Système de Recommandation

Apprentissage chez Junglebike

Komlan Jean-Marie DANTODJI

Université Paris 8, LIASD
Encadrante : Mme Rakia JAZIRI
Tutrice : Mme Alice Battarel

31 mai 2022

Plan

JungleBike

- Start Up de B2C.
- Spécialisée dans le secteur du Vélo.
- Une entreprise de E-commerce dans la vente de matériels de vélos dans le but de faciliter la réparation.
- Elle est fondée en 2018, le site e-commerce a été mis en ligne en 2020.
- Plus d'une vingtaine de fournisseur aujourd'hui.

Solutions proposées

- Mise en ligne des matériels de vélos,
- Enregistrement du vélo permettant d'identifier le modèle et ses différentes pièces afin de faciliter la réparation,
- Mise en relation des clients avec les réparateurs.

Processus de mise en ligne des produits

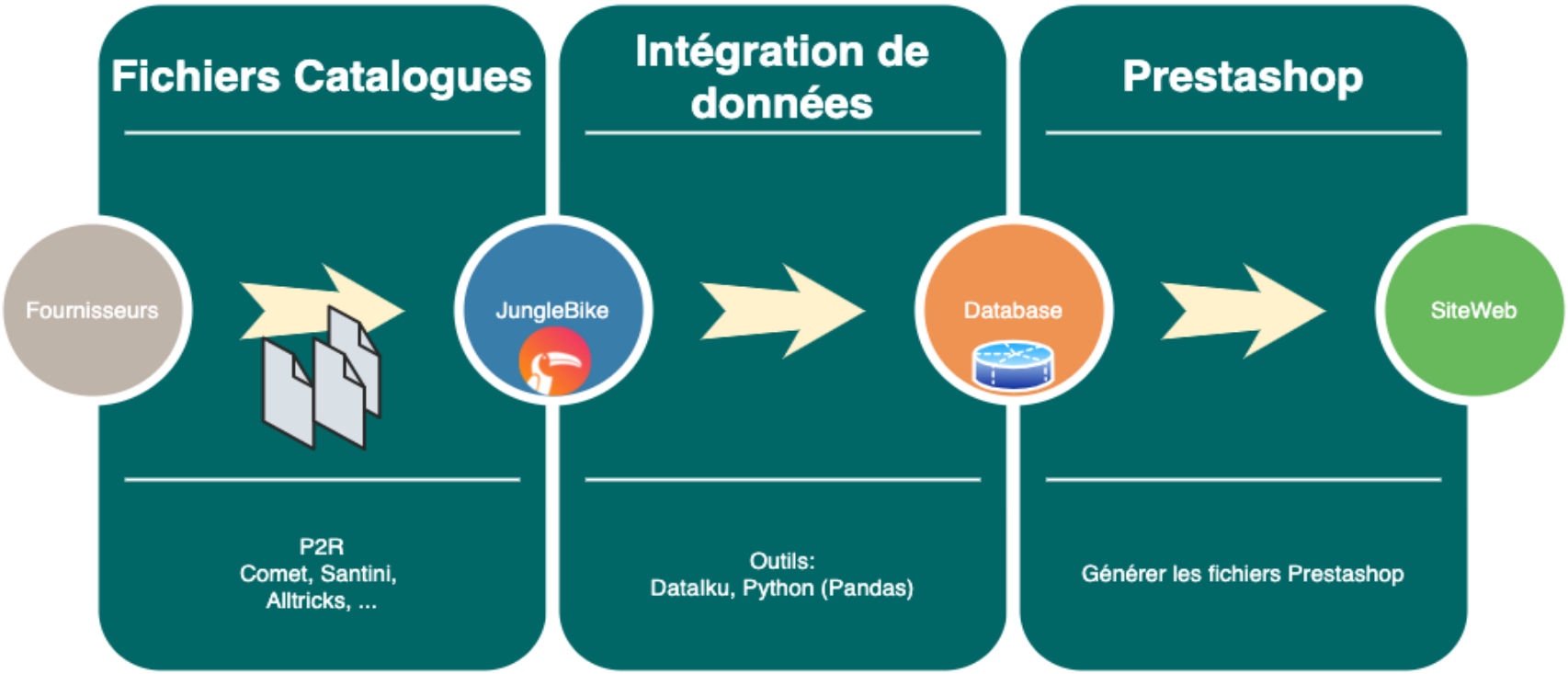


FIG. : Processus de mise en ligne

Contexte RH

- Equipe Data
- Formé de deux data scientistes
- Intégration de donnée
- Construction des algoritmes de catégorisation et d'extration de données.
- Construcition des modèles de recommandation et d'analyse de sentiments.

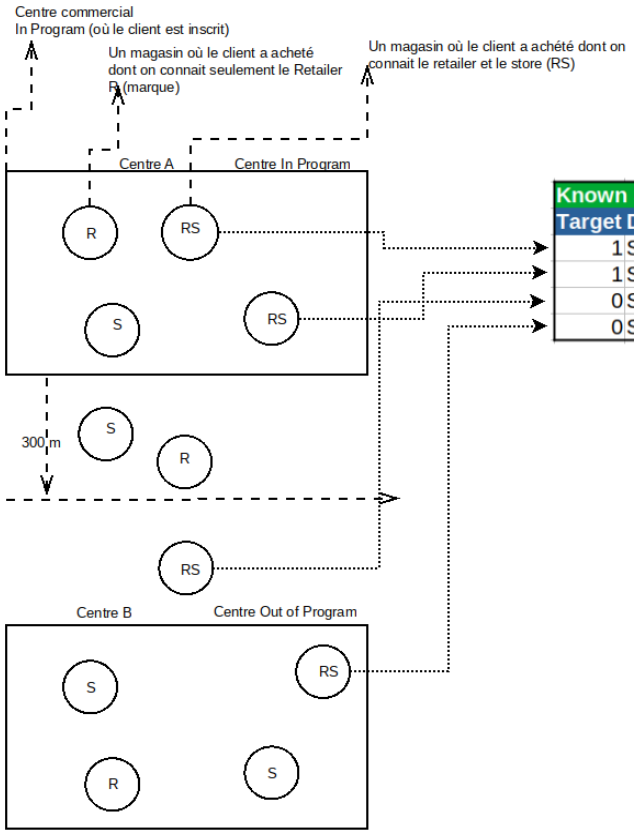
Contexte technique

- Outils : Datalku, DBeaver, .
- Langages et librairies : Python, Scikit Learn, Keras

Problème

- Recommandation des produits basée sur les avis des clients sur les produits.
- Analyse du sentiment des clients basé sur les commentaires des clients sur les produits.

Features Engineering



Known Stores and Retailers						Known Stores				Known Retailers		
Target	DayOfWeek	distance	distance_bin	Amount	Cannib	nb_transac	PM	SM	PNM	PC	PPM	PPNM	PPC
1	Samedi	5.1	(0,10]	10	0.166	10	1	0	1	4	1	1	3
1	Samedi	5.1	(0,10]	40	0.666	10	1	0	1	4	1	1	3
0	Samedi	5.1	(0,10]	5	0.083	10	0	0	0	4	0	0	3
0	Samedi	5.1	(0,10]	5	0.083	10	2	0	0	4	1	0	3

Pour les transactions identifiées (associées à un store et à un retailer)
Target:
Transaction In Program ou Out of Program
dayOfWeek:
jour de transaction extrait de la date de transaction
distance:
distance en (km) entre le centre commercial où le client est inscrit au store le plus proche en déhors de son centre
distance_bin:
classe de distance qu'appartient la distance précédente (a,b]
Amount:
montant dépensé dans le store
Cannib = cannibalisation:
rapport entre le montant dépensé chez le store sur le montant total dépensé dans la journée
nb_transac:
Nombre de transaction effectuées
PM = Purchases_InMall:
Nombre de store dans le meme centre
SM = Services_InMall:
Nombre de services utilisés (amazon locker, parking) à moins de 1km que le centre visité
PNM = Purchases_NextToTheMall:
Nombre de stores visités en déhors du centre et à moins de 300m
PC = Purchases_InCity:
Nombre de store visités dans la même ville que le centre visité
PPM = Purchases_Potential_InMall:
Nombre de store de tous les retailers inconnu dans le meme centre
PPNM = Purchases_Potential_NextToTheMall:
Nombre de stores de tous les retailers visités en déhors du centre et à moins de 300m
PPC = Purchases_Potential_InCity:
Nombre de store de tous les retailers visités dans la même ville que le centre visité

Jeu de données

	target	DayOfWeek	amount	distance	distance_bin	nb_transac	Purchases_InMall	Purchases_NextToTheMall	Services_InMall	Purchases_InCity	Purch
0	1	Monday	39.90	294.316896	(200.0, 500.0]	4	1	1	0	0	
1	1	Monday	64.94	5.605244	(5.0, 10.0]	4	1	1	0	1	
2	1	Monday	35.98	294.114634	(200.0, 500.0]	2	1	1	0	1	
3	1	Monday	75.90	5.458404	(5.0, 10.0]	2	1	1	0	1	
4	1	Saturday	118.96	5.627760	(5.0, 10.0]	6	4	4	0	4	
5	1	Saturday	22.90	5.433361	(5.0, 10.0]	6	4	5	0	4	
6	1	Saturday	10.00	294.203458	(200.0, 500.0]	6	4	4	0	3	
7	1	Saturday	29.00	5.600936	(5.0, 10.0]	6	4	4	0	3	
8	1	Saturday	16.90	5.458404	(5.0, 10.0]	6	4	4	0	3	
9	1	Saturday	30.00	294.206739	(200.0, 500.0]	5	3	2	0	2	
10	1	Saturday	13.98	473.669856	(200.0, 500.0]	5	3	2	0	3	
11	1	Saturday	13.98	294.252744	(200.0, 500.0]	5	3	2	0	2	
12	1	Saturday	18.00	294.247216	(200.0, 500.0]	5	3	3	0	2	
13	0	Saturday	19.00	5.620421	(5.0, 10.0]	5	4	3	0	3	
14	1	Thursday	39.98	294.152306	(200.0, 500.0]	3	2	2	0	2	

FIG. : Transactions considérées

Jeu de données

	Purchases_Potential_InMall	Purchases_Potential_InCity	Purchases_Potential_NextToTheMall	cannibalisation
)	0.0	0.0	0.0	0.593874
	0.0	0.0	0.0	0.418474
	0.0	0.0	0.0	0.484127
	0.0	0.0	0.0	0.376787
	0.0	0.0	0.0	0.070770
	0.0	0.0	0.0	0.355689
	0.0	0.0	0.0	0.569227
	0.0	0.0	0.0	0.222634
	0.0	0.0	0.0	0.375748
	0.0	0.0	0.0	0.379135
	0.0	0.0	0.0	0.661259
	0.0	0.0	0.0	0.252171
	0.0	0.0	0.0	0.934046
	0.0	0.0	0.0	0.210076
	0.0	0.0	0.0	0.528370

FIG. : Transactions considérées

Données

- 483.725 lignes, 14 colonnes
- Données claculées grace au scoring des transactions

Catégories d'algorithmes utilisés

- Support Vector Machine
- Decision Tree
- Random Forest
- K Nearest Neighbor
- Gradient boosting (XGBoost)
- Regression Logistique
- Naive Bayes

Support Vector Machine

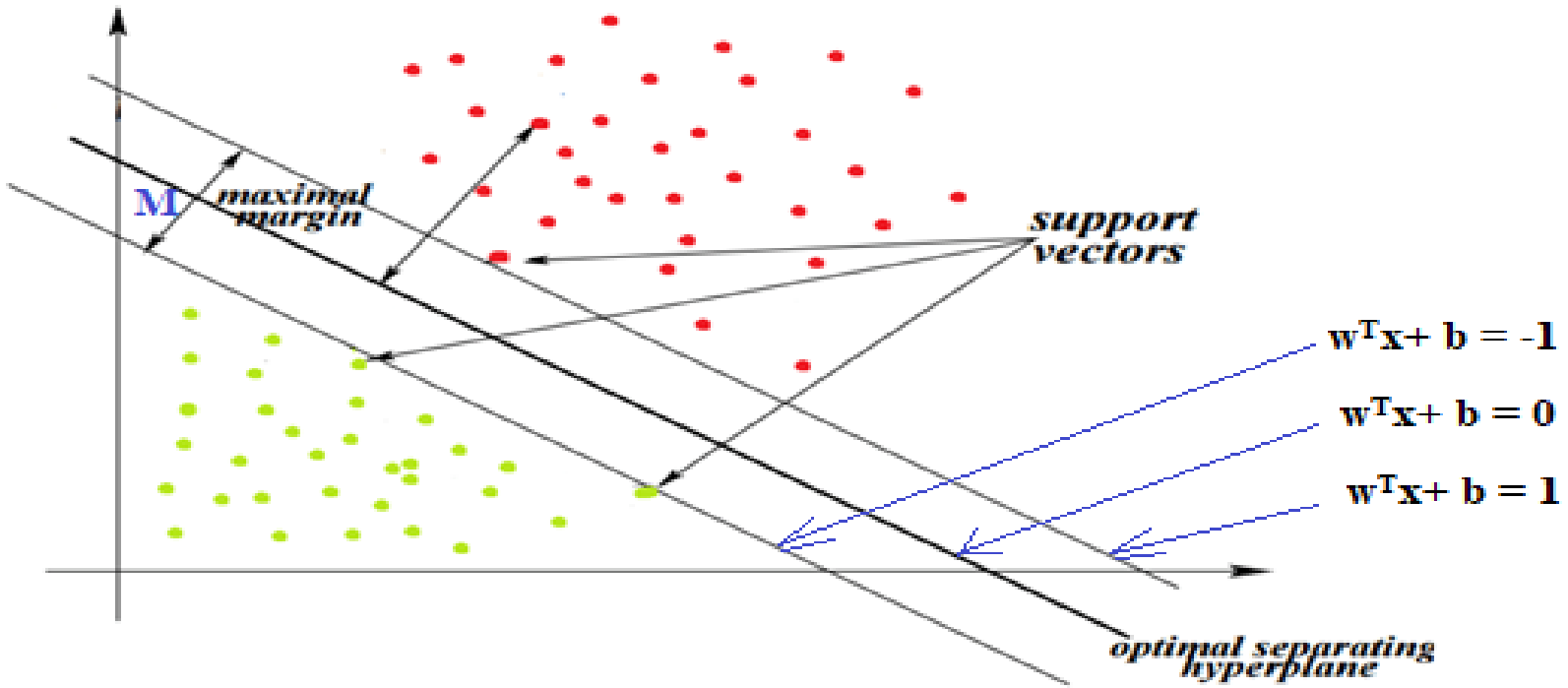


FIG. : Détermination de l'hyperplan

SVM : Déterminsation d'hyperplan

x_0 et x_1 deux vecteurs supports aux deux extrémités,
Soit l' hyperplant

$$(P) : w^T x + b = 0$$

$$M = d(x_0, P) + d(x_1, P) = \frac{| w^T x_0 + b |}{\sqrt{w^T w}} + \frac{| w^T x_1 + b |}{\sqrt{w^T w}}$$

$$= \frac{| 1 |}{\sqrt{w^T w}} + \frac{| -1 |}{\sqrt{w^T w}} = \frac{2}{\sqrt{w^T w}}$$

Maximiser M revient à minimiser

$$\frac{\sqrt{w^T w}}{2} = \frac{\|w\|}{2}$$

Arbre de décision

Time	Rain	Walk
30	1	No
15	1	No
5	1	No
10	0	No
5	0	No
15	0	Yes
20	0	Yes
25	0	Yes
30	0	Yes
30	0	Yes

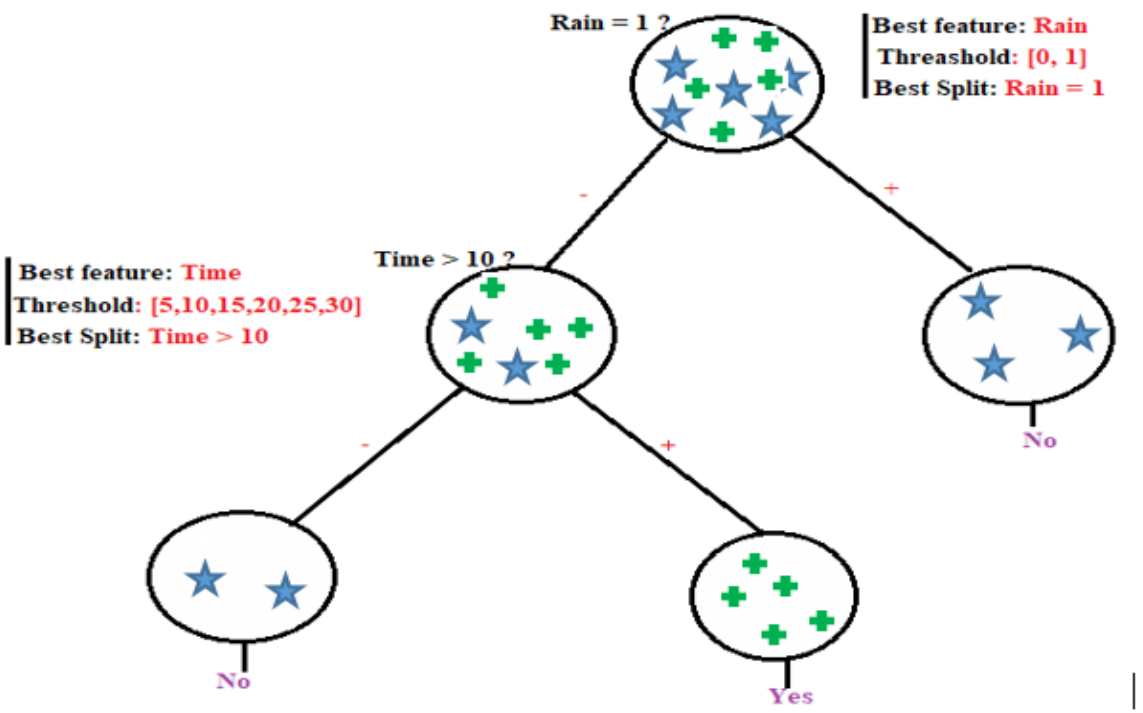


FIG. : Arbre de décision

Arbre de décision

$$Soit\ X_i\ (label) \in [\text{"}Yes\text{"}, \text{"}No\text{"}]$$

$$Posons\ P(X_i) = \frac{nb\_label\_i\_in\_node}{total\_population}$$

$$Pour\ Entropie :\ E = -\sum_{i=0}^{nb\_labels} P(X_i) * log_2(P(X_i))$$

$$Pour\ Gini :\ G = 1 - \sum_{i=0}^{nb\_labels} P(X_i)^2$$

Arbre de décision

Déterminer la meilleure variable et coupure qui correspond au Max(IG) :

$$IG = E(parent) - \sum_{i=0}^{nb\_childs} \frac{total\_population\_in\_node}{total\_population} E(child\_i)$$

$$IG = G(parent) - \sum_{i=0}^{nb\_childs} \frac{total\_population\_in\_node}{total\_population} G(child\_i)$$

Forêt aléatoire

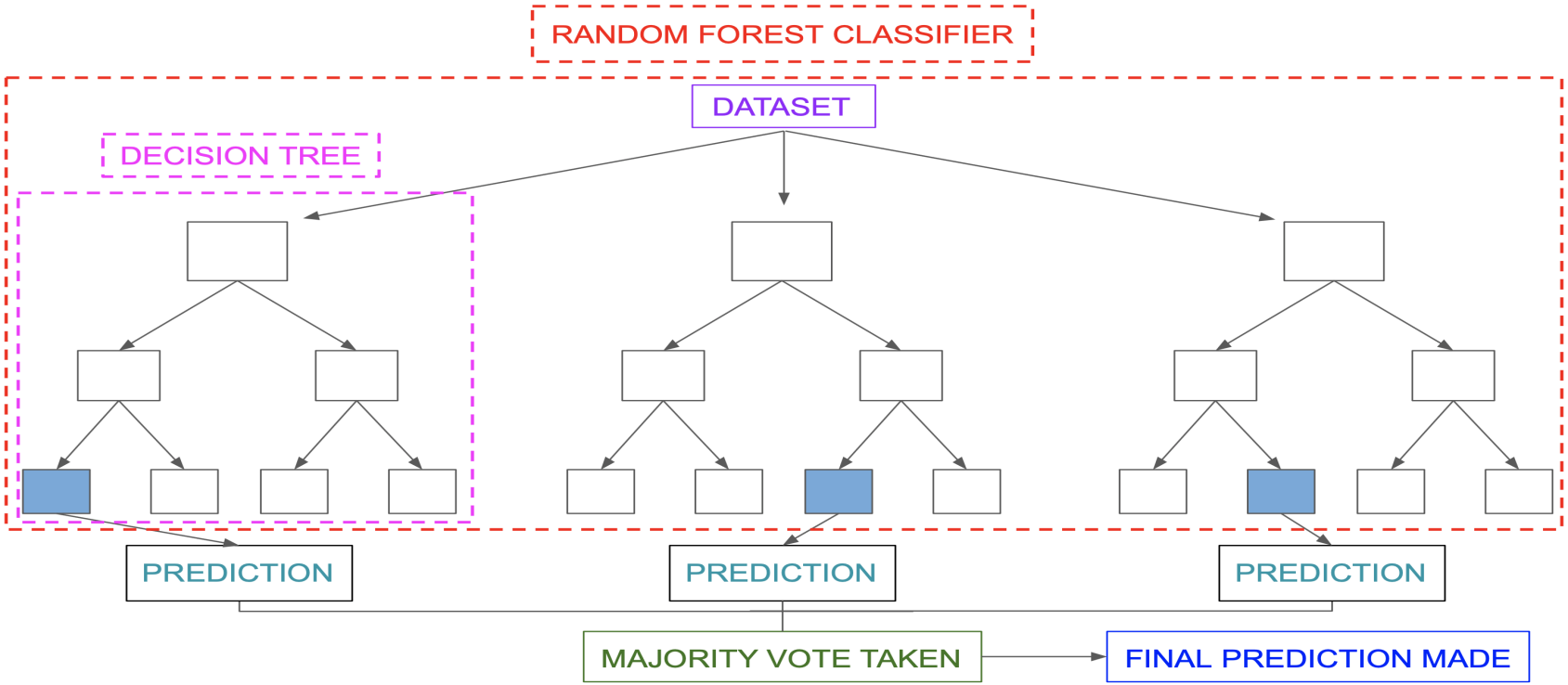


FIG. : Foret aléatoire

Fonctionnement du K-NN

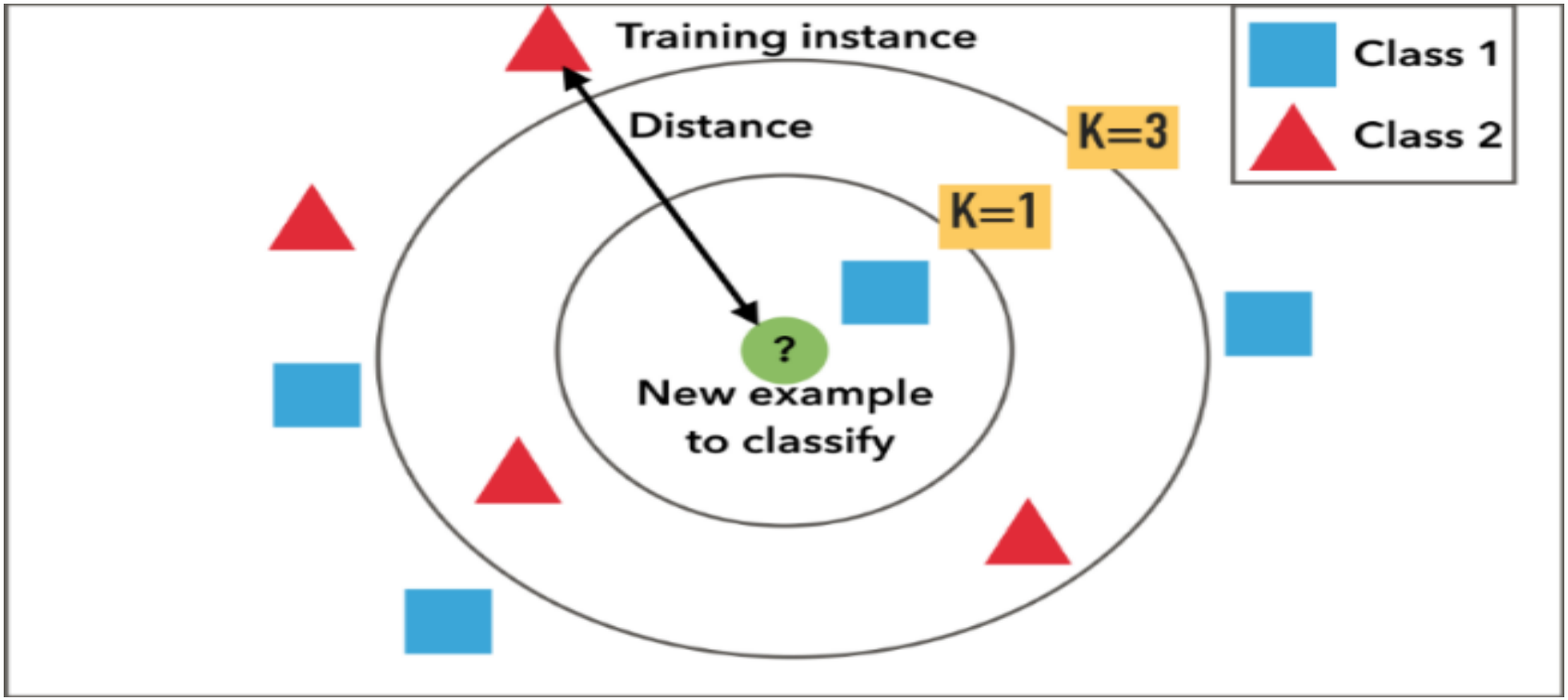


FIG. : K-Nearest Neighbor

Les types de distances

- Distance euclidienne

$$d(A, X) = \sqrt{\sum_{i=1}^{n} (a_i - x_i)^2}$$

- Distance de Manhattan

$$d(A, X) = \sum_{i=1}^{n} \mid a_i - x_i \mid$$

- Distance de Minkowski

$$d(A, X) = \sqrt[p]{\sum_{i=1}^{n} \mid a_i - x_i \mid^p}$$

Choix du paramètre K

- Utilisation de K

$$K = \sqrt{nombre - de - donnees}$$

- Choisir K suivant celui qui donne une meilleure prédiction

Informations données


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 483725 entries, 0 to 483724
Data columns (total 14 columns):
target                                483725 non-null int64
DayOfWeek                             483725 non-null object
amount                                483725 non-null float64
distance                              483725 non-null float64
distance_bin                          483725 non-null object
nb_transac                            483725 non-null int64
Purchases_InMall                      483725 non-null int64
Purchases_NextToTheMall               483725 non-null int64
Services_InMall                       483725 non-null int64
Purchases_InCity                      483725 non-null int64
Purchases_Potential_InMall            483725 non-null float64
Purchases_Potential_InCity            483725 non-null float64
Purchases_Potential_NextToTheMall     483725 non-null float64
cannibalisation                       483725 non-null float64
dtypes: float64(6), int64(6), object(2)
memory usage: 51.7+ MB
```

FIG. : Les types de features

Ici une conclusion qui met en valeur votre travail et indique ce qui reste à faire

Références

- Yingjie Tian, Yong Shi, Xiaohui Liu. RECENT ADVANCES ON SUPPORT VECTOR MACHINES RESEARCH. in TECHNOLOGICAL AND ECONOMIC DEVELOPMENT OF ECONOM, 2012 Volume 18(1) : 5–33
- Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood. Random Forest and Decision Tree. In IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012 ISSN (Online) : 1694-0814

Références

- Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. KNN Model-Based Approach in Classification. In School of Computing and Mathematics, University of Ulster Newtownabbey, BT37 0QB, Northern Ireland, UK
- Ramraj S, Nishant Uzir, Sunil R and Shatadeep Banerjee. Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets. In International Journal of Control Theory and Applications ISSN : 0974–5572 International Science Press Volume 9 ▪ Number 40 , 2016

Références

- C. Mitchell Dayton. LOGISTIC REGRESSION ANALYSIS. Department of Measurement, Statistics and Evaluation. In Room 1230D Benjamin Building University of Maryland September 1992

Merci pour votre attention