accenture

Politecnico
di Torino

1859

APPLIED DATA SCIENCE PROJECT

A.A 2021/2022

PPHI PROJECT

PRESENT BY: TENKAMTE KENGNE ARSÈNE BOLIVAR AND MAMMADLI FIDAN

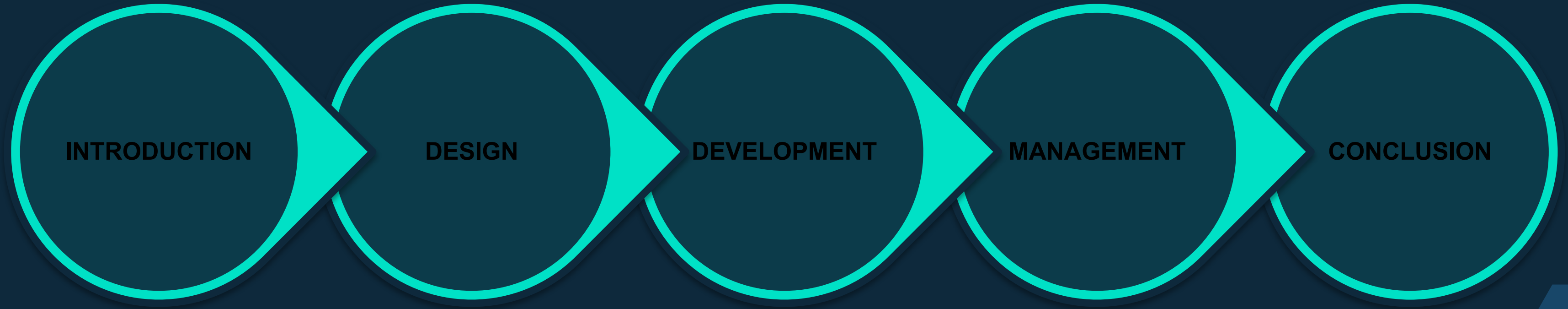
COMPAGNY SUPERVISORS:

- LUCA SCHIATTI
- ANH-DUNG LE
- GIUSEPPE GIORDANO

SUPERVISE BY:

GIUSSEPPE RIZZO

PLAN



INTRODUCTION

DESIGN

DEVELOPMENT

MANAGEMENT

CONCLUSION

1

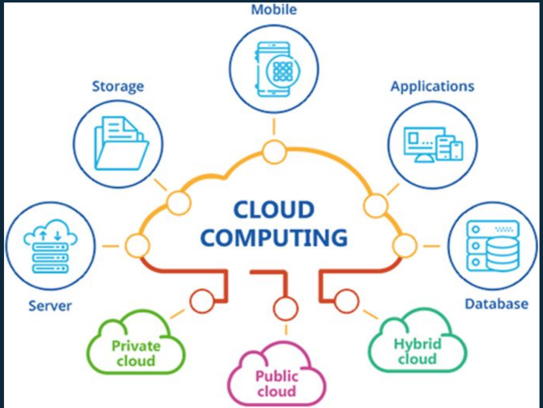
INTRODUCTION OF THE PROJECT

ML Model

predictive model to determine if their Health insurance policyholders would be also interested in a Vehicle Insurance

TEE

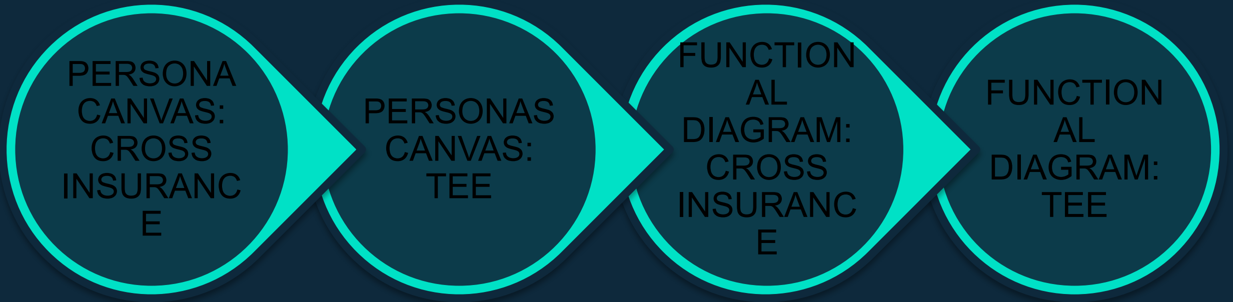
A creation of a secure area that guarantees optimal protection for highly sensitive data in all states, with respect to the confidentiality and integrity



2

DESIGN

Presentation of the general aspect with also the ongoing scenerio of the different point of the projet



PERSONA CANVAS: CROSS INSURANCE

PERSONAS CANVAS: TEE

FUNCTIONAL DIAGRAM: CROSS INSURANCE

FUNCTIONAL DIAGRAM: TEE

Persona Canvas: Cross Insurance Model

Frustrations	Needs
The company don’t know who want to apply to a vehicle insurance	The companies need to optimize it’s business model and revenue.
	A model to predict which customer will apply to the vehicle insurance

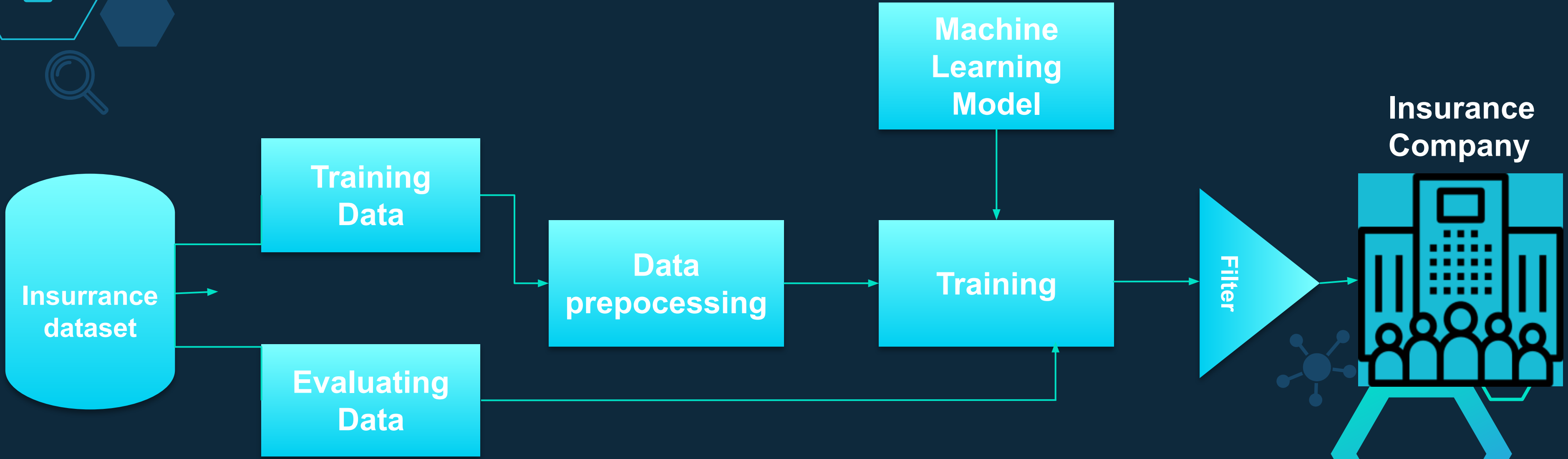
Insurance companies

Persona Canvas: TEE

Companies

Frustrations	Needs
The absence of confidence between the different companies involve on the project. The lack of privacy regarding the sharing of the data through the cloud.	Creation of a secure enclave environment. Storage of the dataset on the database
The lack of security throw the download and the processing of the model. The lack of privacy through the evaluation of the data.	Execution of the model in an enclave environment. Creation of a app that will be launch into the enclave

Functional Diagram: Cross Insurance Model



Insurrance dataset

Training Data

Evaluating Data

Data prepocessing

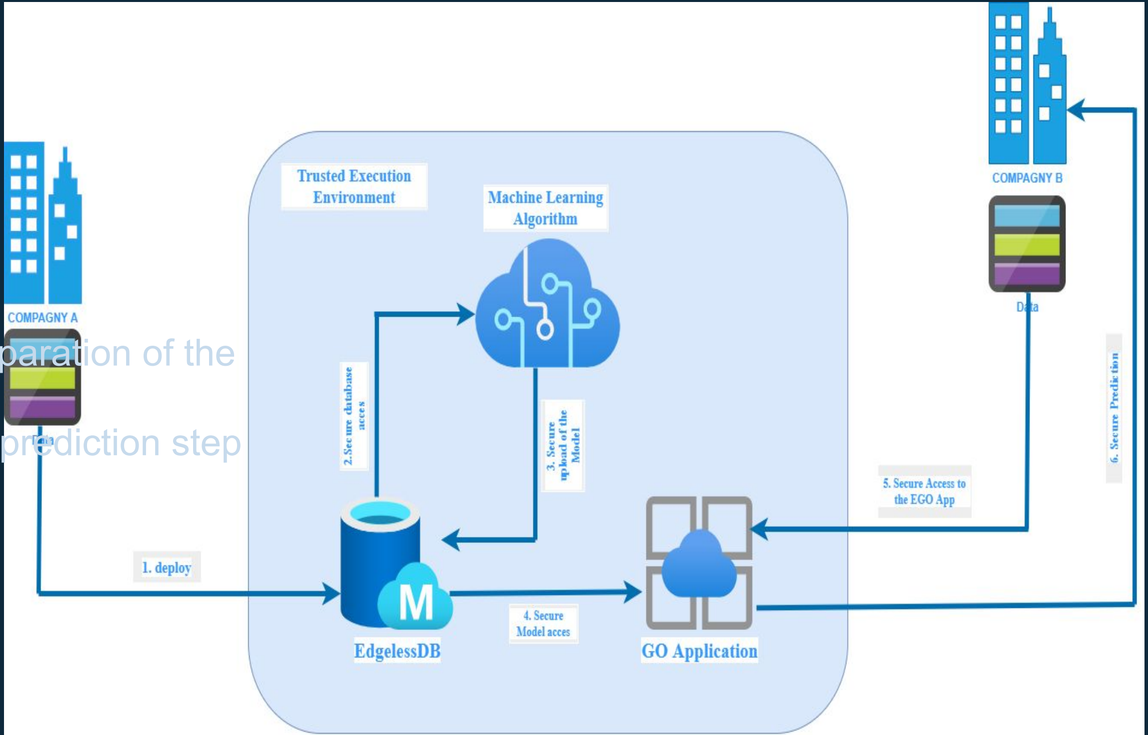
Machine Learning Model

Training

Filter

Insurance Company

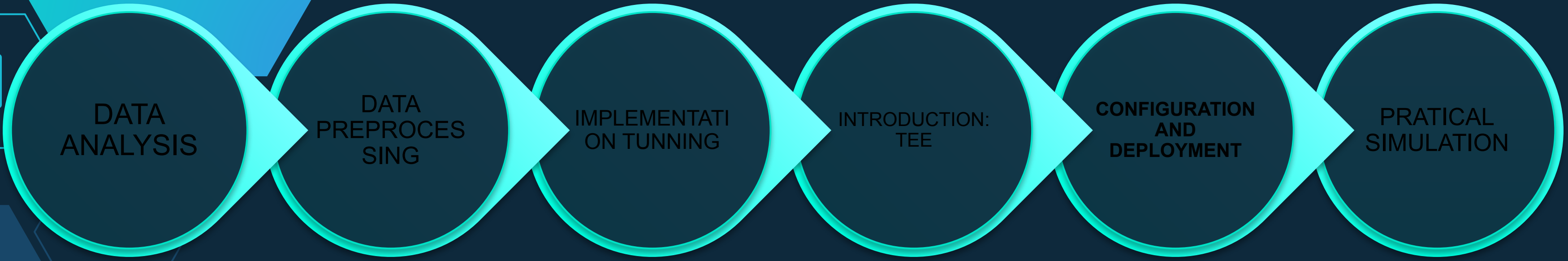
Functional Diagram: TEE



- **First step**: the preparation of the model
- **Second step:** the prediction step

Development

3



DATA ANALYSIS

DATA PREPROCESSING

IMPLEMENTATION TUNNING

INTRODUCTION: TEE

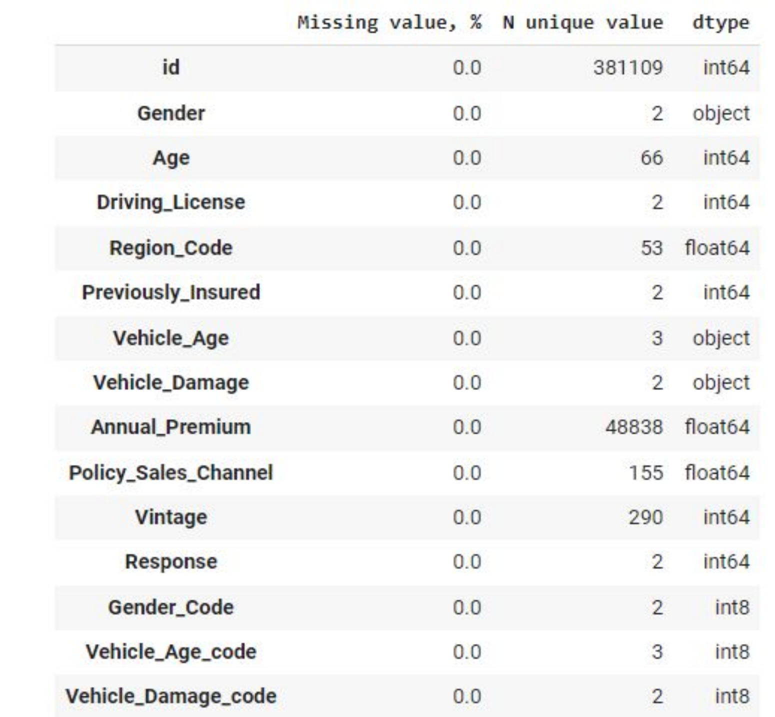
CONFIGURATION AND DEPLOYMENT

PRATICAL SIMULATION

Data Analysis

Numerical attribute

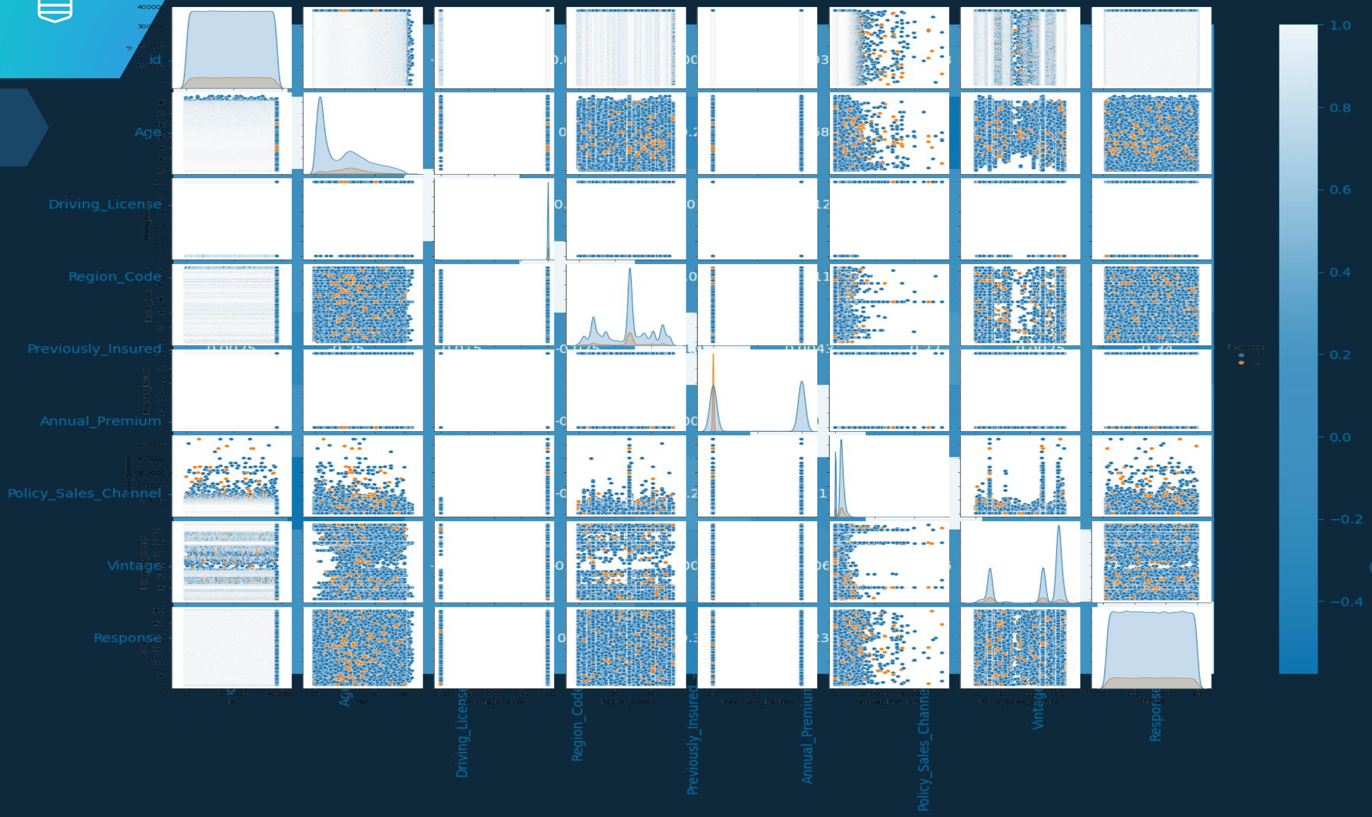
- No missing values on the different features
- Few amount of uniques



	Missing value, %	N unique value	dtype
id	0.0	381109	int64
Gender	0.0	2	object
Age	0.0	66	int64
Driving_License	0.0	2	int64
Region_Code	0.0	53	float64
Previously_Insured	0.0	2	int64
Vehicle_Age	0.0	3	object
Vehicle_Damage	0.0	2	object
Annual_Premium	0.0	48838	float64
Policy_Sales_Channel	0.0	155	float64
Vintage	0.0	290	int64
Response	0.0	2	int64
Gender_Code	0.0	2	int8
Vehicle_Age_code	0.0	3	int8
Vehicle_Damage_code	0.0	2	int8

Data Analysis

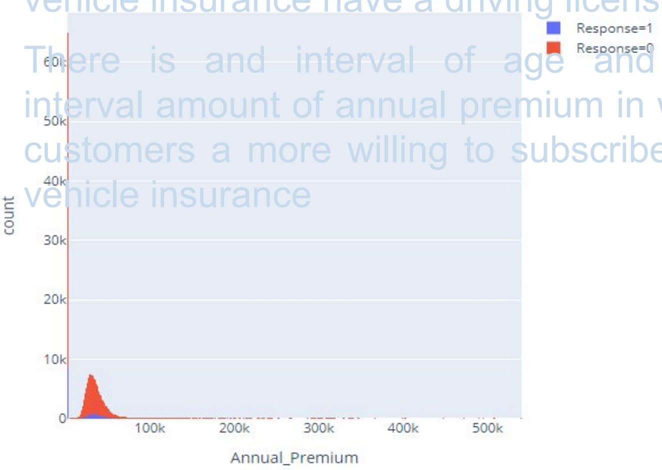
Numerical attribute

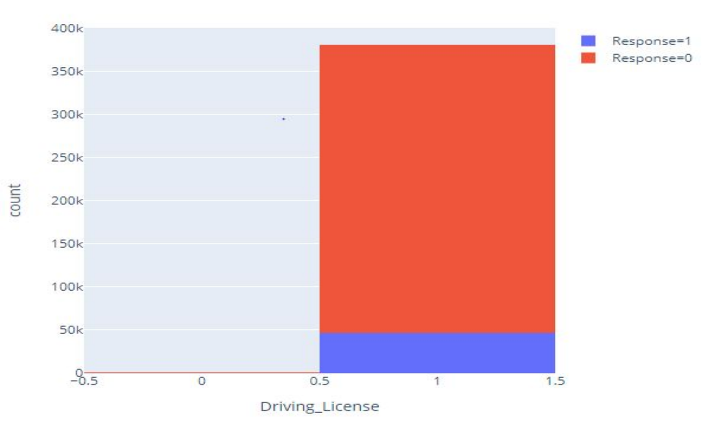


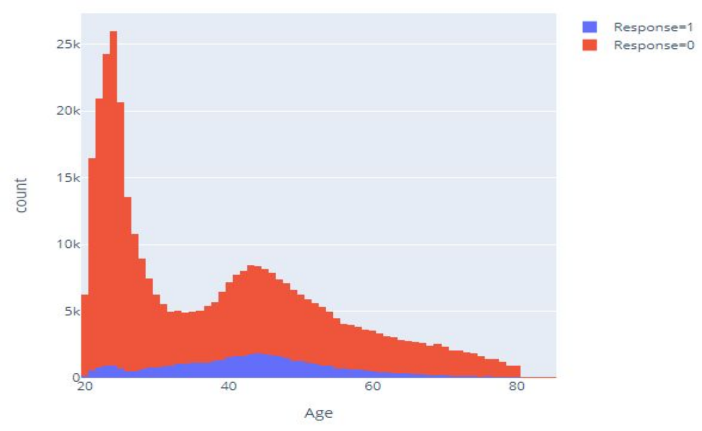
Data Analysis

Numerical attribute

- The majority of the customers that apply to vehicle insurance have a driving license
- There is and interval of age and also interval amount of annual premium in which customers a more willing to subscribe to a vehicle insurance

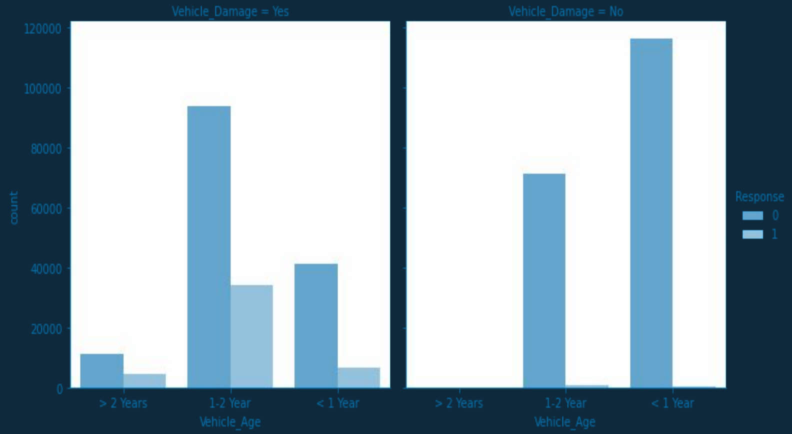


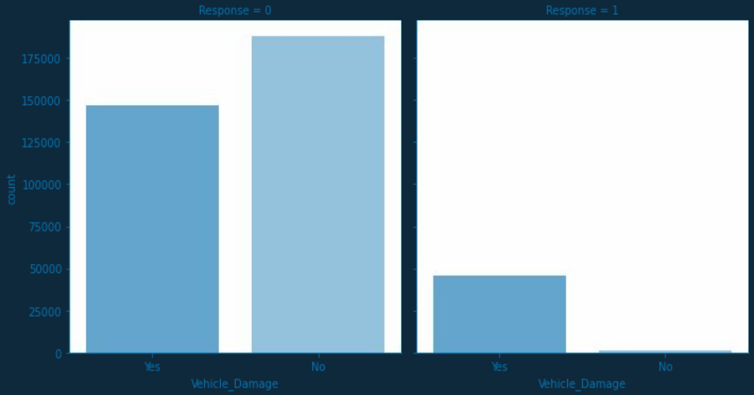


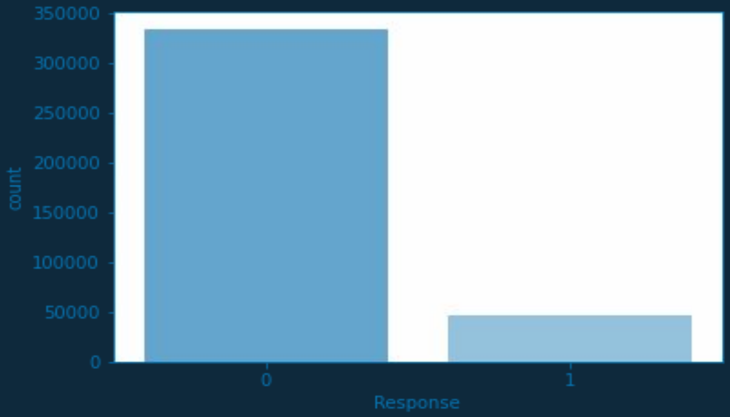


Data Analysis

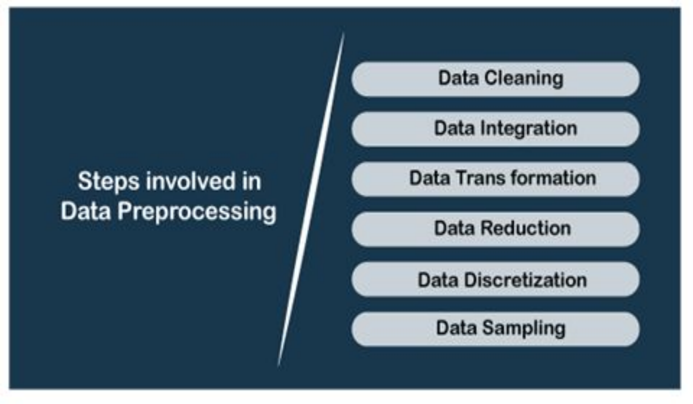
Categorical attribute







Data Preprocessing

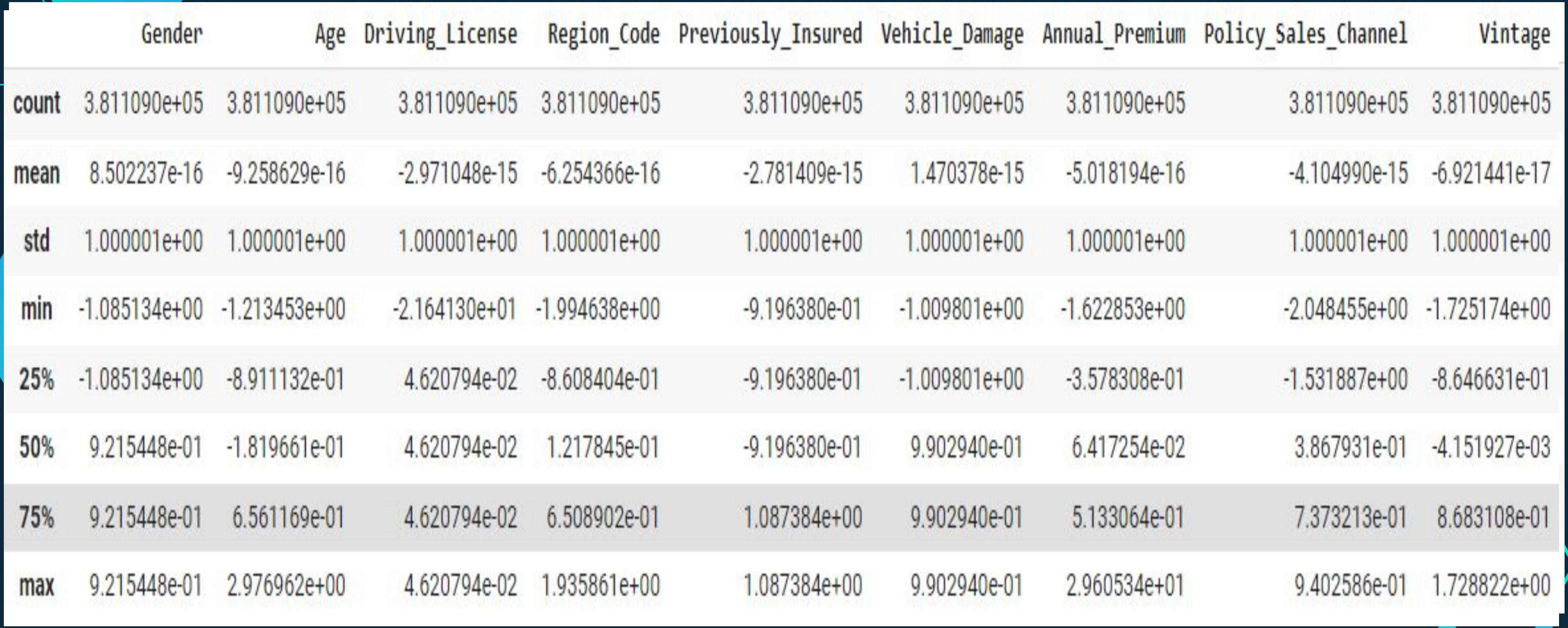


Steps involved in Data Preprocessing

- Data Cleaning
- Data Integration
- Data Trans formation
- Data Reduction
- Data Discretization
- Data Sampling

- **Dropping** of the ID feature
- Use of the **One-hot encoding** on the feature Vehicule_Age
- **Label Encoding**: with the **OrdinalEncoder** for the feature Vehicle_name and Gender
- **Spliting** of the dataset in trainning and test data
  - Unbalanced data
  - Balanced data (SMOTE)
- Four data files to utilize in the construction and the analysis of our models

Normalisation

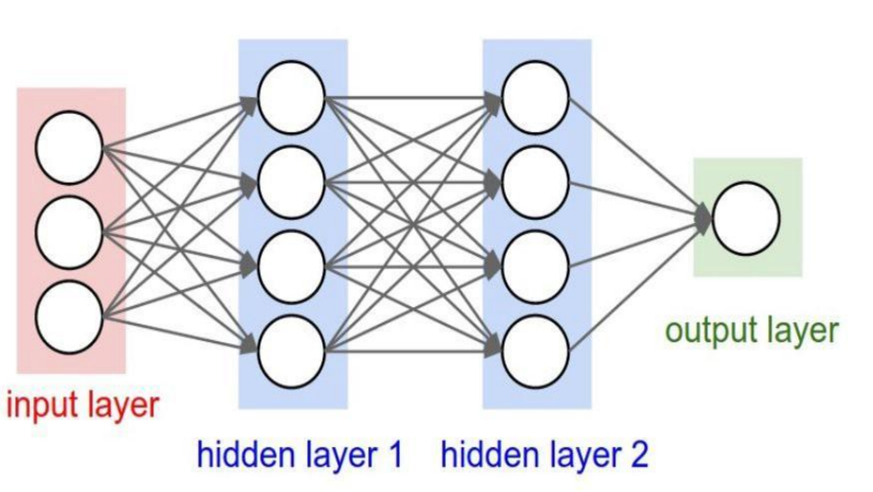


	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage
count	3.811090e+05	3.811090e+05	3.811090e+05	3.811090e+05	3.811090e+05	3.811090e+05	3.811090e+05	3.811090e+05	3.811090e+05
mean	8.502237e-16	-9.258629e-16	-2.971048e-15	-6.254366e-16	-2.781409e-15	1.470378e-15	-5.018194e-16	-4.104990e-15	-6.921441e-17
std	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00
min	-1.085134e+00	-1.213453e+00	-2.164130e+01	-1.994638e+00	-9.196380e-01	-1.009801e+00	-1.622853e+00	-2.048455e+00	-1.725174e+00
25%	-1.085134e+00	-8.911132e-01	4.620794e-02	-8.608404e-01	-9.196380e-01	-1.009801e+00	-3.578308e-01	-1.531887e+00	-8.646631e-01
50%	9.215448e-01	-1.819661e-01	4.620794e-02	1.217845e-01	-9.196380e-01	9.902940e-01	6.417254e-02	3.867931e-01	-4.151927e-03
75%	9.215448e-01	6.561169e-01	4.620794e-02	6.508902e-01	1.087384e+00	9.902940e-01	5.133064e-01	7.373213e-01	8.683108e-01
max	9.215448e-01	2.976962e+00	4.620794e-02	1.935861e+00	1.087384e+00	9.902940e-01	2.960534e+01	9.402586e-01	1.728822e+00

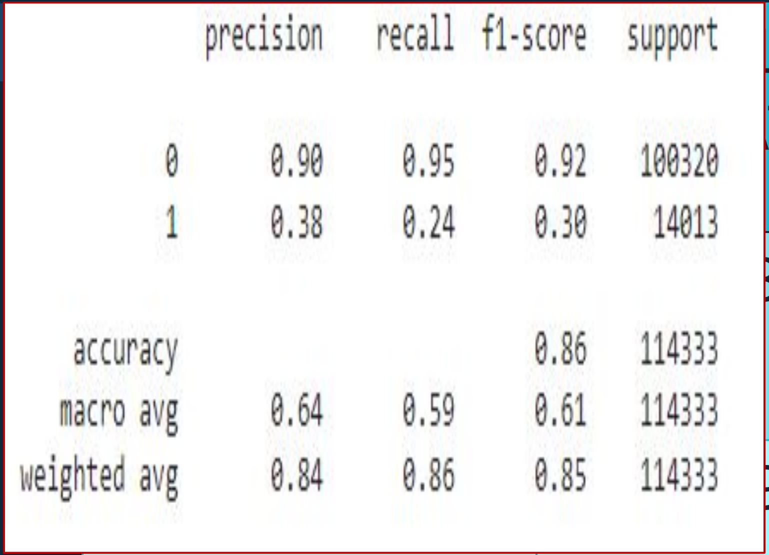
Before the normalisation

After the normalisation

- Decision tree
- Catboost
- Random Forest
- Keras Sequential Model
- Multilayer Perceptron



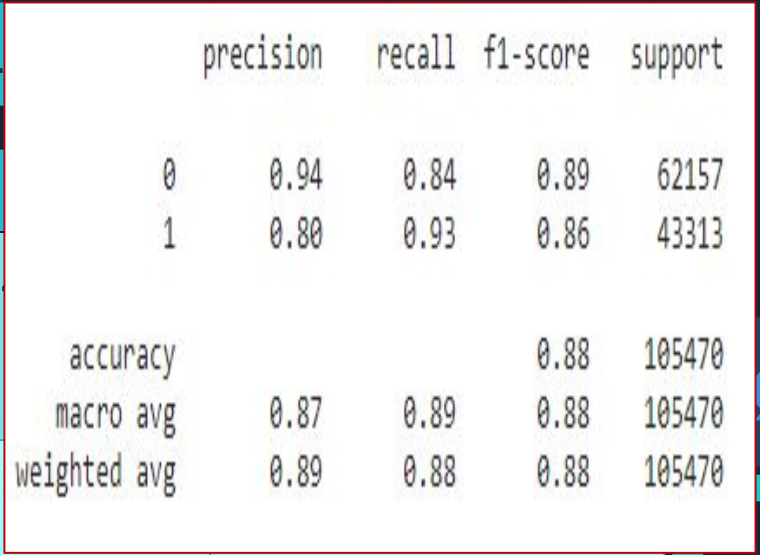
Implementation Tuning



```
              precision    recall  f1-score   support

           0       0.90      0.95      0.92    100320
           1       0.38      0.24      0.30     14013

    accuracy                           0.86    114333
   macro avg       0.64      0.59      0.61    114333
weighted avg       0.84      0.86      0.85    114333
```



```
              precision    recall  f1-score   support

           0       0.94      0.84      0.89     62157
           1       0.80      0.93      0.86     43313

    accuracy                           0.88    105470
   macro avg       0.87      0.89      0.88    105470
weighted avg       0.89      0.88      0.88    105470
```

To find the best parameters of the different models we run a GridsearchCV with a number of cv=3 and which give us the opportunity to bypass the training and check the score on the testing data (validation data)

Implementation Tuning

Random Forest	*max_features* = ['auto'] *criterion* = ['mse'] *bootstrap* = [False] *n_estimators* = [200]	**0.9771245228843219**
CatBoost	*n_estimators = [300]* *depth = [5]* *learning_rate =[0.01]* *border_count =[50]* *ctr_border_count=[100]* *thread_count = 4*	**0.981938**
Keras Sequential model	*epoch=[100]* *InputLayer=[64,32,16,8]*	**0.9575998937222543**
Multilayer Perceptron	*Hidden_layer_sizes = (50,50,50)* *Activation = ['relu']* *Solver =['adam']* *Alpha = [0.0001]* *learning_rate = ['adaptive']*	**0.9634193509552655**

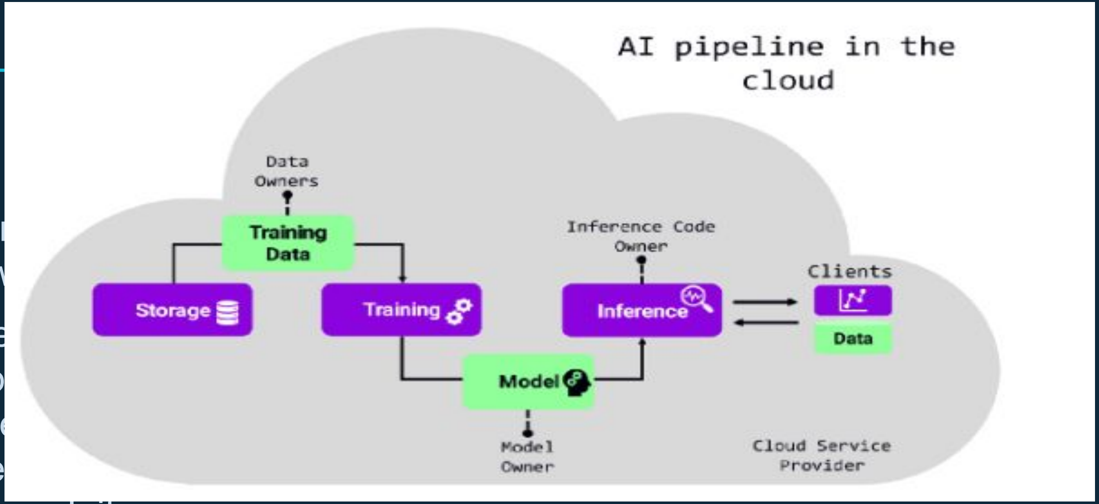
n_estimators = [100, 200, 500, 1000]
n_estimators = [100, 200, 500, 1000]
depth = [illegible],6,4,5,7,8,9,10]
learning_rate =[0.03,0.001,0.01,0.1,0.2,0.3]
border_count=[32,5,10,20,50,100,200]
ctr_border_count=[50,5,10,20,100,200]
thread_count = 4

Activation = ['relu']
Solver =['adam']
Alpha = [0.0001]
learning_rate = ['adaptive']

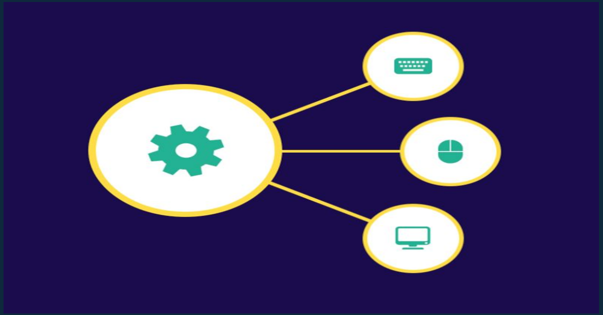
Introduction to the TEE

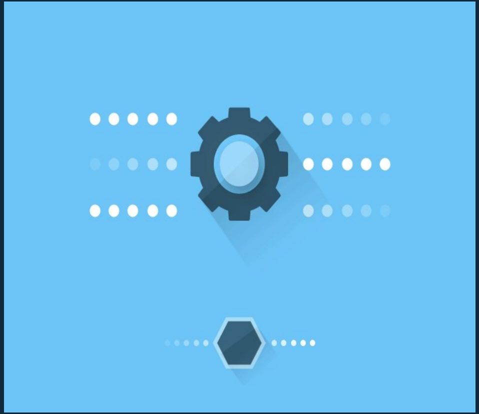
W

- Ther betw
- Ente
- cryp prote type and exploits.



Configuration





FSGSBASE

which is a feature in recent processors which allows direct access to the FS and GS segment base addresses

Intel SGX driver

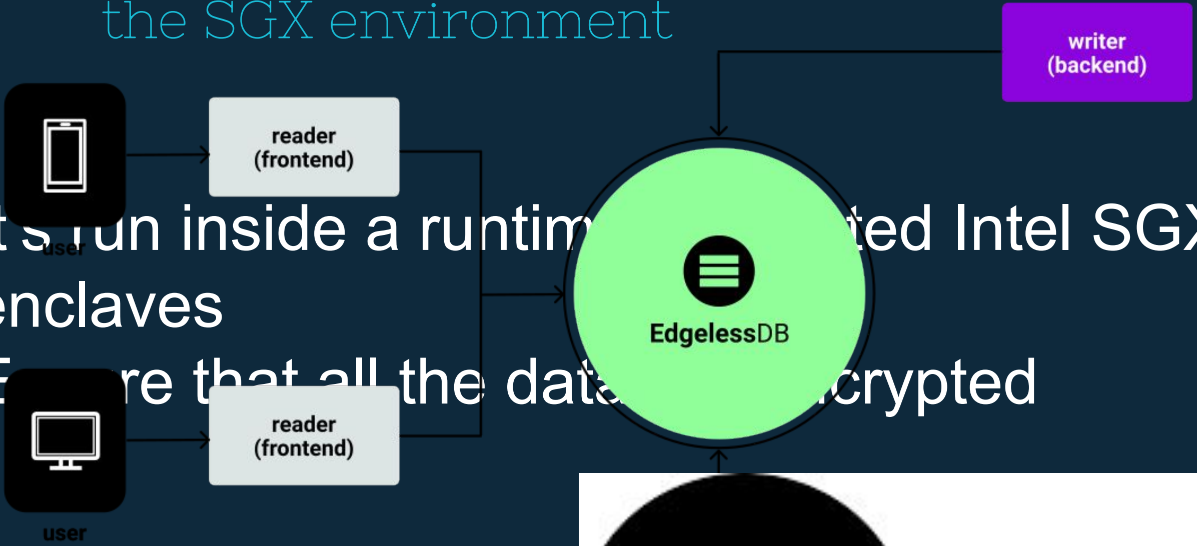
To improve data protection and enhance application code security

EdgelessRT

An SDK for Trusted Execution Environments (TEE)

EdgelessDB

EdgelessDB is a SQL database architected for the SGX environment



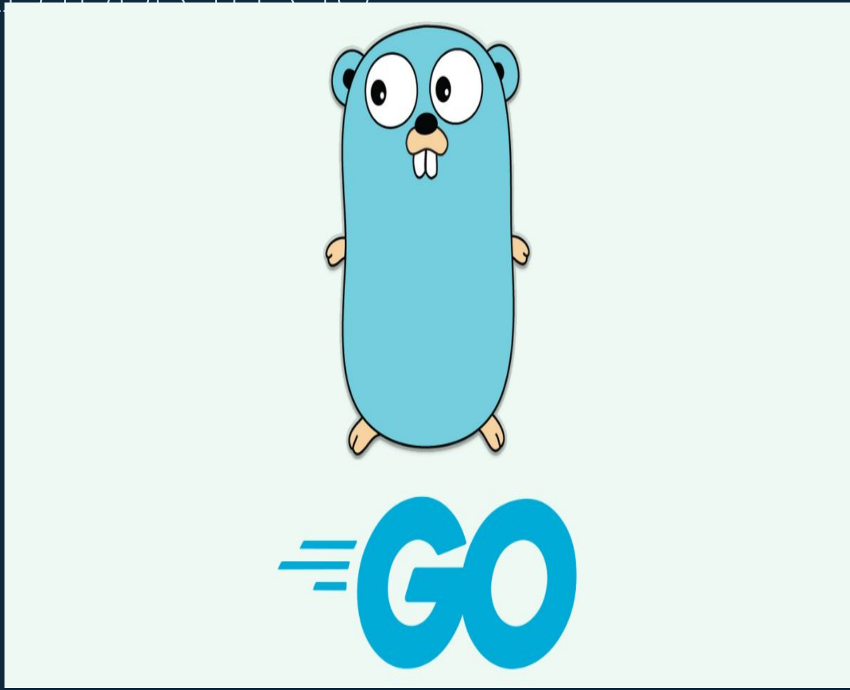
- It s run inside a runtim ted Intel SGX enclaves
- E re that all the data crypted

EdgelessDB

EGO

Ego is a framework for building confidential apps in Go

- Go apps always-encrypted run in encrypted enclaves on Intel SGX-enabled hardware
- It’s simplified enclave deployment
  - Ego-go
  - Ego



Practical simulation

Run the system

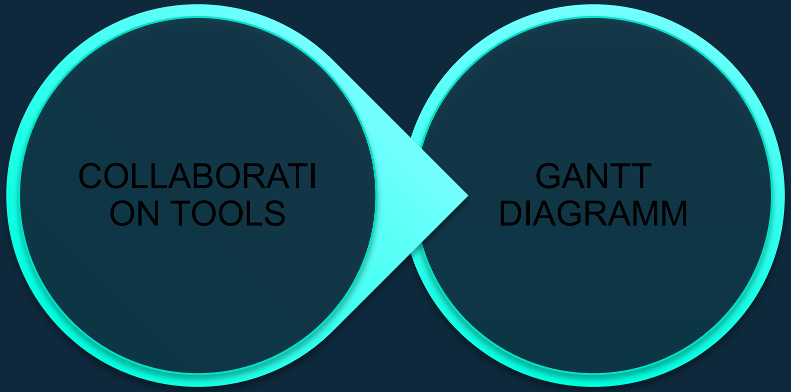
Test the Database

Test the model

4

MANAGEMENT

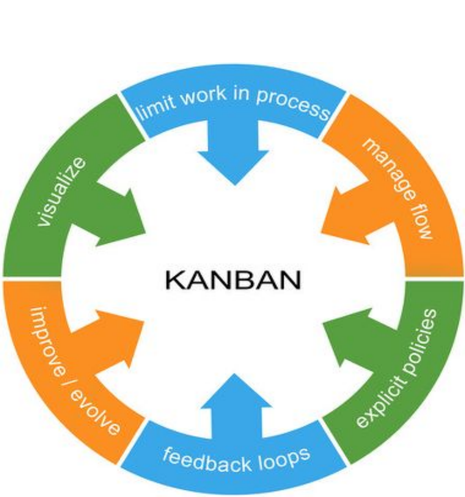
How do we solve the project ? Which has been utilize ?



COLLABORATION TOOLS

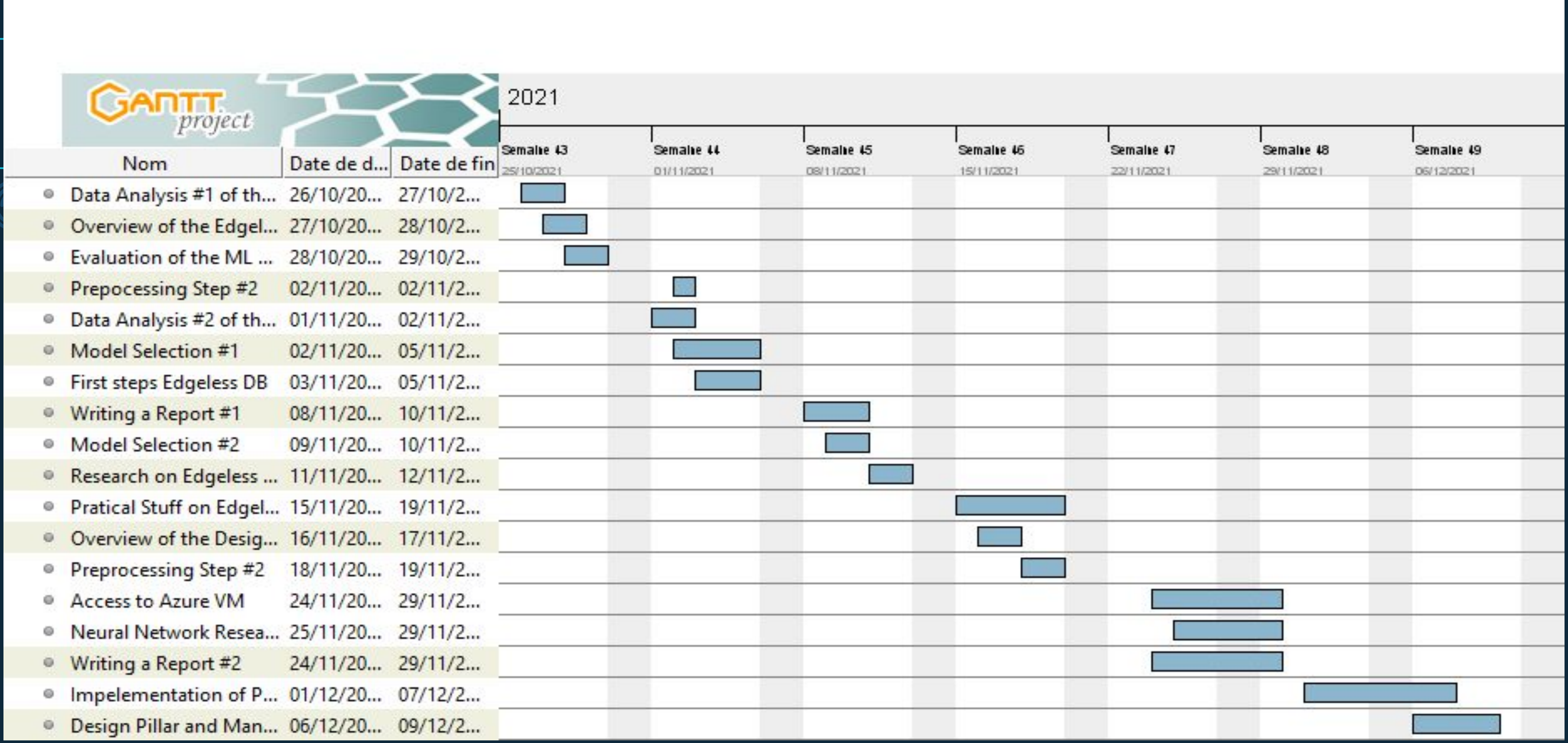
GANTT DIAGRAMM

COLLABORATION TOOLS



- **Kanbanflow** : for the follow up of the team work collaboration.
- **Scrum poker**: to estimate the complexity and effort of the different task.
- **Google Collab:** online cloud-based Jupyter notebook environment that allows us to train our machine learning and deep learning models on CPUs, GPUs, and TPUs.

GANTT DIAGRAMM



Nom	Date de d...	Date de fin
Data Analysis #1 of th...	26/10/20...	27/10/2...
Overview of the Edgel...	27/10/20...	28/10/2...
Evaluation of the ML ...	28/10/20...	29/10/2...
Prepocessing Step #2	02/11/20...	02/11/2...
Data Analysis #2 of th...	01/11/20...	02/11/2...
Model Selection #1	02/11/20...	05/11/2...
First steps Edgeless DB	03/11/20...	05/11/2...
Writing a Report #1	08/11/20...	10/11/2...
Model Selection #2	09/11/20...	10/11/2...
Research on Edgeless ...	11/11/20...	12/11/2...
Pratical Stuff on Edgel...	15/11/20...	19/11/2...
Overview of the Desig...	16/11/20...	17/11/2...
Preprocessing Step #2	18/11/20...	19/11/2...
Access to Azure VM	24/11/20...	29/11/2...
Neural Network Resea...	25/11/20...	29/11/2...
Writing a Report #2	24/11/20...	29/11/2...
Impelementation of P...	01/12/20...	07/12/2...
Design Pillar and Man...	06/12/20...	09/12/2...

1

CONC

Just one last thing to say

Thanks!

- For your time
- And for your attention

Any questions?