جامعة مولاي إسماعيل
ⵜⴰⵙⴷⴰⵡⵜ ⵎⵓⵍⴰⵢ ⵉⵙⵎⴰⵄⵉⵍ
UNIVERSITÉ MOULAY ISMAIL
كلية العلوم
ⵜⴰⵖⴰⵍⵓⵜ ⵏ ⵜⵣⵔⴼⵜ
FACULTÉ DES SCIENCES

Data mining project

cars prediction project

Prepared By

HADDANY ABDELKHALEK

KARMOUT ISMAIL

Project Context

we have been contacted by a car dealership to help them better target vehicles that may interest their clients. For this purpose, they provide you with:

• Their catalog of vehicles “**Catalogue.csv**”

	A	B	C	D	E	F	G
1	marque,nom,puissance,longueur,nbPlaces,nbPortes,couleur,occasion,prix						
2	Volvo,S80 T6,272,très longue,5,5,blanc,false,50500						
3	Volvo,S80 T6,272,très longue,5,5,noir,false,50500						
4	Volvo,S80 T6,272,très longue,5,5,rouge,false,50500						
5	Volvo,S80 T6,272,très longue,5,5,gris,true,35350						
6	Volvo,S80 T6,272,très longue,5,5,bleu,true,35350						
7	Volvo,S80 T6,272,très longue,5,5,gris,false,50500						
8	Volvo,S80 T6,272,très longue,5,5,bleu,false,50500						
9	Volvo,S80 T6,272,très longue,5,5,rouge,true,35350						
10	Volvo,S80 T6,272,très longue,5,5,blanc,true,35350						

• Their customer file for purchases in the current year “**Clients.csv**”

	A	B	C	D	E	F	G	H
1	age,sexe,taux,situationFamiliale,nbEnfantsAcharge,2eme voiture,immatriculation							
2	25,F,159,En Couple,2,false,3467 SB 72							
3	53,M,594,En Couple,2,false,113 LY 42							
4	20,F,949,En Couple,1,false,925 WK 87							
5	29,M,571,En Couple,2,false,3279 RV 81							
6	47,M,502,En Couple,1,false,82 RZ 54							
7	29,F,503,En Couple,3,false,8290 SD 41							
8	52,F,211,En Couple,4,true,9339 BW 87							
9	58,M,536,Célibataire,0,false,3696 JS 92							
10	21,M,211,En Couple,4,false,6484 MS 45							

- Access to all information on registrations made this year “**Immatriculations.csv**”

	A	B	C	D	E	F	G	H	I
1	immatriculation,marque,nom,puissance,longueur,nbPlaces,nbPortes,couleur,occasion,prix								
2	3176 TS 67,Renault,Laguna 2.0T,170,longue,5,5,blanc,false,27300								
3	3721 QS 49,Volvo,S80 T6,272,très longue,5,5,noir,false,50500								
4	9099 UV 26,Volkswagen,Golf 2.0 FSI,150,moyenne,5,5,gris,true,16029								
5	3563 LA 55,Peugeot,1007 1.4,75,courte,5,5,blanc,true,9625								
6	6963 AX 34,Audi,A2 1.4,75,courte,5,5,gris,false,18310								
7	5592 HQ 89,Skoda,Superb 2.8 V6,193,très longue,5,5,bleu,false,31790								
8	674 CE 26,Renault,Megane 2.0 16V,135,moyenne,5,5,gris,false,22350								
9	1756 PR 31,Mercedes,A200,136,moyenne,5,5,noir,true,18130								
10	6705 GX 50,BMW,120i,150,moyenne,5,5,noir,true,25060								
11	4487 DR 75,Saab,9.3 1.8T,150,longue,5,5,gris,true,27020								

- A brief documentation of the data

Catalogue.csv : catalogue de véhicules

Attribut	Type	Description	Domaine de valeurs
Marque	caractères	Nom de la marque du véhicule	Audi, BMW, Dacia, Daihatsu, Fiat, Ford,Honda, Hyundaï, Jaguar, Kia, Lancia,Mercedes , Mini, Nissan, Peugeot, Re- nault, Saab, Seat, Skoda, Volkswagen, Volvo
Nom	caractères	Nom du modèle de véhicule	S80 T6, Touran 2.0 FSI, Polo 1.2 6V, New Beatle 1.8, Golf 2.0 FSI, Superb 2.8 V6, Toledo 1.6, 9.3 1.8T, Vel Satis 3.5 V6, Megane 2.0 16V, Laguna 2.0T, Espace 2.0T, 1007 1.4, Primera 1.6, Maxima 3.0 V6, Almera 1.8, Copper 1.6 16V, S500, A200, Ypsilon 1.4 16V, Pi- canto 1.1, X-Type 2.5 V6, Matrix 1.6FR-V 1.7, Mondeo 1.8, Croma 2.2, Cuore 1.0, Logan 1.6 MPI, M5, 120i, A3 2.0 FSI, A2 1.4
Puissance	numérique	Puissance en chevaux Din	[55, 507]
Longueur	catégoriel	Catégorie de longueur	courte, moyenne, longue, très longue
NbPlaces	numérique	Nombre de places	[5, 7]
NbPortes	numérique	Nombre de portes	[3, 5]
Couleur	catégoriel	Couleur	blanc, bleu, gris, noir, rouge
Occasion	booléen	Véhicule d’occasion ?	true, false
Prix	numérique	Prix de vente en euros	[7500, 101300]

Immatriculations.csv : informations sur les immatriculations effectuées cette année

Attribut	Type	Description	Domaine de valeurs
Immatriculation	caractères	Numéro unique d'immatriculation du véhicule	Texte au format « 9999 AA 99 »
Marque	caractères	Nom de la marque du véhicule	Audi, BMW, Dacia, Daihatsu, Fiat, Ford, Honda, Hyundaï, Jaguar, Kia, Lancia,Mercedes , Mini, Nissan, Peugeot, Re- nault, Saab, Seat, Skoda, Volkswagen, Volvo
Nom	caractères	Nom du modèle de véhicule	S80 T6, Touran 2.0 FSI, Polo 1.2 6V, New Beatle 1.8, Golf 2.0 FSI, Superb 2.8 V6, Toledo 1.6, 9.3 1.8T, Vel Satis 3.5 V6, Megane 2.0 16V, Laguna 2.0T, Espace 2.0T, 1007 1.4, Primera 1.6, Maxima 3.0 V6, Almera 1.8, Copper 1.6 16V, S500, A200, Ypsilon 1.4 16V, Picanto 1.1, X-Type 2.5 V6, Matrix 1.6FR-V 1.7, Mondeo 1.8, Croma 2.2, Cuore 1.0, Logan 1.6 MPI, M5, 120i, A3 2.0 FSI, A2 1.4
Puissance		Puissance en chevaux Din	[55, 507]
Longueur		Catégorie de longueur	courte, moyenne, longue, très longue
NbPlaces	numérique	Nombre de places	[5, 7]
NbPortes	numérique	Nombre de portes	[3, 5]
Couleur	catégoriel	Couleur	blanc, bleu, gris, noir, rouge
Occasion	booléen	Véhicule d'occasion ?	true, false
Prix	numérique	Prix de vente en euros	[7500, 101300]

• An interview with a salesperson (see the interview below)

Your client will be satisfied if you propose a way for:

1. A salesperson can quickly assess the type of vehicle most likely to interest clients who come to the dealership
2. They can send precise documentation on the most suitable vehicle for clients selected by their marketing department.

Our work

For this requirement, I will attempt to explain my process in the following steps. Before delving into the steps, I will provide a summary of my work to help better understand its nature.

The primary focus of my work involves constructing two predictive models. The first model aims to **predict the cluster** of a car based on its information. The second model is designed to **predict the optimal car type**, or "cluster," for each client, taking into account their individual information.

Initially, I conducted preprocessing on the catalog dataset file to facilitate model training using the KMeans algorithm. For determining the optimal number of clusters, I employed the Elbow method, which aided in identifying four distinct clusters:

0: 'Family Cars',
1: 'Sport Utility Cars',
2: 'City Cars',
3: 'Luxury Cars'

After that, I did the preprocessing on the client's dataset and the matriculation, I merge them into a single data frame to facilitate the training of my second model.
after that, i train my model using three algorithms

1. Multinomial Logistic Regression
2. Random Forest Classifier
3. XGBoost

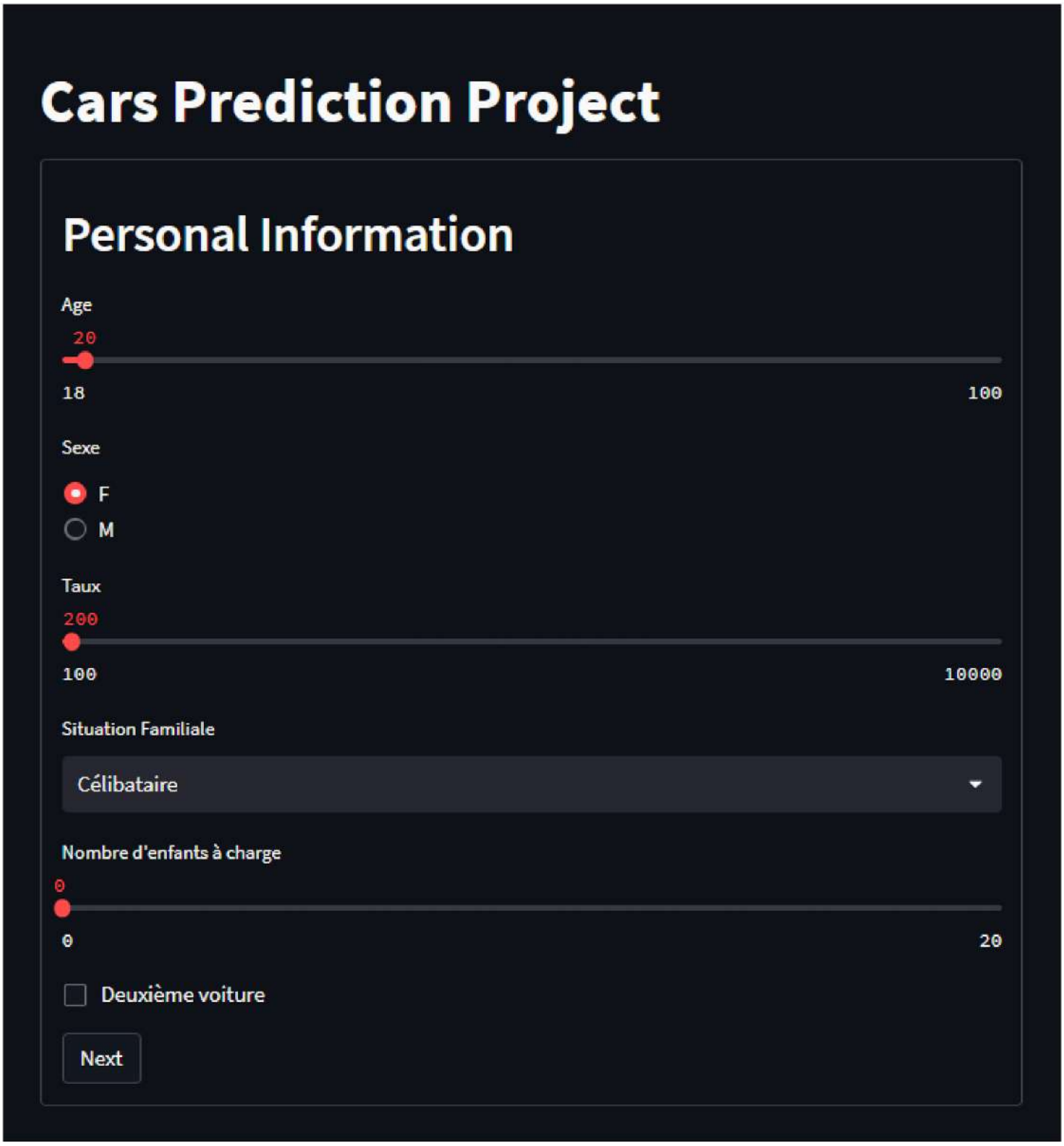
and after that, I try to choose one of those models depending on his performance using the result below:

```
Multinomial Logistic Regression Metrics:
Accuracy: 0.79
Precision: 0.77
Recall: 0.79
F1-score: 0.76

Random Forest Classifier Metrics:
Accuracy: 0.76
Precision: 0.75
Recall: 0.76
F1-score: 0.75

XGBoost Classifier Metrics:
Accuracy: 0.82
Precision: 0.82
Recall: 0.82
F1-score: 0.78
```

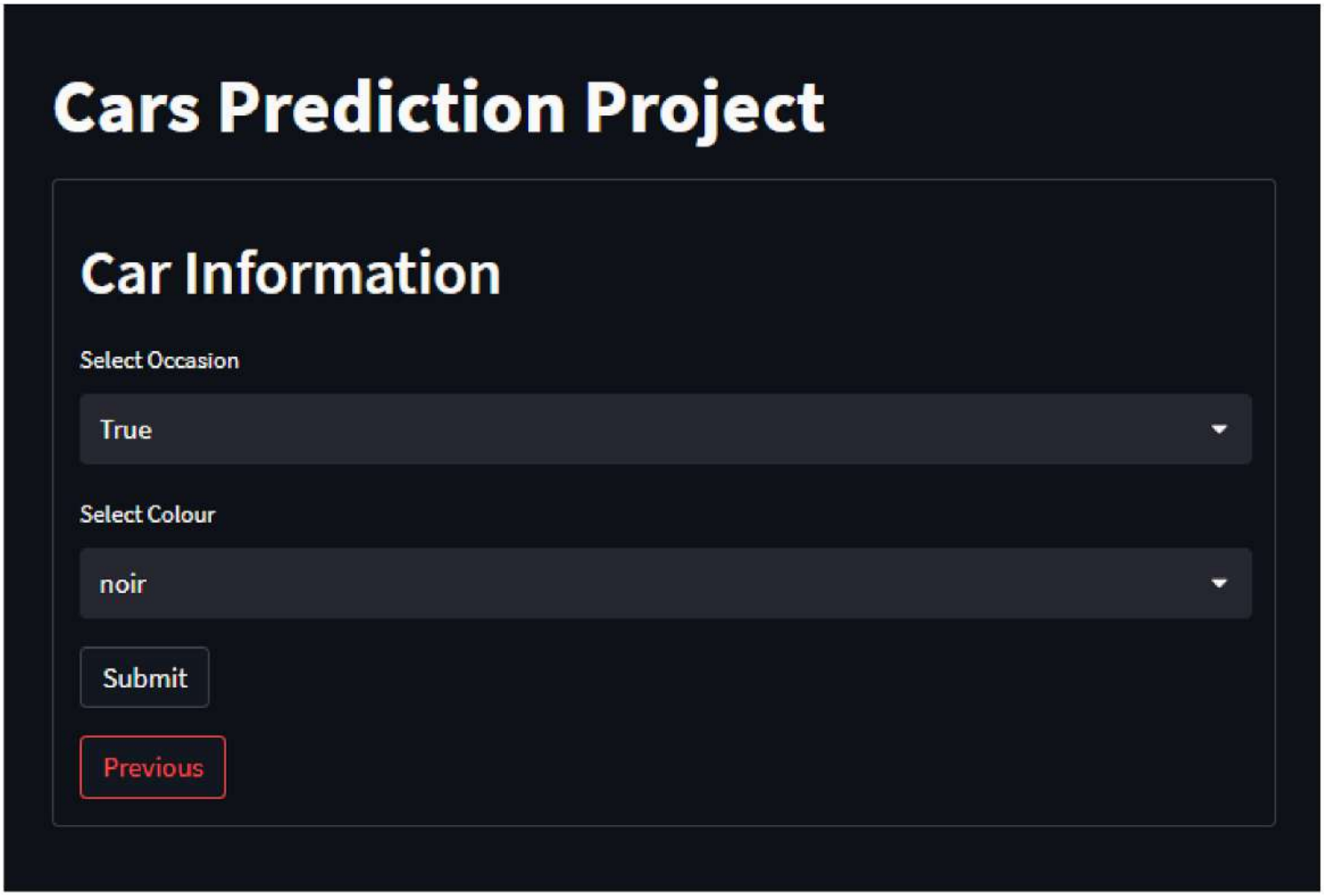
it seems that the XGBoost Classifier is the best when i tested them I found that the best model is **Logistic Regression**

Then I export this model to use it in my StreamLit interface

for my StreamLit interface is content two pages the first one is for the client information



The second page is for the car information and the results



When I click submit the interface shows me the results of which content is the best cluster for the user or the client depending on his information and it shows the proposed cars depending on the color and the occasion status all thats appear in the image bellow:

Car Information

Select Occasion

True

Select Colour

noir

Submit

Predicted Car Type: Sport Utility Cars

	marque	nom	puissance	longueur	nbPlaces	nbPortes	prix
9	Volvo	S80 T6	272	très longue	5	5	3535
31	Volkswagen	New Beatle 1.8	110	moyenne	5	5	1864
44	Volkswagen	Golf 2.0 FSI	150	moyenne	5	5	1602
64	Saab	9.3 1.8T	150	longue	5	5	2702
71	Renault	Vel Satis 3.5 V6	245	très longue	5	5	3444
84	Renault	Megane 2.0 16V	135	moyenne	5	5	1564
94	Renault	Laguna 2.0T	170	longue	5	5	1911
118	Peugeot	1007 1.4	75	courte	5	5	962
138	Mini	Copper 1.6 16V	115	courte	5	5	1274
159	Mercedes	A200	136	moyenne	5	5	1813

Previous

Conclusion

Upon the completion of this project, we found it gratifying due to its real-world nature. We dedicated effort to identifying a solution and determining the most effective approach for its implementation, striving to maximize accuracy through various datasets. Ultimately, our observations underscored the importance of having a substantial and balanced dataset to achieve optimal accuracy.