889  0.600000  0.590643        14
weighted avg   0.669841  0.642857  0.653300        14
```

In our prediction, we can see that the accuracy of the Decision Tree model is equal to **0.64**.

logistic regression

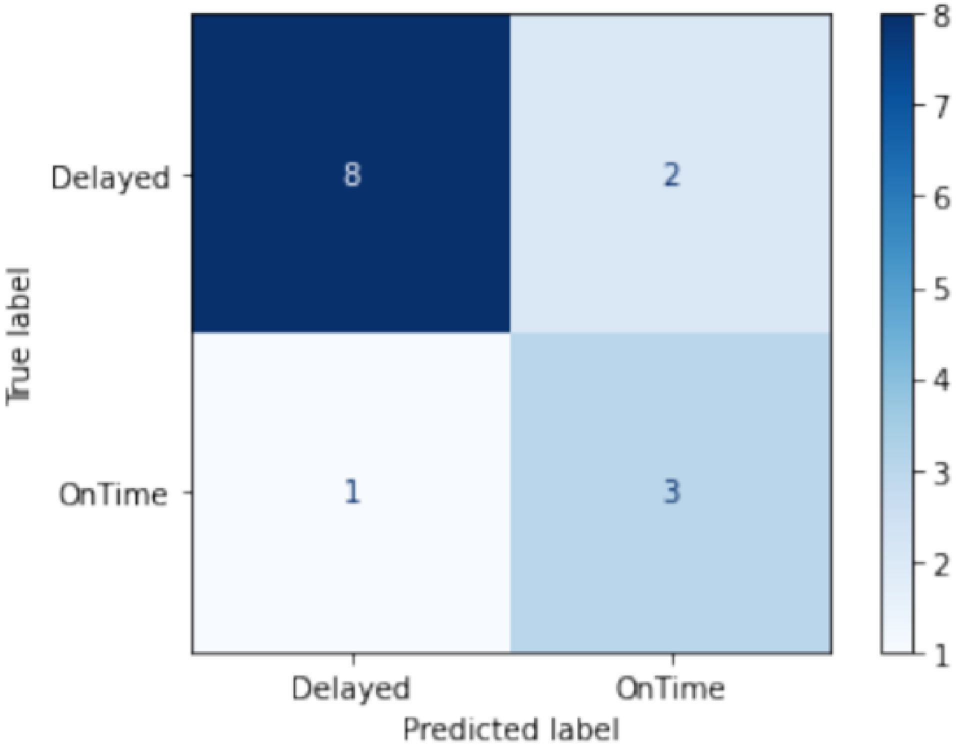
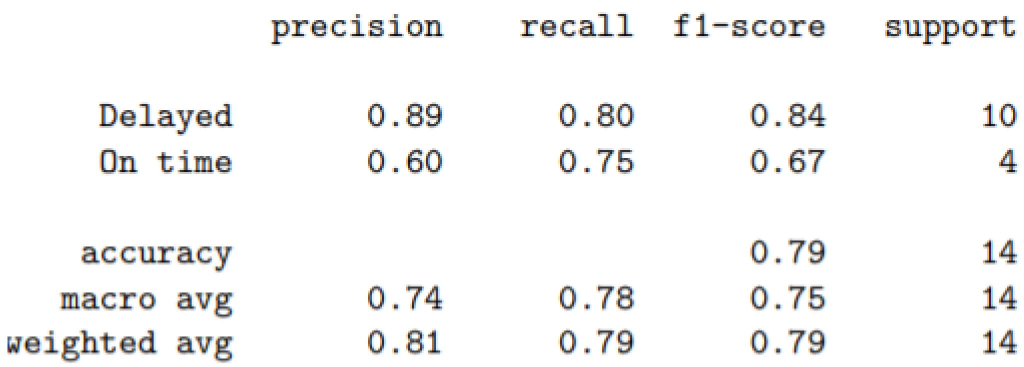


Figure 41:confusion matrix of logistic regression



```
              precision    recall  f1-score   support

     Delayed       0.89      0.80      0.84        10
     On time       0.60      0.75      0.67         4

    accuracy                           0.79        14
   macro avg       0.74      0.78      0.75        14
weighted avg       0.81      0.79      0.79        14
```

In our prediction, we can see that the accuracy of the logistic regression model is equal to **0.79.**

Classification of the models :

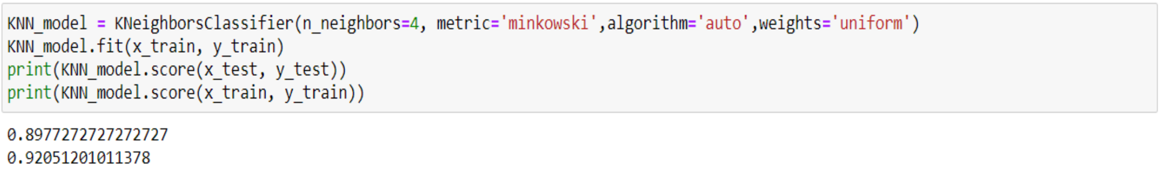
- ✓ KNN:86%
- ✓ Logistic Regression: 79%
- ✓ Decision Tree: 64%

B. external data

A churn model is a statistical representation of how the enterprise is influenced by churn. Calculations of churn are based on current data (the number of customers who left 45

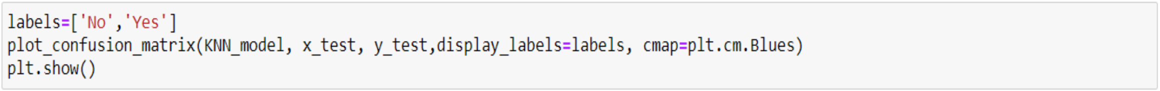
your service during a given time period). On this information, a predictive churn model extrapolates to show future potential churn rates. Prediction models of consumer churn aim to identify consumers with a high propensity for churn. Three main aspects of a churn prediction model are predictive precision, comprehensibility, and justifiability.

KNN



```
KNN_model = KNeighborsClassifier(n_neighbors=4, metric='minkowski',algorithm='auto',weights='uniform')
KNN_model.fit(x_train, y_train)
print(KNN_model.score(x_test, y_test))
print(KNN_model.score(x_train, y_train))
```

```
0.8977272727272727
0.92051201011378
```



```
labels=['No','Yes']
plot_confusion_matrix(KNN_model, x_test, y_test,display_labels=labels, cmap=plt.cm.Blues)
plt.show()
```

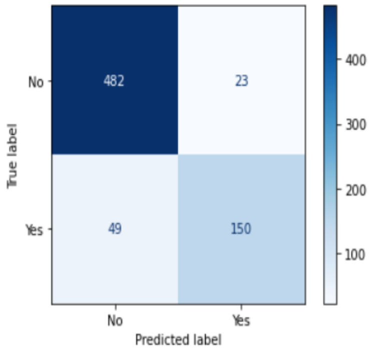
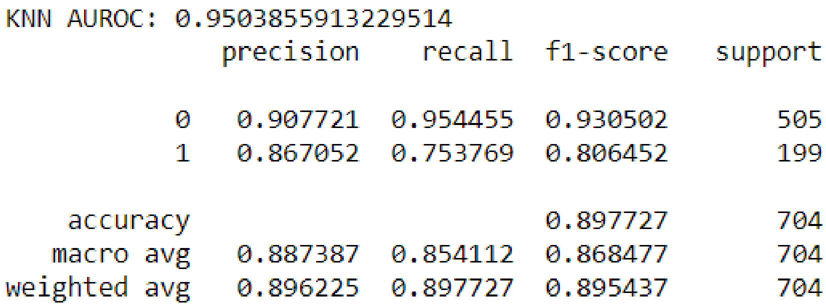


Figure 42:confusion matrix of KNN



```
KNN AUROC: 0.9503855913229514
              precision    recall  f1-score   support

           0   0.907721  0.954455  0.930502       505
           1   0.867052  0.753769  0.806452       199

    accuracy                       0.897727       704
   macro avg   0.887387  0.854112  0.868477       704
weighted avg   0.896225  0.897727  0.895437       704
```

Starting with the NO row of the confusion matrix, we can see that 482 customers did not churn and were correctly predicted not to have churned. Only 23 customers who did not churn and predicted to have churned. In the second row , we have 49 people who did churn and predicted not to have churned. And finally , 150 customers churned and correctly predicted to have churned.

In our prediction, we can see that the accuracy of the KNN model is equal to **0.89.**

Decision Tree

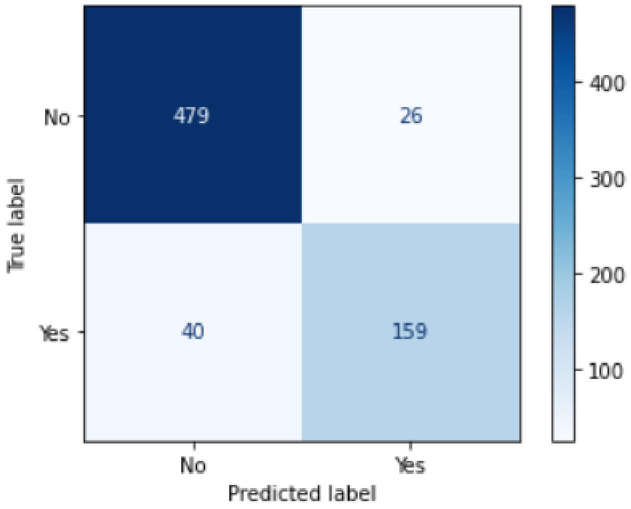
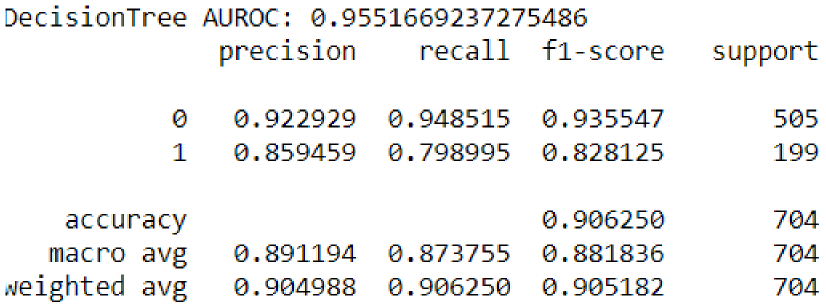


Figure 43:confusion matrix of Decision tree model



```
DecisionTree AUROC: 0.9551669237275486
              precision    recall  f1-score   support

           0   0.922929  0.948515  0.935547       505
           1   0.859459  0.798995  0.828125       199

    accuracy                       0.906250       704
   macro avg   0.891194  0.873755  0.881836       704
weighted avg   0.904988  0.906250  0.905182       704
```

Starting with the NO row of the confusion matrix, we can see that 479 customers did not churn and were correctly predicted not to have churned. Only 26 customers who did not churn and predicted to have churned. In the second row , we have 40 people who did churn and predicted not to have churned. And finally , 159 customers churned and correctly predicted to have churned.

In our prediction, we can see that the accuracy of the Decision TREE model is equal to **0.91**.

Logistic Regression

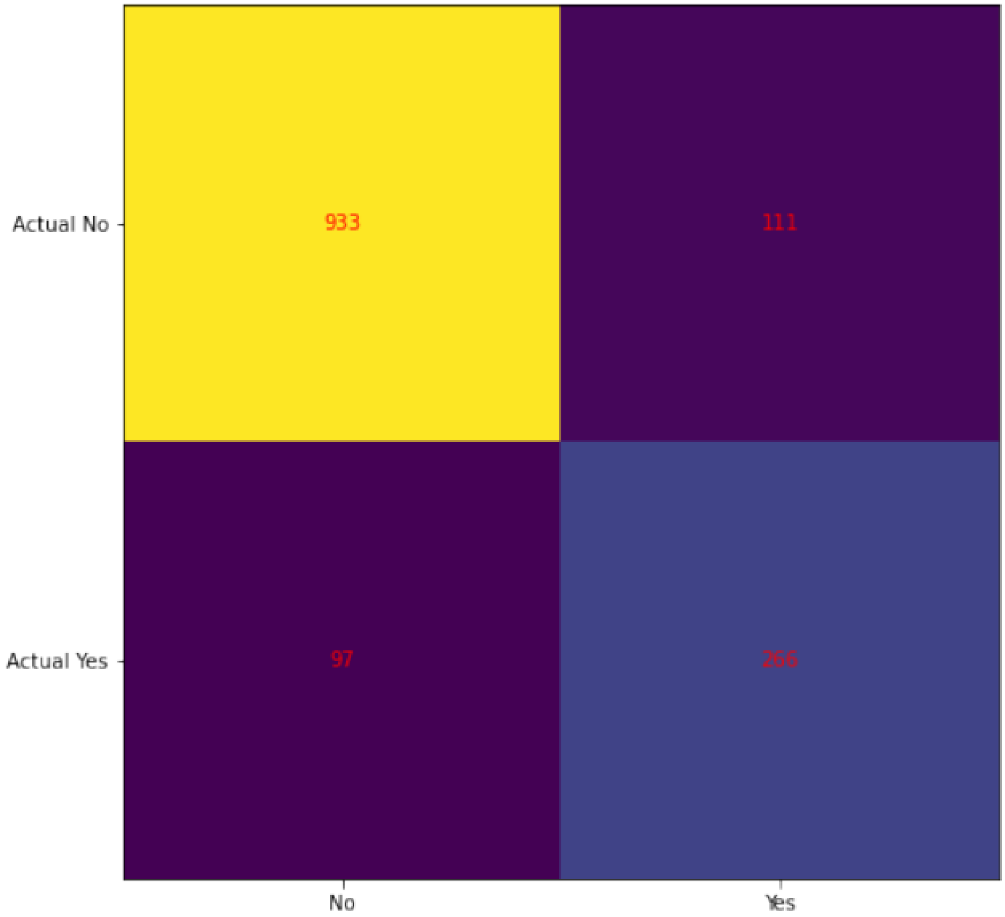
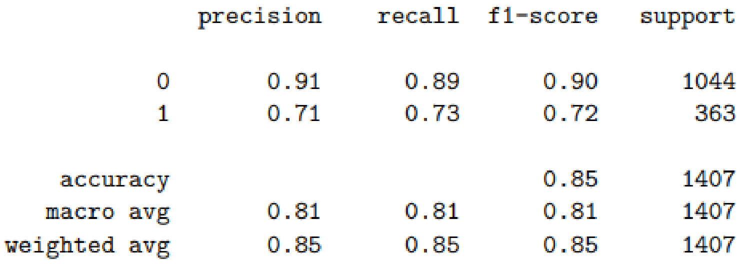


Figure 44:confusion matrix of logistic regression



```
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      1044
           1       0.71      0.73      0.72       363

    accuracy                           0.85      1407
   macro avg       0.81      0.81      0.81      1407
weighted avg       0.85      0.85      0.85      1407
```

Starting with the NO row of the confusion matrix, we can see that 933 customers did not churn and were correctly predicted not to have churned. Only 111 customers who did not churn and predicted to have churned. In the second row , we have 97 people who did churn and predicted not to have churned. And finally , 266 customers churned and correctly predicted to have churned.

In our prediction, we can see that the accuracy of the logistic regression model is equal to **0.85.**

Classification of the models :

- ✓ Decision Tree: 91%
- ✓ KNN: 89%
- ✓ Logistic Regression: 85%

5. Conclusion

By the end of this chapter we used different tools and technologies to succeed in both descriptive and predictive analytics wether utilizing internal or external data.

General Conclusion

These months of work have allowed us to situate ourselves in a professional context and to work on a large scale project as a cooperative team. our curriculum was particularly the technical perspective of the trainer.

We have strengthened our bases in power bi and talend with which we can connect , transform and prepare our data for processing later on and above all we discovered the new world of business intelligence and reporting.

In terms of implementation, we have met the objectives established by TALYS company despite all the challenges that we have encountered which mainly affect the data sources structures. One of the difficulties, which was the most arduous of them all, concerns the missing data, and as a result it took us a while to understand the data structures.

Webography

- https://www.talys-consulting.com/
- https://www.projeqtor.org/fr/
- https://cloudzone.io/the-great-data-warehousing-debate/