LOAN PREDICTION PROBLEM DATASET

Ibrahima BARRY

1. INTRODUCTION

- Dans ce TP, nous étudierons les données bancaires afin de pouvoir prédire, en utilisant des techniques d'apprentissage automatique, si une personne peut ou non faire un emprunt bancaire en fonction des informations qu'il va donner à travers une interface graphique.
- Pour cela il est nécessaire de passer par la phase de préparation du dataset qui consiste à améliorer la qualité de données afin d'en extraire tout le potentiel.

LES DONNÉES:

- Les données utilisées, provenant de la plateforme Kaggle, contiennent 983 lignes et 13 colonnes qui sont divisées en données d'entrainement et données de test
- Les colonnes sont: Loan ID, Gender, Married, Dependents, Education, Self Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term Credit_History, Property_Area et Loan_Status
- Lien: https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y
5	LP001011	Male	Yes	2	Graduate	Yes	5417	4196.0	267.0	360.0	1.0	Urban	Y
6	LP001013	Male	Yes	0	Not Graduate	No	2333	1516.0	95.0	360.0	1.0	Urban	Y
7	LP001014	Male	Yes	3+	Graduate	No	3036	2504.0	158.0	360.0	0.0	Semiurban	N
8	LP001018	Male	Yes	2	Graduate	No	4006	1526.0	168.0	360.0	1.0	Urban	Y
9	LP001020	Male	Yes	1	Graduate	No	12841	10968.0	349.0	360.0	1.0	Semiurban	N

2. LA PRÉPARATION DES DONNÉES:

Afin de préparer les dataset, les différentes techniques utilisées sont:

- D'abord séparer les variables catégoriques des variables numériques
- Pour les variables catégoriques, les valeurs manquantes sont remplacées sont remplacées par les valeurs qui se répètent le plus
- Pour les variables numériques, les valeurs manquantes sont remplacées par les valeurs précédentes
- Les valeurs des variables catégoriques sont converties en valeurs numériques

2. LA PRÉPARATION DES DONNÉES:

Le dataset obtenu se présente comme suit:

	Gender	Married	Dependents	Education	Self_Employed	Property_Area	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
0	1	0	0	0	0	2	5849.0	0.0	128.0	360.0	1.0
1	1	1	1	0	0	0	4583.0	1508.0	128.0	360.0	1.0
2	1	1	0	0	1	2	3000.0	0.0	66.0	360.0	1.0
3	1	1	0	1	0	2	2583.0	2358.0	120.0	360.0	1.0
4	1	0	0	0	0	2	6000.0	0.0	141.0	360.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...
609	0	0	0	0	0	0	2900.0	0.0	71.0	360.0	1.0
610	1	1	3	0	0	0	4106.0	0.0	40.0	180.0	1.0
611	1	1	1	0	0	2	8072.0	240.0	253.0	360.0	1.0
612	1	1	2	0	0	2	7583.0	0.0	187.0	360.0	1.0
613	0	0	0	0	1	1	4583.0	0.0	133.0	360.0	0.0

614 rows × 11 columns

3. ENTRAÎNEMENT ET ÉVALUATION DE MODÈLE

Dans ce travail, les 3 modèles utilisés sont:

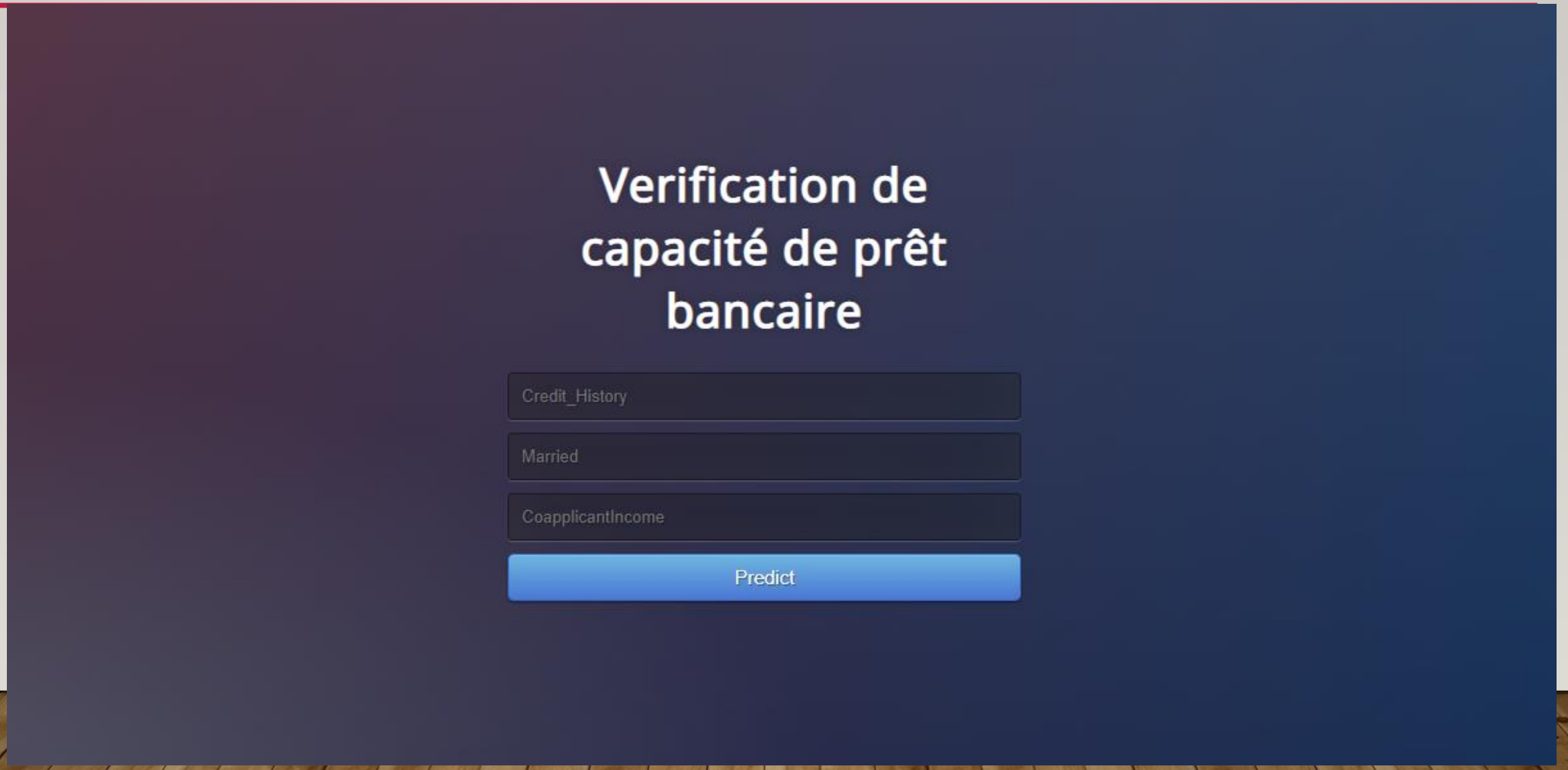
• Logistic Regression	85,36%
• k plus proches voisins (k-nearest neighbors)	65%
• Decision Tree	84%

Le modele choisi est donc la Logistic Regression

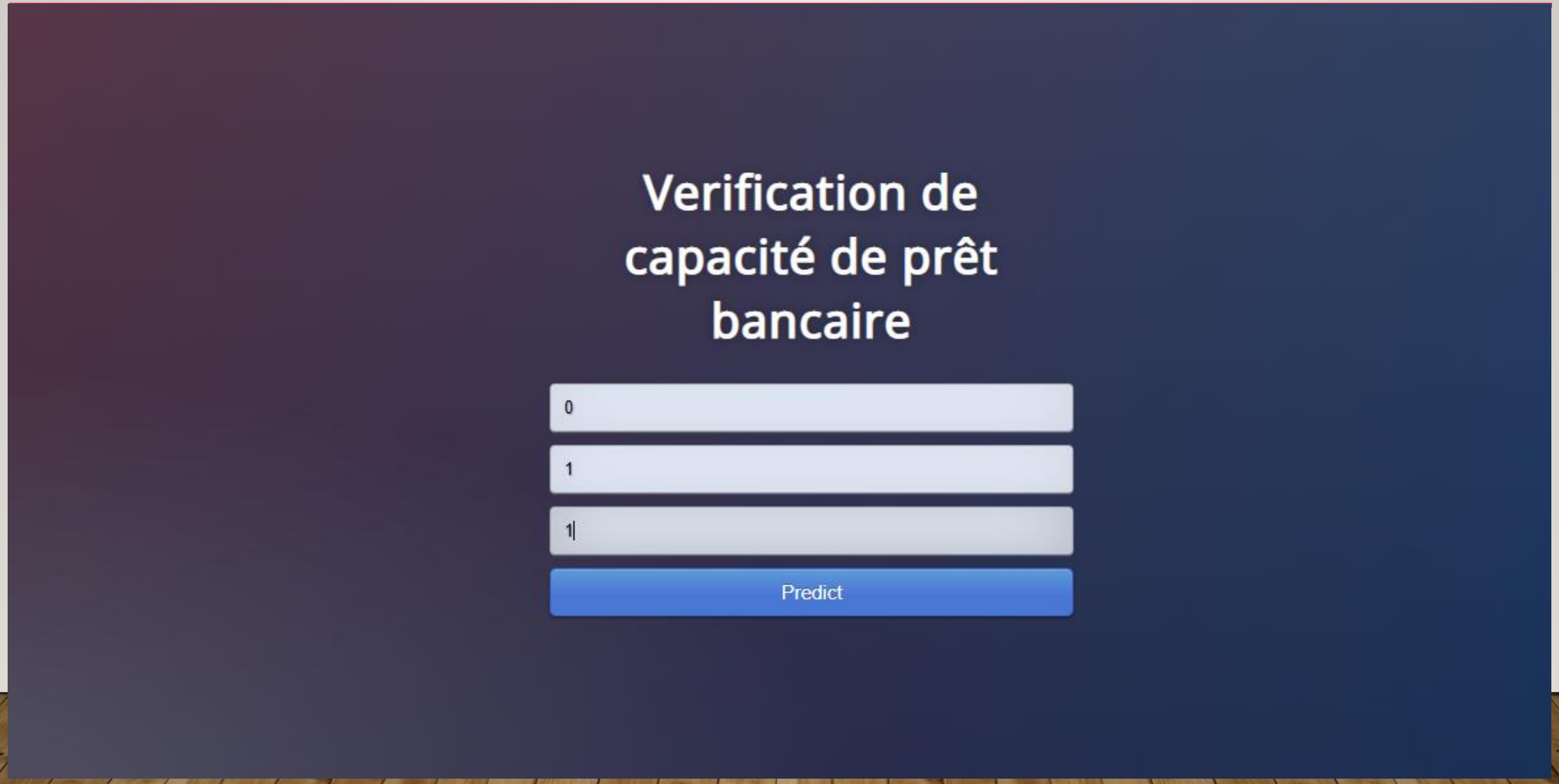
3. TEST ET DÉPLOIEMENT:

- Cette étape de machine learning permet de confronter notre model obtenu dans la phase précédente à la réalité du terrain.
- Pour cela on se sert généralement du dataset de test, mais pour ce travail il a été décidé d'utiliser une interface graphique qui permet à tout le monde d'interagir avec notre travail
- Pour créer cette interface graphique, nous avons utilisé Flask qui est un micro framework web python qui permet de créer des application web

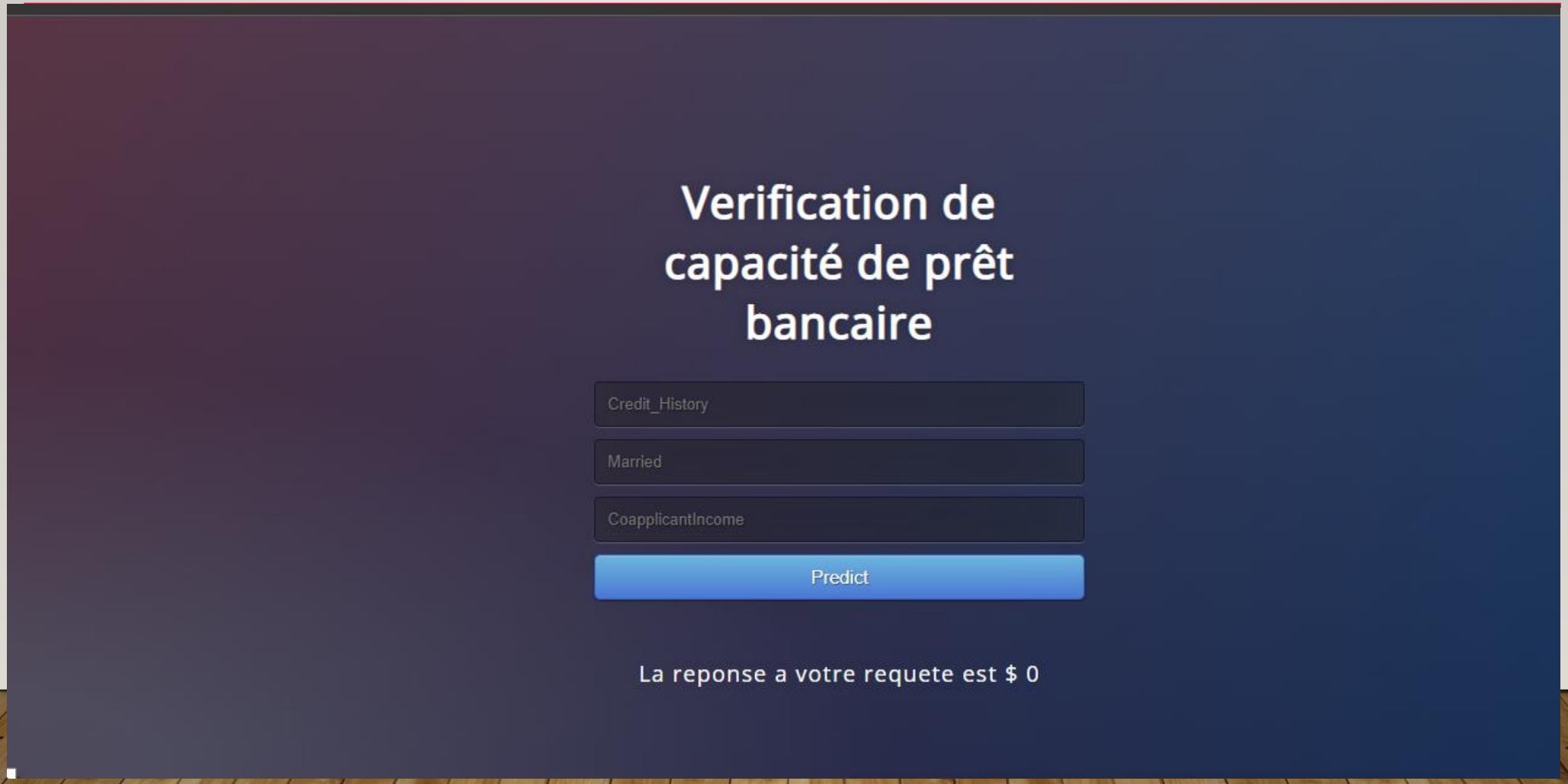
3.TEST ET DÉPLOIEMENT:



3.TEST ET DÉPLOIEMENT:



3.TEST ET DÉPLOIEMENT:



CONCLUSION