

# RAPPORT DE STAGE D'ÉTÉ- STAGE IGÉNIEUR

PRÉSENTÉ À

ECOLE NATIONALE D'ELECTRONIQUE ET DE TÉLÉCOMMUNICATIONS DE  
SFAX

INGÉNIERIE DES DONNÉES ET SYSTÈMES DÉCISIONNELS



## Système d'autocorrection des chèques

Réalisé par:

# Hadil SAHRAOUI

Encadrant Professionnels :

**Mrs. Islem SAOUDI**

**2023-2024**



# Remerciements

Je souhaite exprimer ma profonde gratitude à tous ceux qui ont contribué à la réussite de mon stage et qui m'ont soutenu tout au long de cette expérience enrichissante.

Tout d'abord, je tiens à remercier **l'ENET'COM** pour m'avoir offert cette opportunité de vivre un stage d'été à la fois professionnellement intéressant et formateur.

Mes sincères remerciements vont également à mon encadrante au sein de l'entreprise, **Madame Islem SAOUDI**, pour sa disponibilité, son accompagnement constant, et pour avoir répondu à toutes mes interrogations au cours de cette période.

Je tiens aussi à remercier toute l'équipe de **l'ATB (Arab Tunisian Bank)** pour leur accueil chaleureux et pour avoir créé un environnement propice à l'apprentissage et au développement de mes compétences.

# Table des matières

<b>Introduction Generale</b>	<b>1</b>
<b>1 Contexte générale</b>	<b>2</b>
Introduction . . . . .	3
1.1 Présentation de l'entreprise d'accueil . . . . .	3
1.1.1 Présentation de Arab Tunisien Bank-ATB . . . . .	3
1.1.2 Département IT de ATB . . . . .	4
1.2 Cadre de projet . . . . .	4
1.2.1 L'étude de l'existant . . . . .	4
1.2.2 Problématique . . . . .	5
1.2.3 Solution proposée . . . . .	5
1.3 Méthodologie de travail . . . . .	6
1.3.1 SEMMA . . . . .	6
1.3.2 KDD . . . . .	7
1.3.3 CRISP-DM . . . . .	8
1.3.4 La méthodologie choisie . . . . .	9
1.4 Planning de travail . . . . .	9
Conclusion . . . . .	10
<b>2 Compréhension du métier et des données</b>	<b>11</b>
Introduction . . . . .	12
2.1 Traitement des chèques . . . . .	12
2.1.1 Enjeux et Défis . . . . .	12
2.1.2 Impacts du Système d'Autocorrection des Chèques . . . . .	13
2.1.3 Objectifs Métiers . . . . .	13
2.1.4 Objectifs de la Data Science . . . . .	13
2.2 Compréhension des données . . . . .	14
2.2.1 Collecte des données . . . . .	14
2.2.2 Exploration et visualisation des données . . . . .	14

---

2.2.2.1	Visualisation des Images des Types de Chèques . . . . .	14
2.2.2.2	Distribution des Images selon les Banques . . . . .	15
Conclusion	. . . . .	16
<b>3</b>	<b>Préparation des Données</b>	<b>17</b>
Introduction	. . . . .	18
3.1	Préparation des données . . . . .	18
3.2	Prétraitement des Données . . . . .	19
3.2.1	Segmentation des données . . . . .	19
3.2.1.1	Extraction du Montant en Chiffres . . . . .	19
3.2.1.2	Extraction du Montant en Lettres . . . . .	19
3.2.1.3	Extraction du Nom du Client . . . . .	19
3.2.1.4	Extraction de la Date . . . . .	19
3.2.1.5	Extraction de l'identifiant de chèque . . . . .	20
3.2.2	Extraction et Tri des Chiffres . . . . .	20
3.2.2.1	Prétraitement des Images de Chiffres . . . . .	20
3.3	Extraction des Données . . . . .	21
3.3.1	k-Nearest Neighbors (KNN) . . . . .	21
3.3.1.1	Modèle . . . . .	21
3.3.1.2	Évaluation . . . . .	22
3.3.2	Paddle OCR . . . . .	25
3.3.3	PyTesseract . . . . .	25
3.3.4	Modèles Adaptés . . . . .	26
Conclusion	. . . . .	27
<b>4</b>	<b>Modélisation et evaluation</b>	<b>28</b>
Introduction	. . . . .	29
4.1	Etude théorique . . . . .	29
4.1.1	LSTM (Long Short-Term Memory) . . . . .	29
4.1.2	Seq2Seq (Sequence-to-Sequence) . . . . .	30
4.1.3	Mécanisme d'attention . . . . .	30
4.1.4	Transformers . . . . .	31

---

4.2	Modélisation	31
4.2.1	Création et Prétraitement des Données	32
4.2.2	Développement du Modèle	33
4.2.2.1	Choix du Modèle	33
4.2.2.2	Construction et Entraînement	33
4.2.2.3	Architecture Complète d'Autocorrection	34
4.3	Évaluation du Modèle	35
4.3.1	Visualisations des Performances	35
	Conclusion	36
<b>5</b>	<b>Déploiement</b>	<b>37</b>
	Introduction	39
5.1	Les technologies utilisés	39
5.1.1	Python	39
5.1.2	Django	39
5.1.3	SQLite	40
5.1.4	PowerBi	40
5.2	Pipeline de projet	41
5.2.1	Architecture du projet	41
5.2.2	Sitemap du site web	42
5.3	Les interfaces web de l'application	42
5.3.1	Page d'accueil	42
5.3.2	Page de connexion	43
5.3.3	Page d'inscription	43
5.3.4	Dashboard Administrateur	44
5.3.5	Interface de téléchargement des chèques	45
5.3.6	Interface de résultats du traitement des chèques	45
5.3.7	Interface d'informations supplémentaires	46
5.3.8	Table des Clients	47
5.3.9	Table des Chèques	47
5.3.10	Table des Employés	48
5.3.11	Données des clients	49

---

5.3.12	Données des chèques . . . . .	49
5.3.13	Données des employés . . . . .	50
5.3.14	Profil utilisateur . . . . .	51
Conclusion	. . . . .	51
<b>Conclusion Generale</b>		<b>52</b>

# Table des figures

1.1	Les étapes de SEMMA . . . . .	7
1.2	KDD . . . . .	7
1.3	Les étapes de crisp dm . . . . .	8
1.4	Diagramme de gantt . . . . .	10
2.1	Le processus de traitement des chèques . . . . .	12
2.2	Échantillon d'images de chèques . . . . .	15
2.3	Distribution des chèques par banque . . . . .	15
3.1	Préparation des images . . . . .	18
3.2	Montant en Chiffres . . . . .	19
3.3	Montant en Lettres . . . . .	19
3.4	Nom du Client . . . . .	19
3.5	la date . . . . .	20
3.6	l'identifiant de chèque . . . . .	20
3.7	Illustration des chiffres après prétraitement . . . . .	21
3.8	Architecture de modele KNN . . . . .	22
3.9	Illustration de modèle PaddlOCR . . . . .	23
3.10	Illustration de modèle PaddlOCR . . . . .	24
3.11	Illustration de modèle PaddlOCR . . . . .	24
3.12	Illustration de modèle PaddlOCR . . . . .	25
3.13	Illustration de modèle Pytesseract de Tesseract OCR . . . . .	26
4.1	Architecture LSTM . . . . .	29
4.2	Architecture Seq2Seq . . . . .	30
4.3	Architecture de Mécanisme d'attention . . . . .	30
4.4	Architecture de Transformers . . . . .	31
4.5	Architecture du Modèle . . . . .	33
4.6	Architecture Complète d'Autocorrection . . . . .	34
4.7	Evolution de la précision du modèle . . . . .	35

4.8	Perte pour l'entraînement et la validation au fil des époques . . . . .	36
5.1	Architecture du projet . . . . .	41
5.2	Sitemap du site web . . . . .	42
5.3	Page d'accueil . . . . .	43
5.4	Page de connexion . . . . .	43
5.5	Page d'inscription . . . . .	44
5.6	Dashboard Administrateur . . . . .	44
5.7	Interface de téléchargement des chèques . . . . .	45
5.8	Interface de résultats du traitement des chèques . . . . .	46
5.9	Interface d'informations supplémentaires . . . . .	46
5.10	Table des Banques . . . . .	47
5.11	Table des Chèques . . . . .	48
5.12	Table des Employés . . . . .	48
5.13	Données des clients . . . . .	49
5.14	Données des chèques . . . . .	50
5.15	Données des employés . . . . .	50
5.16	Profil utilisateur . . . . .	51

# Liste des tableaux

1.1 Tableau comparatif des méthodologies . . . . .	9
--	---

# Introduction Générale

Dans un environnement bancaire en constante évolution, l'efficacité opérationnelle et la précision des transactions sont devenues des impératifs pour garantir la satisfaction des clients et renforcer la compétitivité. Ce projet s'inscrit dans cette démarche en cherchant à moderniser le processus de vérification des chèques au sein de la Banque Arabe Tunisienne (ATB). Actuellement, la vérification des montants inscrits sur les chèques repose sur des méthodes manuelles, ce qui peut entraîner des erreurs humaines, des retards dans le traitement, et par conséquent, une diminution de la satisfaction client.

L'objectif principal de ce projet est de développer une solution innovante utilisant la reconnaissance optique de caractères (OCR) et les techniques d'apprentissage automatique pour automatiser la vérification et la correction des montants des chèques. Cette solution vise à minimiser les erreurs, à accélérer le processus de traitement, et à offrir une meilleure expérience aux clients de la banque.

Pour atteindre cet objectif, nous commencerons par recueillir et analyser les données relatives aux chèques traités par la banque, en mettant l'accent sur les erreurs couramment rencontrées lors de la vérification manuelle. Ensuite, nous développerons un modèle capable de détecter et de corriger automatiquement ces erreurs. Ce modèle sera intégré dans un système global qui permettra non seulement d'améliorer l'efficacité opérationnelle de la banque, mais aussi de renforcer la confiance des clients envers les services offerts.

Les résultats de ce projet permettront à ATB de moderniser un processus clé, d'améliorer la qualité de ses services, et de mieux répondre aux attentes de sa clientèle, tout en réduisant les coûts opérationnels liés aux erreurs de traitement..

# CONTEXTE GÉNÉRALE

---

## Plan

<b>Introduction . . . . .</b>	<b>3</b>
<b>1 Présentation de l'entreprise d'accueil . . . . .</b>	<b>3</b>
1.1 Présentation de Arab Tunisien Bank-ATB . . . . .	3
1.2 Département IT de ATB . . . . .	4
<b>2 Cadre de projet . . . . .</b>	<b>4</b>
2.1 L'étude de l'existant . . . . .	4
2.2 Problématique . . . . .	5
2.3 Solution proposée . . . . .	5
<b>3 Méthodologie de travail . . . . .</b>	<b>6</b>
3.1 SEMMA . . . . .	6
3.2 KDD . . . . .	7
3.3 CRISP-DM . . . . .	8
3.4 La méthodologie choisie . . . . .	9
<b>4 Planning de travail . . . . .</b>	<b>9</b>
<b>Conclusion . . . . .</b>	<b>10</b>

## Introduction

L'analyse d'un projet constitue une approche stratégique qui permet d'obtenir une perspective globale sur celui-ci afin de faciliter son bon déroulement.

Ce premier chapitre sera donc consacré à cette étude qui vise à situer le projet dans son contexte global. D'abord, il s'agit d'une présentation de l'organisme d'accueil, puis nous exposons le contexte du projet, puis nous présentons la méthodologie de travail, et nous terminons par le planning du travail.

### 1.1 Présentation de l'entreprise d'accueil

#### 1.1.1 Présentation de Arab Tunisien Bank-ATB

La Banque Arabe Tunisienne (ATB) est une banque de renom en Tunisie, créée en 1982. ATB a joué un rôle crucial dans le développement économique du pays depuis sa fondation en proposant une vaste gamme de services bancaires et financiers à ses clients, qu'ils soient particuliers ou professionnels. [1]



- **Historique :** Le secteur bancaire tunisien a été renforcé grâce à une initiative commune entre des investisseurs tunisiens et étrangers, connue sous le nom d'ATB. Depuis lors, elle n'a cessé de se développer et de varier ses prestations afin de satisfaire les besoins évolutifs de sa clientèle. ATB a réussi à s'ajuster aux avancées technologiques et réglementaires au fil du temps, tout en maintenant une approche axée sur le client.

- **Mission :** ATB a pour objectif de proposer des services bancaires de qualité supérieure, en mettant l'accent sur l'innovation, la satisfaction des clients et la responsabilité sociale. L'objectif d'ATB est de devenir la principale banque en Tunisie, connue pour son excellence opérationnelle et son engagement envers le développement économique et social du pays.
- **Valeurs :** Intégrité, clarté, créativité et dévouement envers les clients.

### 1.1.2 Département IT de ATB

Le département des TIC de l'Arab Tunisian Bank (ATB) occupe une place centrale dans la transition numérique de la banque. Son rôle consiste à concevoir, mettre en place et gérer les systèmes d'information et les technologies qui soutiennent les opérations bancaires. Ce service joue un rôle crucial dans la préservation de la compétitivité de l'ATB dans un contexte bancaire en perpétuelle mutation.

Le 1er mars 2017, la banque Arabe Tunisienne reçoit la certification ISO 27001 pour ses services de banque en ligne et de banque mobile. Ainsi, l'ATB devient la seule banque tunisienne à être certifiée de cette manière et la troisième à l'échelle mondiale. [1]

## 1.2 Cadre de projet

Ce travail s'inscrit dans le cadre de mon projet de stage d'été chez Arab Tunisien Bank (ATB). ATB m'a donné l'opportunité et la confiance pour la réalisation de ce projet, qui porte sur le développement d'une application dédiée à l'autocorrection des montants sur les chèques bancaires. Ce projet a pour but de proposer une solution innovante visant à automatiser le processus de vérification et de correction des montants inscrits sur les chèques, afin de réduire les erreurs humaines et d'améliorer l'efficacité des opérations bancaires.

### 1.2.1 L'étude de l'existant

Dans le cadre de l'étude de l'existant, il est important de souligner que les méthodes actuelles de vérification des chèques chez ATB reposent principalement sur des processus manuels. Ces méthodes, bien que robustes, sont sujettes à des erreurs humaines et peuvent entraîner des retards dans le traitement des chèques. De plus, ces processus manuels manquent de précision et de rapidité, ce qui peut affecter la satisfaction des clients et l'efficacité globale des opérations bancaires. Les principales lacunes des méthodes actuelles sont :

- **Vérification manuelle des montants sur les chèques :** Cette méthode est susceptible d'erreurs humaines, pouvant entraîner des incohérences dans les transactions.
- **Processus de traitement long et inefficace :** Les opérations manuelles sont chronophages, ce qui ralentit le traitement global des chèques.
- **Absence d'une solution intégrée pour la correction automatique des erreurs :** Actuellement, il n'existe pas de mécanisme automatique pour corriger les erreurs potentielles dans les montants inscrits sur les chèques.

Face à ces lacunes, ATB s'est fixé les objectifs suivants pour moderniser et optimiser ses processus :

- **Automatiser la vérification et la correction des montants :** Réduire les erreurs humaines en introduisant une solution automatisée qui détecte et corrige les incohérences.
- **Améliorer l'efficacité et la rapidité du traitement des chèques :** Accélérer le traitement des chèques en réduisant la dépendance aux méthodes manuelles.
- **Mettre en place une solution basée sur l'IA et le Machine Learning :** Intégrer des technologies avancées pour l'autocorrection des montants, garantissant ainsi une meilleure précision et fiabilité des opérations bancaires.

ATB cherche ainsi à moderniser ses processus en adoptant des technologies avancées pour améliorer la précision et la rapidité des opérations liées aux chèques bancaires.

### 1.2.2 Problématique

Dans un environnement bancaire de plus en plus compétitif, il est crucial pour les institutions financières de minimiser les erreurs et d'optimiser leurs processus internes pour offrir un service de haute qualité à leurs clients. Les erreurs dans la transcription des montants sur les chèques peuvent entraîner des problèmes significatifs, tant pour la banque que pour les clients, tels que des retards dans le traitement des paiements ou des incohérences dans les comptes. Ainsi, la nécessité de développer une solution automatisée pour la détection et la correction des erreurs sur les chèques devient une priorité pour ATB.

### 1.2.3 Solution proposée

Pour répondre à cette problématique, nous proposons de développer une application dédiée à l'autocorrection des montants inscrits sur les chèques, basée sur les technologies de l'intelligence

artificielle (IA) et de l'apprentissage automatique (Machine Learning). La solution se décompose en plusieurs étapes :

1. **Prétraitement des Images** : Les chèques seront scannés et les images seront prétraitées pour améliorer la lisibilité des montants inscrits, en utilisant des techniques telles que la conversion en niveaux de gris, la binarisation et le nettoyage des images.
2. **Extraction des Montants** : L'application utilisera des modèles OCR (reconnaissance optique de caractères) pour extraire les montants en chiffres et en lettres des images des chèques.
3. **Vérification et Correction Automatique** : Un algorithme d'autocorrection analysera les montants extraits et détectera les éventuelles incohérences entre le montant en chiffres et celui en lettres. En cas de divergence, l'application proposera une correction automatique basée sur des modèles de traitement du langage naturel (NLP).
4. **Intégration et Visualisation** : Une interface utilisateur permettra aux agents bancaires de superviser le processus, de valider les corrections proposées, et de générer des rapports sur les opérations traitées.

Cette solution permettra à ATB de moderniser et d'automatiser son processus de traitement des chèques, réduisant ainsi les erreurs humaines, améliorant l'efficacité opérationnelle, et augmentant la satisfaction client.

### 1.3 Méthodologie de travail

#### 1.3.1 SEMMA

SEMMA est un acronyme qui signifie Sample,Explore,Modify,Model et Access.C'est une méthode pour réaliser des projets en data science.Cette méthode simplifie l'application des techniques d'exploration statistique,la sélection et la transformation des variables prédictives les plus primordiales,la modélisation des variables pour prédire les résultats et la confirmation de la précision du modèle.

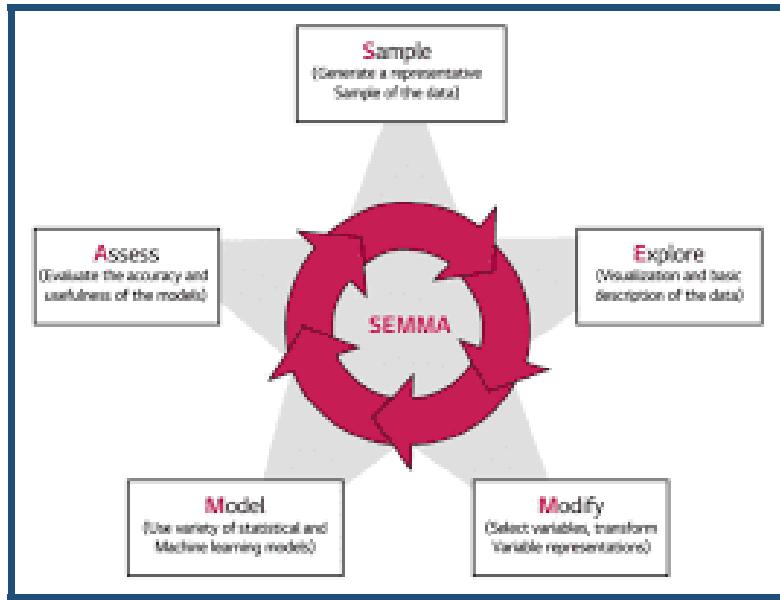


FIGURE 1.1 : Les étapes de SEMMA

### 1.3.2 KDD

KDD est le processus itératif et interactif du projet de recherche proposé par Osama Fayyad en 1996. Il s'agit d'une méthode qui permet aux experts d'extraire des modèles ou des informations requises à partir des données.

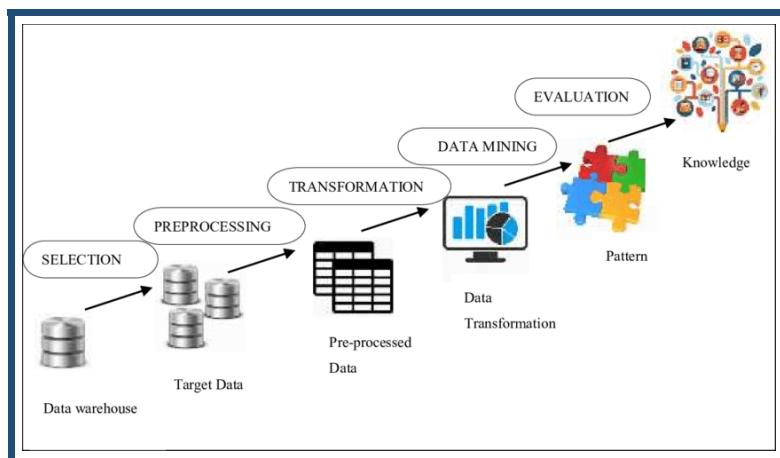


FIGURE 1.2 : KDD

Elle comprend cinq étapes : sélection, pré-traitement, transformation, exploration de données, et interprétation/évaluation.

### 1.3.3 CRISP-DM

CRISP-DM signifie le Processus standard intersectoriel pour l'exploration de données. Cette méthodologie a été développée par IBM pour réaliser des projets en data science. Aujourd'hui elle reste l'unique méthode adoptée pour tous les projets Data Science. Ce modèle a la même nature cyclique que KDD et SEMMA, la principale différence dans la structure est que les transitions entre les étapes peuvent être inversées. Ainsi, si lors de la phase de modélisation, le spécialiste trouve que les données ne sont pas suffisantes pour résoudre l'objectif du projet, il peut revenir à la phase de préparation des données et sélectionner différentes variables cibles, générer des fonctionnalités, etc., sans revenir au début du cycle.

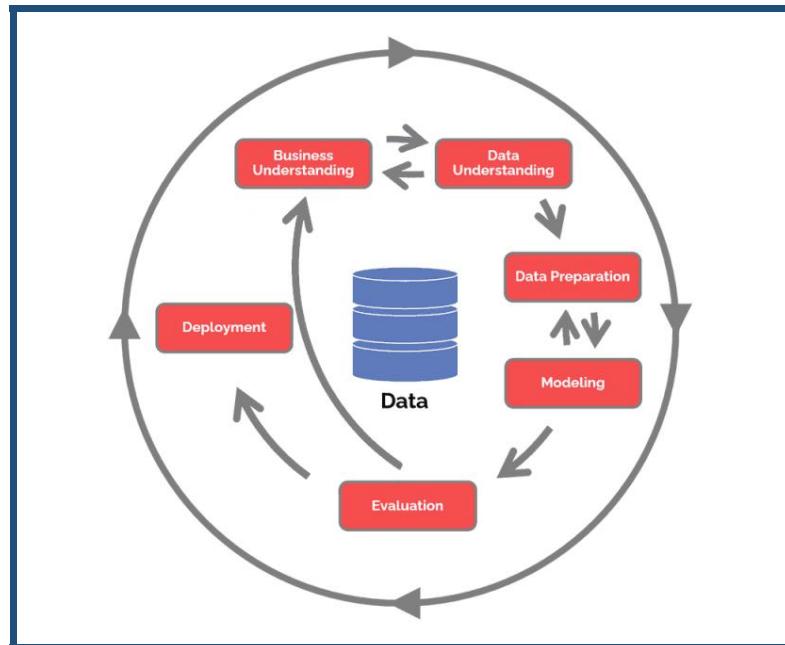


FIGURE 1.3 : Les étapes de crisp dm

La méthode CRISP-DM comprend six étapes, allant de la compréhension du problème au déploiement : Compréhension du métier, Compréhension des données, Préparation des données, Modélisation, Évaluation et Déploiement.

### 1.3.4 La méthodologie choisie

Dans le cadre de notre projet, nous avons évalué différentes méthodologies pour guider les différentes étapes de notre projet. Nous avons comparé trois approches couramment utilisées : SEMMA, KDD et CRISP-DM. Le tableau 1.1 présente une comparaison des avantages et des inconvénients de ces méthodologies.

Méthodologie	Avantages	Inconvénients
SEMMA	-Processus itératif -Flexibilité dans les étapes	-Manque de standardisation - Dépendance à l'expertise humaine
KDD	-Approche holistique -Utilisation efficace des données	-Complexité -Nécessite de grandes quantités de données
CRISP-DM	- Structure en 6 phases -Prise en compte de l'aspect métier	- Temps et coûts élevés -Difficulté à s'adapter aux changements

**TABLEAU 1.1 :** Tableau comparatif des méthodologies[2]

Nous avons choisi de travailler avec la méthodologie CRISP-DM en raison des points suivants :

- Cette méthodologie est rentable car elle comprend un certain nombre de processus permettant d'éliminer les tâches simples d'exploration de données, et les processus sont bien établis dans l'industrie.
- CRISP-DM encourage les meilleures pratiques et permet aux projets de se reproduire. Cette méthodologie fournit un cadre uniforme pour la planification et la gestion d'un projet.
- En tant que norme intersectorielle, CRISP-DM peut être mise en œuvre dans tout projet de science des données, quel que soit son domaine.

## 1.4 Planning de travail

Afin de garantir une bonne conduite de notre projet, il est essentiel de s'organiser et de maintenir un équilibre entre le temps et l'avancement du travail, tout en respectant les différentes phases du projet. Pour cela, nous allons utiliser un outil couramment utilisé en gestion de projet :

## Chapitre 1. Contexte générale

---

le diagramme de Gantt. Cet outil nous permettra de représenter visuellement l'état d'avancement des différentes tâches qui constituent notre projet, ainsi que de modéliser les tâches nécessaires à sa réalisation.

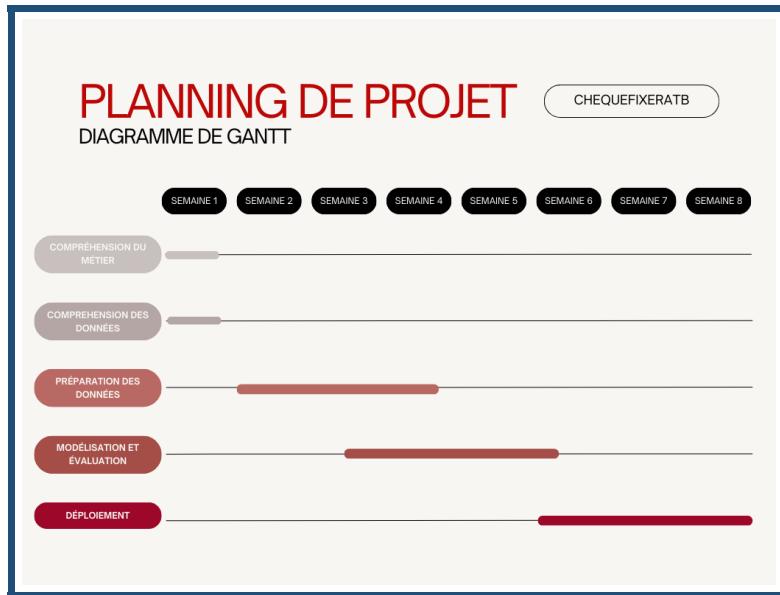


FIGURE 1.4 : Diagramme de gantt

## Conclusion

Ce chapitre introductif nous a permis de présenter l'entreprise d'accueil et de situer le projet dans son cadre général. À partir d'une étude théorique qui nous a donné une vue d'ensemble sur l'état actuel, la problématique et la solution ont été identifiées. Après avoir établi les objectifs, cette vision sera approfondie dans la partie sur la compréhension du métier et des données.

---

*CHAPITRE 2*

---

# COMPRÉHENSION DU MÉTIER ET DES DONNÉES

---

## Plan

<b>Introduction</b> . . . . .	<b>12</b>
<b>1 Traitement des chèques</b> . . . . .	<b>12</b>
1.1 Enjeux et Défis . . . . .	12
1.2 Impacts du Système d'Autocorrection des Chèques . . . . .	13
1.3 Objectifs Métiers . . . . .	13
1.4 Objectifs de la Data Science . . . . .	13
<b>2 Compréhension des données</b> . . . . .	<b>14</b>
2.1 Collecte des données . . . . .	14
2.2 Exploration et visualisation des données . . . . .	14
<b>Conclusion</b> . . . . .	<b>16</b>

## Introduction

Dans ce chapitre, nous abordons la compréhension du métier, des données et leur préparation, les premières phases de CRISP-DM. Cette étape permet de définir les besoins des parties prenantes, de clarifier les objectifs métier et data science, et de présenter les types de systèmes de recommandation, axe central de notre étude. Nous détaillons également la collecte et la compréhension des données.

### 2.1 Traitement des chèques

Le processus de traitement des chèques dans les banques implique actuellement une saisie manuelle des montants en chiffres et en lettres à partir des images des chèques. Ce processus est essentiellement manuel, ce qui le rend sujet aux erreurs humaines et à des délais potentiels dans le traitement des transactions financières.



**FIGURE 2.1 :** Le processus de traitement des chèques

#### 2.1.1 Enjeux et Défis

Les erreurs de montant sur les chèques posent des défis majeurs pour les opérations bancaires, affectant finances et satisfaction client :

- **Impact sur les Opérations :** Elles introduisent des retards et de l'incertitude dans le traitement des transactions.

- **Coûts et Satisfaction Client** : Les corrections entraînent des coûts supplémentaires et peuvent réduire la satisfaction client.

Il est essentiel de développer des solutions pour réduire ces impacts et améliorer la précision des transactions.

### 2.1.2 Impacts du Système d'Autocorrection des Chèques

- **Réduction des erreurs** : Automatiser la correction des montants diminue les erreurs humaines, garantissant des transactions plus fiables.
- **Efficacité et réduction des coûts** : Cette automatisation accélère le traitement des chèques et réduit les ressources nécessaires, entraînant des économies pour la banque.
- **Satisfaction client accrue** : Des transactions plus précises et rapides améliorent l'expérience client et réduisent les risques d'erreurs financières.

### 2.1.3 Objectifs Métiers

- **Réduire les erreurs de saisie** : Implémentation d'un système OCR avancé pour extraire précisément les montants des chèques et détection automatique des incohérences.
- **Améliorer l'efficacité** : Développement d'algorithmes pour automatiser la lecture et la validation des chèques, et utilisation du machine learning pour accélérer la vérification.
- **Minimiser les coûts** : Mise en place d'un système de correction automatique des erreurs pour réduire les coûts liés aux erreurs et aux pénalités.

### 2.1.4 Objectifs de la Data Science

- **Collecte des données** : Collecter les images de chèques et extraire les montants écrits en chiffres et en lettres.
- **Préparation des données** : Nettoyer, prétraiter et normaliser les images des chèques pour améliorer la qualité de l'extraction des montants.
- **Modélisation des données** : Utiliser des modèles OCR préentraînés pour extraire avec précision les montants en chiffres et en lettres des images des chèques et intégrer des techniques avancées de NLP pour une correction automatique.
- **Déploiement** : Intégrer le système d'extraction automatique des montants dans une application bancaire pour une utilisation en temps réel et sécurisée.

## 2.2 Compréhension des données

### 2.2.1 Collecte des données

Pour mener à bien notre projet, nous avons utilisé le IDRBT Cheque Image Dataset, une source de données exhaustive et fiable pour notre analyse. Voici une description détaillée de cette source de données :

**Images de chèques** : Le dataset contient 112 images de chèques provenant de quatre banques différentes en Inde. Les chèques ont été écrits par neuf volontaires utilisant sept stylos bleus et sept stylos noirs, créant une diversité de combinaisons de styles d'écriture et d'encre. Chaque chèque a été scanné à une résolution de 300 dpi pour garantir une qualité optimale pour l'extraction des montants.

**Métadonnées associées** : Pour chaque image de chèque, des métadonnées détaillées sont fournies, incluant :

- **ID du chèque** : Un identifiant unique pour chaque chèque.
- **Type de stylo** : Information sur le type et la couleur du stylo utilisé.
- **Banque émettrice** : La banque qui a émis le chèque.
- **Montants en lettres et en chiffres** : Les montants exacts en lettres et en chiffres tels qu'écrits sur le chèque.
- **Volontaires** : Identifiant des volontaires ayant écrit les chèques, permettant d'analyser les variations individuelles dans l'écriture.

La collecte de ces données a été réalisée en respectant les normes de confidentialité et de sécurité des informations, garantissant la protection des données sensibles des clients.

### 2.2.2 Exploration et visualisation des données

L'exploration des données permet de comprendre et d'analyser les tendances et modèles cachés. Les techniques de visualisation, telles que les graphiques et tableaux de bord, aident à vérifier la qualité des données et à détecter les erreurs ou données manquantes.

#### 2.2.2.1 Visualisation des Images des Types de Chèques

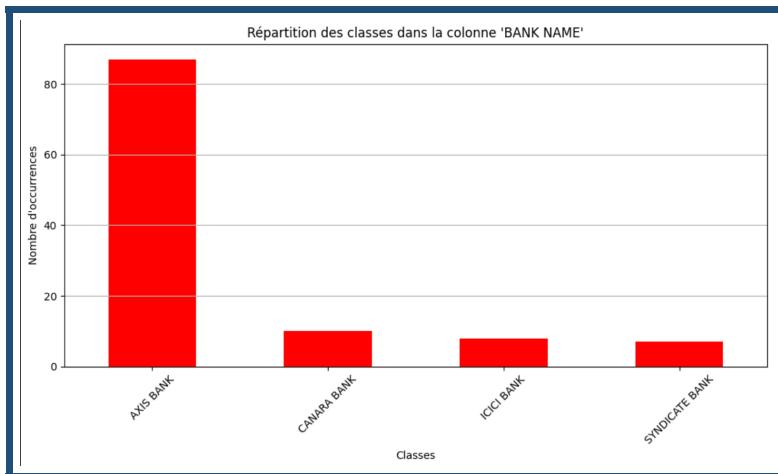
Nous avons commencé par visualiser un échantillon des images de chèques pour comprendre la diversité des types de chèques présents dans notre dataset.



**FIGURE 2.2 :** Échantillon d'images de chèques

### 2.2.2.2 Distribution des Images selon les Banques

Ensuite, nous avons analysé la distribution des images de chèques selon les banques pour comprendre la répartition de notre dataset.



**FIGURE 2.3 :** Distribution des chèques par banque

Le graphique ci-joint illustre la répartition des occurrences des différents noms de banques dans la colonne 'BANK NAME'. Il met en évidence une distribution inégale du nombre de chèques émis par ces banques. AXIS BANK se démarque avec le plus grand nombre de chèques, dépassant 80 occurrences, tandis que CANARA BANK, ICICI BANK et SYNDICATE BANK ont chacune environ 10 chèques ou moins. Cette visualisation initiale nous permet de comprendre la diversité des échantillons de chèques provenant de différentes institutions bancaires, ce qui est crucial pour l'analyse et la préparation ultérieure des données.

## Conclusion

Dans ce deuxième chapitre, nous avons abordé la première phase de la méthodologie CRISP-DM en étudiant et identifiant les besoins de notre projet. Nous avons résolu notre problématique en identifiant les objectifs métier et les objectifs DataScience de notre application .Puis, nous avons passé vers la partie compréhension des données. Enfin, nous avons abordé l'analyse des données selon la méthodologie CRISP-DM.

# PRÉPARATION DES DONNÉES

---

## Plan

<b>Introduction . . . . .</b>	<b>18</b>
<b>1      Préparation des données . . . . .</b>	<b>18</b>
<b>2      Prétraitement des Données . . . . .</b>	<b>19</b>
2.1     Segmentation des données . . . . .	19
2.2     Extraction et Tri des Chiffres . . . . .	20
<b>3      Extraction des Données . . . . .</b>	<b>21</b>
3.1     k-Nearest Neighbors (KNN) . . . . .	21
3.2     Paddle OCR . . . . .	25
3.3     PyTesseract . . . . .	25
3.4     Modèles Adaptés . . . . .	26
<b>Conclusion . . . . .</b>	<b>27</b>

## Introduction

Dans ce chapitre, nous allons aborder les étapes cruciales pour présenter les techniques de préparation et de pré-traitement des données collectées. Tout au long de ce chapitre, nous explorerons l'importance de la préparation des données pour garantir que les données extraites soient de la plus haute qualité possible, facilitant ainsi leur utilisation dans les modèles de reconnaissance de caractères et d'analyse.

### 3.1 Préparation des données

La préparation des données est une étape critique dans le traitement des images de chèques, visant à optimiser la qualité et la pertinence des informations extraites. Nous utilisons plusieurs techniques pour prétraiter les images afin de faciliter l'extraction des données nécessaires : **le chargement de l'image** à partir du chemin spécifié, **la suppression des éléments non pertinents** tels que le logo de la banque et les mentions inutiles pour simplifier l'image, **la conversion de l'image couleur en niveaux de gris** pour conserver les informations d'intensité lumineuse, **l'application d'un filtre** qui remplace chaque pixel par la médiane de ses voisins pour réduire le bruit, **la conversion de l'image en niveaux de gris** en une image binaire, avec des pixels blancs ou noirs selon un seuil, **le seuillage adaptif**, **l'érosion** pour réduire le bruit et dilatation pour lisser les frontières et combler les trous.

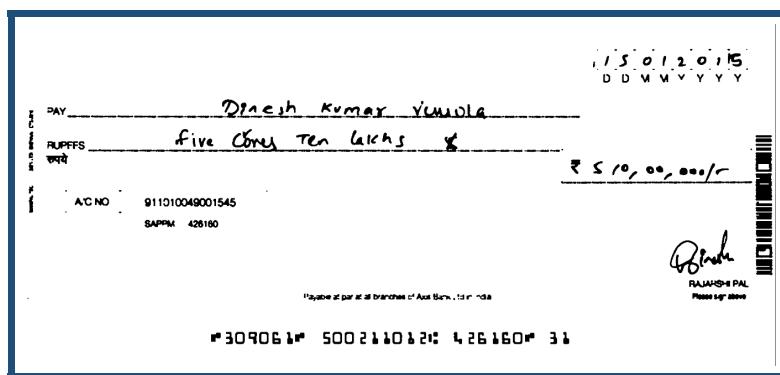


FIGURE 3.1 : Préparation des images

## 3.2 Prétraitement des Données

### 3.2.1 Segmentation des données

La segmentation et l'extraction sont des étapes cruciales dans le traitement des images de chèques, visant à isoler et à extraire spécifiquement les informations pertinentes telles que l'identifiant du chèque, les montants en lettres et en chiffres, le nom du client, et la date.

#### 3.2.1.1 Extraction du Montant en Chiffres

Cette étape consiste à extraire le montant numérique figurant sur le chèque.

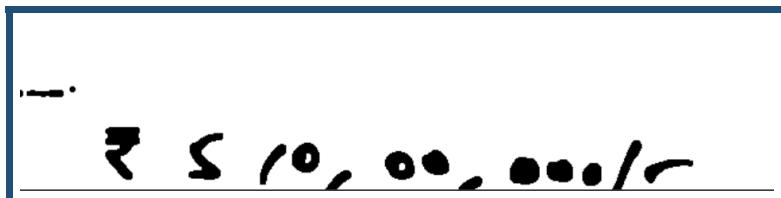


FIGURE 3.2 : Montant en Chiffres

#### 3.2.1.2 Extraction du Montant en Lettres

Cette étape consiste à extraire le montant écrit en lettres figurant sur le chèque.

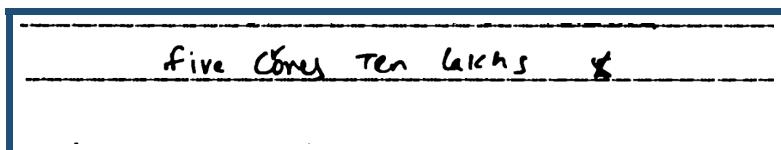


FIGURE 3.3 : Montant en Lettres

#### 3.2.1.3 Extraction du Nom du Client

Cette étape consiste à identifier et à extraire le nom du client mentionné sur le chèque.

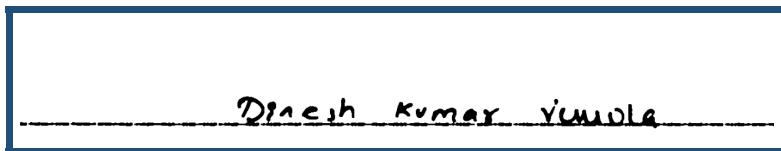


FIGURE 3.4 : Nom du Client

#### 3.2.1.4 Extraction de la Date

Cette étape consiste à identifier et à extraire la date indiquée sur le chèque.



FIGURE 3.5 : la date

### 3.2.1.5 Extraction de l'identifiant de chèque

Cette étape consiste à identifier et à extraire la date indiquée sur le chèque.



FIGURE 3.6 : l'identifiant de chèque

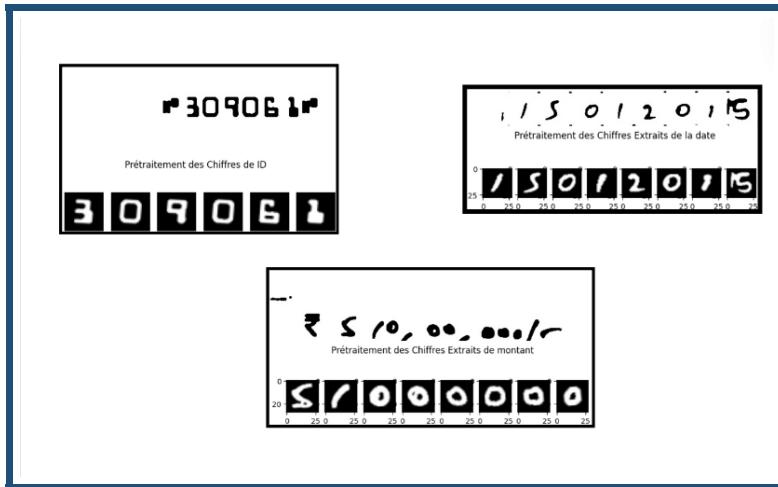
## 3.2.2 Extraction et Tri des Chiffres

L'extraction et le tri des chiffres sont des étapes essentielles dans le prétraitement des images de chèques, particulièrement pour les montants numériques. Cette section décrit le processus en détail en se basant sur les techniques de traitement d'image appliquées. Il est important de noter que cette partie concerne spécifiquement le traitement des chiffres, tandis que les textes en lettres sont traités séparément.

### 3.2.2.1 Prétraitement des Images de Chiffres

Le prétraitement est crucial pour isoler les chiffres de manière efficace. Voici les étapes suivies pour préparer les images contenant des chiffres : **Chargement et Conversion en Niveaux de Gris** qui fait éliminer les informations de couleur et laisser uniquement les variations d'intensité lumineuse nécessaires, **un flou gaussien** est appliqué pour réduire le bruit dans l'image facilitant ainsi la détection des contours, **détection des contours, extraction des chiffres** où chaque chiffre est recadré et redimensionné à une taille uniforme de 28x28 pixels.

Voici une illustration des chiffres après prétraitement, montrant les résultats finaux pour les montants, les identifiants, et les dates :



**FIGURE 3.7 :** Illustration des chiffres après prétraitement

Cette image présente les chiffres extraits après toutes les étapes de prétraitement, fournissant un aperçu clair des données prêtes pour la reconnaissance par le modèle.

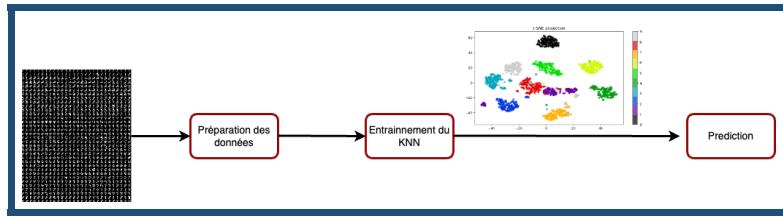
### 3.3 Extraction des Données

#### 3.3.1 k-Nearest Neighbors (KNN)

Le modèle k-Nearest Neighbors (KNN) est un algorithme de classification basé sur la similarité. Lorsqu'une nouvelle donnée doit être classifiée, le KNN recherche les K exemples les plus proches dans l'espace de caractéristiques et effectue une prédiction en fonction de la majorité des classes parmi ces voisins. Ce modèle est particulièrement adapté à la reconnaissance de chiffres manuscrits, car il peut identifier des motifs similaires dans des images prétraitées en utilisant des distances métriques.

##### 3.3.1.1 Modèle

Nous avons appliqué le KNN sur les chiffres extraits des images de chèques après un prétraitement approfondi. Ce choix a été motivé par la simplicité et l'efficacité du KNN pour des tâches de reconnaissance de chiffres où des données augmentées améliorent la robustesse du modèle.



**FIGURE 3.8 : Architecture de modèle KNN**

### 3.3.1.2 Évaluation

Dans cette section, nous analyserons les performances de notre modèle de classification KNN en utilisant plusieurs métriques d'évaluation.

- **Résumé des Performances**

Le résumé des performances d'un modèle de classification permet d'évaluer son efficacité à prédire les classes correctes à partir des données d'entrée. Les principales métriques utilisées pour résumer les performances comprennent l'Accuracy, la Précision, le Rappel et le F1-Score.

- **Accuracy (Exactitude)** : L'accuracy mesure la proportion d'échantillons correctement classés parmi l'ensemble des échantillons.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Résultat : L'accuracy obtenue est de 94,32%, indiquant que 94,32% des prédictions du modèle étaient correctes.

- **Précision (Precision)** : La précision évalue la proportion des prédictions positives qui sont correctes.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Résultat : La précision obtenue est de 0,94, ce qui signifie que 94% des prédictions positives étaient correctes.

- **Rappel (Recall)** : Le rappel mesure la capacité du modèle à détecter toutes les instances positives réelles.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Résultat : Le rappel obtenu est de 0,94, indiquant que 94% des échantillons positifs réels ont été correctement détectés.

- **F1-Score** : Le F1-score est une mesure combinée de la précision et du rappel, offrant une

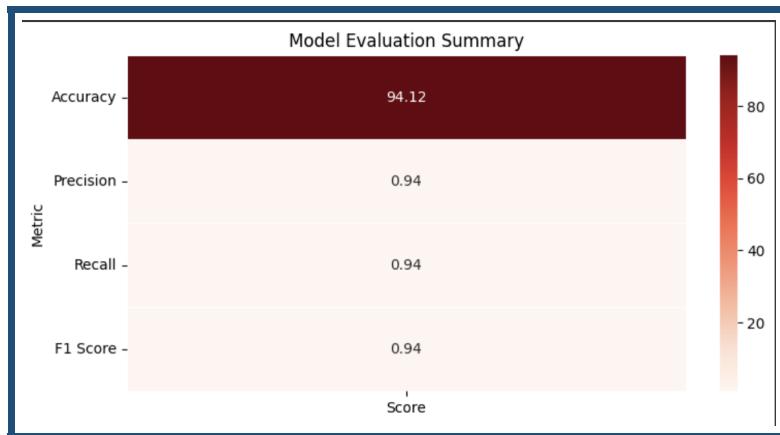
évaluation globale de la qualité des détections.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Résultat : Le F1-score obtenu est de 0,94, reflétant un équilibre solide entre la précision et le rappel.

Où :

- > **TP (True Positives)** : Les True Positives (TP) représentent le nombre d'échantillons correctement classifiés comme positifs par le modèle.
- > **FP (False Positives)** : Les False Positives (FP) représentent le nombre d'échantillons incorrectement classifiés comme positifs alors qu'ils sont en réalité négatifs.
- > **FN (False Negatives)** : Les False Negatives (FN) représentent le nombre d'échantillons incorrectement classifiés comme négatifs alors qu'ils sont réellement positifs.
- > **TN (True Negatives)** : Les True Negatives (TN) représentent le nombre d'échantillons correctement classifiés comme négatifs.



**FIGURE 3.9 :** Illustration de modèle PaddlOCR

#### • Rapport de Classification

Le rapport de classification fournit une vue d'ensemble des performances du modèle pour chaque classe. Il inclut des métriques telles que la précision, le rappel, le F1-score et le nombre de support (c'est-à-dire le nombre d'échantillons de chaque classe). Ce rapport est essentiel pour comprendre comment le modèle se comporte pour chaque classe individuelle.

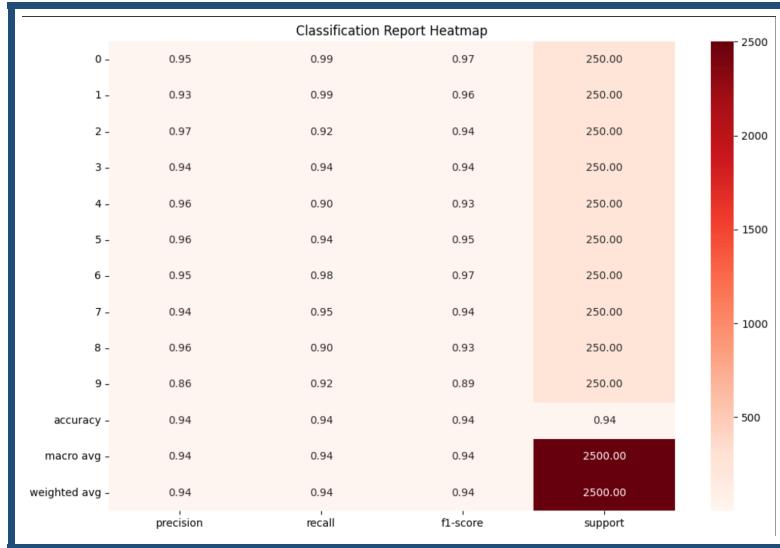


FIGURE 3.10 : Illustration de modèle PaddlOCR

- Matrice de Confusion

La matrice de confusion est un outil utilisé pour évaluer la performance d'un modèle de classification en comparant les prédictions du modèle aux valeurs réelles.

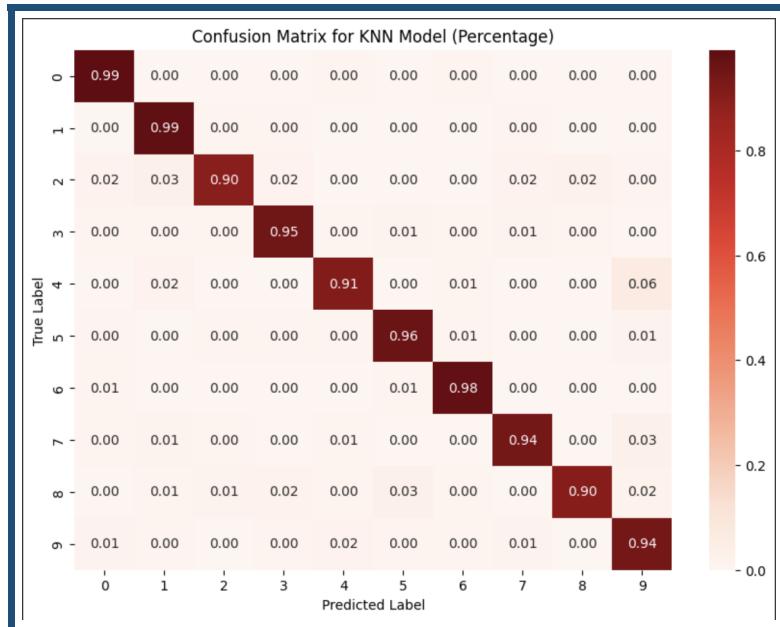


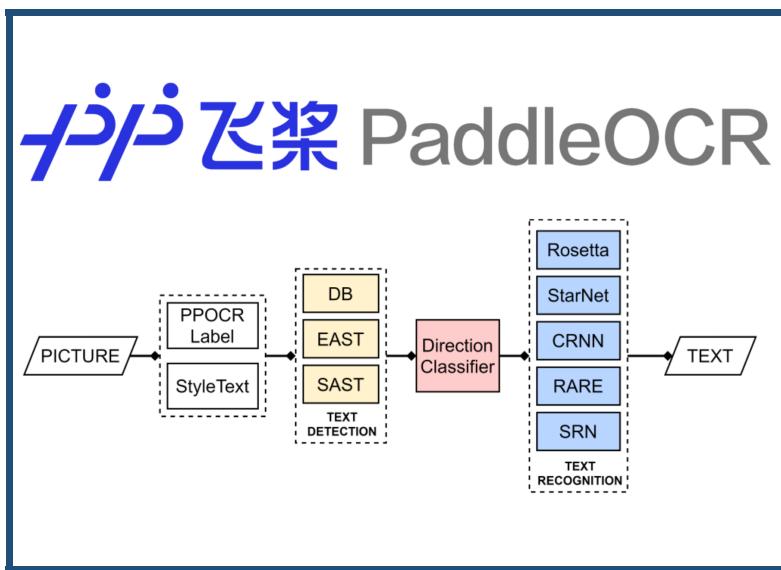
FIGURE 3.11 : Illustration de modèle PaddlOCR

Cette matrice de confusion présente les valeurs représentées par des nuances de couleurs. Par exemple, des couleurs plus vives ou plus foncées peuvent indiquer une détection plus précise du modèle, tandis que des couleurs plus claires peuvent indiquer une détection moins précise du modèle. Cette représentation visuelle permet de mettre en évidence les zones où le modèle

performe bien (valeurs élevées sur la diagonale principale) et les zones où le modèle a des difficultés à prédire correctement les classes (valeurs élevées en dehors de la diagonale principale).

### 3.3.2 Paddle OCR

PaddleOCR est une solution OCR basée sur le framework PaddlePaddle, spécialement conçue pour traiter des textes dans des images en utilisant des réseaux de neurones profonds. Ce modèle offre des fonctionnalités avancées telles que la détection de texte, la reconnaissance de texte et la correction d'orientation. Nous avons sélectionné PaddleOCR pour l'extraction des identifiants de chèques, car il excelle dans la détection et la reconnaissance de texte dans des contextes variés et complexes, tels que les documents manuscrits ou les images de chèques avec des formats différents. Sa précision et sa capacité à gérer divers styles de texte en font un choix optimal pour cette tâche.



**FIGURE 3.12 :** Illustration de modèle PaddleOCR

### 3.3.3 PyTesseract

PyTesseract est un wrapper pour le moteur OCR Tesseract, largement reconnu pour sa capacité à extraire du texte à partir d'images. Ce modèle utilise des techniques de traitement d'image avancées pour convertir des pixels en texte lisible. Nous avons utilisé PyTesseract pour extraire les montants écrits en lettres et les noms des clients sur les chèques, en raison de sa flexibilité et de son efficacité dans la reconnaissance de texte en utilisant des configurations pré-définies et des ajustements de seuils pour améliorer la précision. PyTesseract s'est révélé performant pour les tâches

où le texte est relativement clair et bien segmenté, ce qui est crucial pour la reconnaissance de texte sur les chèques.

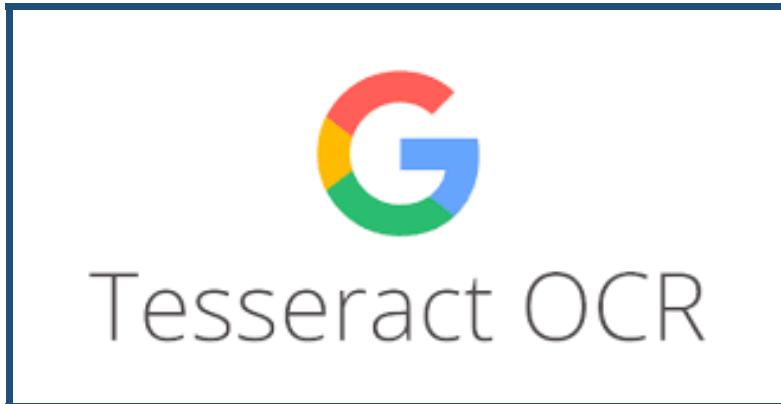


FIGURE 3.13 : Illustration de modèle Pytesseract de Tesseract OCR

### 3.3.4 Modèles Adaptés

Après une évaluation approfondie des différents modèles, nous avons sélectionné les suivants pour notre projet :

- **KNN pour la reconnaissance des chiffres** : Le modèle KNN a été choisi pour sa capacité à classifier efficacement les chiffres extraits des images de chèques après prétraitement. Sa simplicité et son efficacité dans la reconnaissance des chiffres manuscrits ont été des facteurs déterminants.
- **PyTesseract pour l'extraction des textes** : PyTesseract a été retenu pour extraire les montants écrits en lettres et les noms des clients, en raison de sa flexibilité et de son efficacité dans la reconnaissance de texte à partir d'images claires et bien segmentées.
- **PaddleOCR pour l'extraction des identifiants des chèques** : PaddleOCR a été sélectionné pour sa précision dans la détection et la reconnaissance des identifiants de chèques. Sa capacité à traiter des textes dans des contextes variés et complexes a été un atout majeur pour cette tâche.

Ces choix ont été guidés par la performance des modèles lors des tests, leur capacité à traiter efficacement les données spécifiques des chèques, et leur compatibilité avec les exigences de notre projet. KNN a montré une bonne capacité à classifier les chiffres après prétraitement, PyTesseract a prouvé son efficacité dans l'extraction de texte, et PaddleOCR s'est distingué par sa précision dans la détection des identifiants des chèques.

## Conclusion

En conclusion, ce chapitre a présenté les différentes techniques de préparation et de prétraitement des images de chèques. Nous avons abordé les étapes cruciales pour nettoyer et transformer ces images, rendant les données prêtes pour l'extraction et l'analyse. Le processus de prétraitement, incluant des méthodes comme le flou médian, la binarisation, et le seuillage adaptatif, a été essentiel pour améliorer la qualité des données. Cette préparation rigoureuse nous permet d'avancer avec confiance vers les phases d'extraction des données, où des outils comme KNN, Paddle OCR et PyTesseract seront utilisés pour une reconnaissance précise des informations.

# MODÉLISATION ET EVALUATION

---

## Plan

<b>Introduction</b> . . . . .	<b>29</b>
<b>1 Etude théorique</b> . . . . .	<b>29</b>
1.1 LSTM (Long Short-Term Memory) . . . . .	29
1.2 Seq2Seq (Sequence-to-Sequence) . . . . .	30
1.3 Mécanisme d'attention . . . . .	30
1.4 Transformers . . . . .	31
<b>2 Modélisation</b> . . . . .	<b>31</b>
2.1 Création et Prétraitement des Données . . . . .	32
2.2 Développement du Modèle . . . . .	33
<b>3 Évaluation du Modèle</b> . . . . .	<b>35</b>
3.1 Visualisations des Performances . . . . .	35
<b>Conclusion</b> . . . . .	<b>36</b>

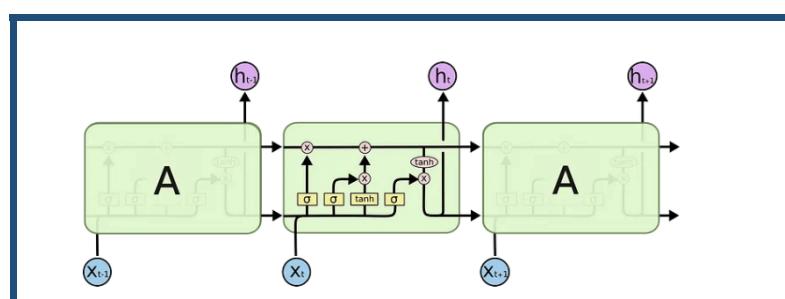
## Introduction

Après avoir préparé les données des chèques et réalisé les prétraitements nécessaires, nous entrons dans la phase de modélisation pour l'autocorrection des montants écrits en lettres. Cette étape est cruciale pour transformer efficacement les montants manuscrits en chiffres corrects. Nous explorerons et testerons différents modèles, notamment Seq2Seq, LSTM, et Transformers, afin d'identifier celui qui offre la meilleure performance en termes de précision et de robustesse pour la correction automatique des montants. L'objectif est de sélectionner le modèle le plus performant pour garantir une correction fiable et efficace des erreurs dans les chèques.

### 4.1 Etude théorique

#### 4.1.1 LSTM (Long Short-Term Memory)

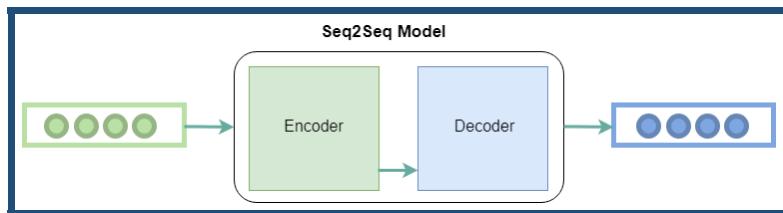
Long Short-Term Memory (LSTM) : LSTM est un type de réseau neuronal récurrent (RNN) utilisé en apprentissage en profondeur pour traiter des séquences de données. Conçu par Hochreiter et Schmidhuber, il a été développé pour résoudre le problème du "vanishing gradient" rencontré par les RNN traditionnels et les algorithmes d'apprentissage automatique. Une caractéristique distinctive de LSTM est l'incorporation de mécanismes de portes. Ces portes permettent de contrôler le flux d'informations dans le réseau, ce qui permet de préserver et de gérer efficacement les informations pertinentes sur de longues séquences. Contrairement aux RNN classiques, qui ont tendance à perdre de l'information sur de longues périodes, LSTM peut capturer des dépendances à long terme dans les données séquentielles. [3]



**FIGURE 4.1 :** Architecture LSTM

#### 4.1.2 Seq2Seq (Sequence-to-Sequence)

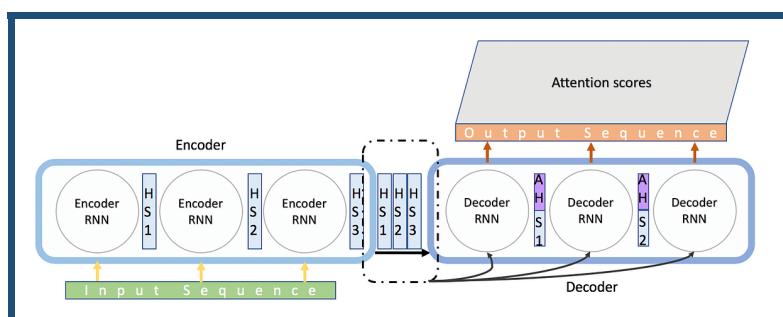
Le modèle Seq2Seq est une architecture de réseau de neurones conçue pour transformer une séquence d'entrée en une séquence de sortie. Le modèle Seq2Seq se compose de deux parties principales : l'encodeur et le décodeur. L'encodeur traite la séquence d'entrée et la convertit en une représentation vectorielle appelée "contexte", tandis que le décodeur génère la séquence de sortie à partir de cette représentation. Cette architecture est particulièrement adaptée aux tâches de traduction automatique et de génération de texte où les longueurs des séquences d'entrée et de sortie peuvent varier. ...[4]



**FIGURE 4.2 :** Architecture Seq2Seq

#### 4.1.3 Mécanisme d'attention

Le mécanisme d'attention est crucial dans de nombreuses architectures de réseaux neuronaux modernes, en particulier dans les tâches impliquant des séquences telles que le traitement du langage naturel et la vision par ordinateur. L'attention fonctionne en calculant un poids pour chaque élément de la séquence d'entrée, en fonction de sa pertinence pour l'élément actuellement généré par le décodeur. Ce poids est utilisé pour ajuster la contribution de chaque élément de la séquence d'entrée à la représentation contextuelle utilisée par le décodeur.

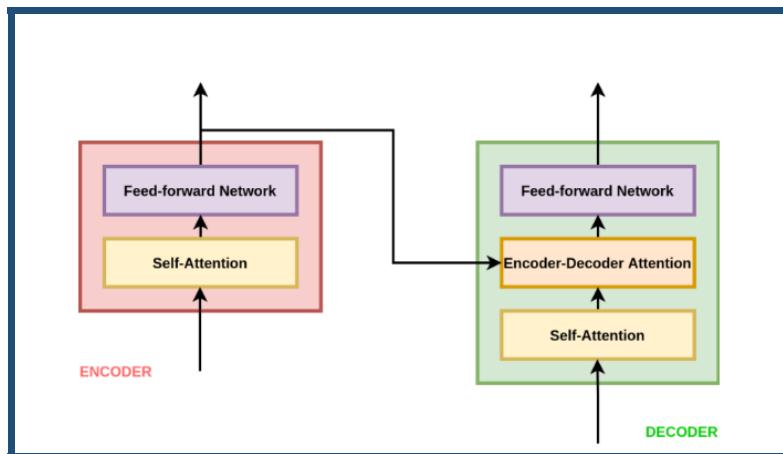


**FIGURE 4.3 :** Architecture de Mécanisme d'attention

#### 4.1.4 Transformers

Les transformateurs(Transformers) représentent une avancée majeure dans le domaine de l'apprentissage profond, ayant un impact significatif sur les modèles Seq2Seq, en particulier dans les tâches de traitement du langage naturel (NLP). Contrairement aux modèles Seq2Seq traditionnels basés sur des réseaux neuronaux récurrents (RNN), les transformateurs utilisent des mécanismes d'auto-attention pour capturer les dépendances à longue portée dans les séquences d'entrée. Cette approche permet aux transformateurs de traiter les séquences en parallèle et de modéliser des relations complexes avec une grande efficacité grâce à l'attention multi-têtes.

L'architecture des transformateurs améliore considérablement la capacité du modèle à comprendre et générer du texte avec une précision élevée. Ce gain en performance est crucial pour des tâches telles que la correction des montants en lettres, où une compréhension contextuelle précise est essentielle pour une correction efficace et fiable. [5]



**FIGURE 4.4 :** Architecture de Transformers

## 4.2 Modélisation

Dans cette section, nous détaillons les étapes de la modélisation de notre système de correction des montants en lettres. Cette phase est cruciale pour transformer les données prétraitées en un modèle performant capable de corriger automatiquement les erreurs dans les montants écrits en lettres.

### 4.2.1 Création et Prétraitement des Données

#### 1. Crédit du Dataset

Dans cette phase, nous avons généré un dataset synthétique destiné à l'entraînement de notre modèle de correction de texte. Ce dataset se compose de montants financiers présentés sous deux formes : correcte et erronée. Les montants numériques ont été convertis en texte écrit en Hindi, et des erreurs ont été introduites pour simuler des situations réelles où des erreurs peuvent se produire. Montants Corrects et Erronés

A. Montants Corrects : Les montants financiers ont été générés de manière incrémentielle, couvrant une large gamme de valeurs pour garantir une diversité représentative. Ces montants servent de référence pour évaluer les corrections apportées par le modèle.

B. Montants Erronés : Des erreurs ont été introduites de manière aléatoire pour simuler diverses fautes. Ces erreurs comprennent :

- **Erreurs Typographiques** : Des fautes de frappe aléatoires dans le texte, telles que des lettres manquantes ou des caractères incorrects.
- **Erreurs Sémantiques** : Modifications de mots courants, par exemple, en remplaçant des termes comme “thousand” par “thausand” pour simuler des erreurs de compréhension.
- **Erreurs de Formatage** : Suppression d'espaces, ajout de symboles incorrects ou autres erreurs de mise en forme.
- **Caractères Supplémentaires** : Insertion de mots ou de caractères non désirés pour simuler des erreurs de saisie, créant des textes perturbés ou incohérents.

En combinant ces montants corrects et erronés, nous avons pu créer un dataset robuste pour entraîner et évaluer notre modèle de correction de texte, visant à améliorer l'exactitude des montants extraits des chèques et à automatiser le processus de correction dans un contexte bancaire.

#### 2. Préparation des Données

Les données ont été préparées pour l'entraînement du modèle Seq2Seq. Cela a impliqué la tokenisation des textes corrects et erronés, ainsi que la transformation de ces textes en séquences numériques. Les séquences ont ensuite été remplies pour garantir une longueur uniforme, facilitant ainsi leur traitement par le modèle.

- Tokenisation : Les textes ont été convertis en séquences numériques à l'aide d'un tokeniser.
- Remplissage des Séquences : Les séquences ont été remplies pour garantir que toutes les entrées du modèle aient la même longueur, ce qui est crucial pour l'entraînement efficace du modèle.

## 4.2.2 Développement du Modèle

### 4.2.2.1 Choix du Modèle

Après une évaluation approfondie des différentes architectures, nous avons choisi le modèle **Seq2Seq (Sequence-to-Sequence)** pour notre tâche de correction automatique des montants en lettres. Cette décision est basée sur la capacité du modèle Seq2Seq à transformer efficacement des séquences d'entrée en séquences de sortie, ce qui est essentiel pour notre application où nous devons convertir des montants erronés en montants corrects. Comparé aux architectures LSTM et Transformers, Seq2Seq offre un bon compromis entre complexité et performance, avec une mise en œuvre plus simple et des exigences computationnelles réduites. Sa capacité à gérer les variations de longueur des séquences et son efficacité démontrée dans des tâches similaires, telles que la traduction automatique, ont renforcé notre choix pour cette architecture.

### 4.2.2.2 Construction et Entraînement

La construction et l'entraînement du modèle de correction automatique des montants écrits en lettres impliquent plusieurs étapes essentielles. L'architecture du modèle repose sur une combinaison de couches d'encodage, de décodage, et d'attention.

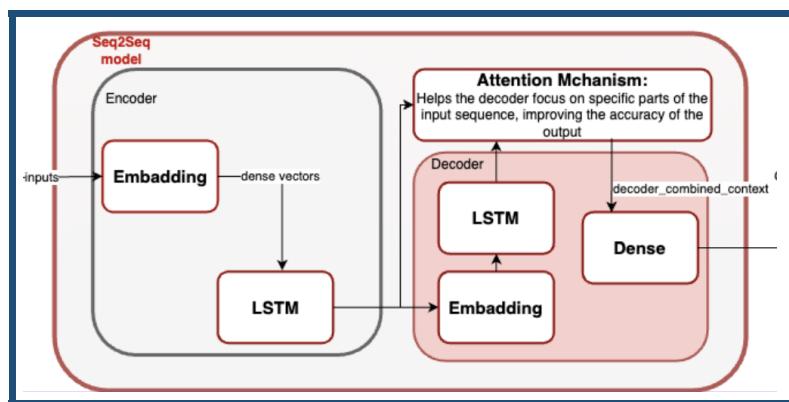


FIGURE 4.5 : Architecture du Modèle

L'encodeur joue un rôle crucial en capturant les caractéristiques sémantiques des séquences de

montants à corriger. Cette transformation est effectuée par une couche d'embedding qui représente chaque mot ou caractère en un vecteur dense, suivi d'une couche LSTM qui mémorise les dépendances séquentielles, assurant ainsi que les relations entre les différentes parties du montant (comme les unités, dizaines, centaines, etc.) sont bien comprises.

Le décodeur, quant à lui, reçoit ces représentations vectorielles et les traite pour générer la version corrigée du montant en lettres. Il commence par convertir les vecteurs en embeddings avant de les passer à travers un LSTM, qui génère la séquence de sortie, c'est-à-dire le montant corrigé. Une couche dense est ensuite utilisée pour affiner cette sortie en un texte bien formé et précis.

Le mécanisme d'attention ajoute une couche supplémentaire de précision en permettant au modèle de se concentrer sur les parties critiques de la séquence d'entrée lors de la génération de la sortie. Par exemple, s'il y a une confusion dans la transcription d'un chiffre ou d'une partie spécifique du montant, l'attention aide le modèle à identifier et corriger cette erreur en se focalisant sur les informations pertinentes fournies par l'encodeur.

#### 4.2.2.3 Architecture Complète d'Autocorrection

L'architecture complète du système d'autocorrection est conçue pour fournir une solution intégrée capable de traiter, corriger et valider les montants écrits en lettres extraits des chèques.

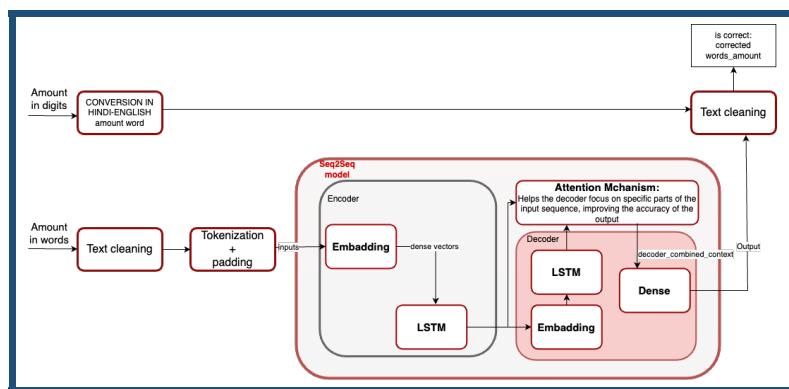


FIGURE 4.6 : Architecture Complète d'Autocorrection

L'architecture complète du système d'autocorrection est conçue pour traiter, corriger et valider les montants écrits en lettres extraits des chèques. Le processus commence par une étape de comparaison où le montant écrit en lettres est confronté au montant en chiffres pour vérifier leur cohérence. Cette comparaison initiale permet d'identifier les éventuelles erreurs ou incohérences dans le texte. Ensuite, le texte est soumis à un processus de correction automatisée, qui ajuste et reformule les montants en lettres pour garantir leur exactitude. Enfin, une validation finale est effectuée pour

s'assurer que le montant corrigé est correctement formaté et conforme aux attentes.

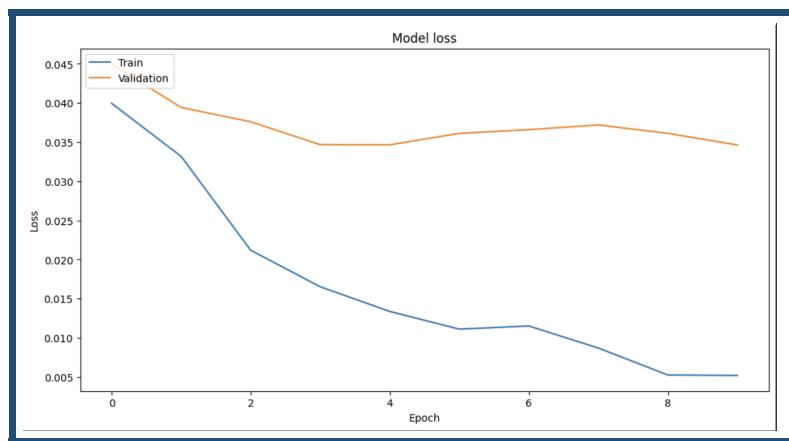
## 4.3 Évaluation du Modèle

Dans cette section, nous analysons les performances de notre modèle de correction des montants en lettres en utilisant plusieurs métriques d'évaluation. Cette analyse est essentielle pour comprendre l'efficacité du modèle et son aptitude à corriger automatiquement les erreurs dans les montants financiers écrits en lettres.

### 4.3.1 Visualisations des Performances

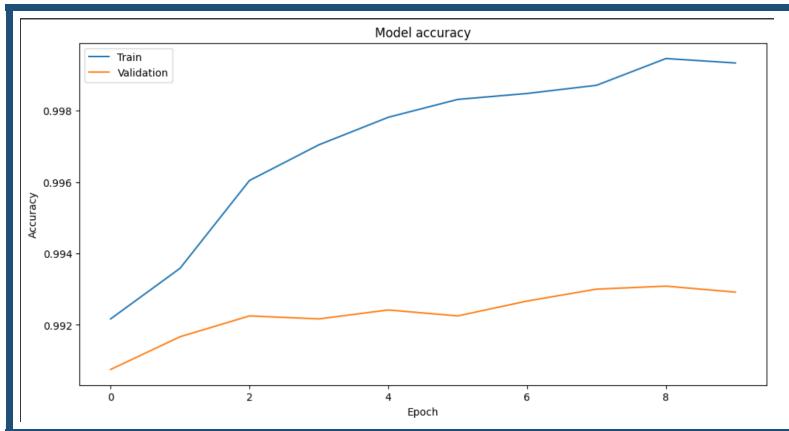
Pour visualiser les performances du modèle, nous avons généré les graphiques suivants :

Le graphique ci-dessus montre l'évolution de la précision du modèle au fil des époques pour les ensembles d'entraînement et de validation. Une bonne performance est indiquée par une précision croissante et stable pour les deux ensembles.



**FIGURE 4.7 :** Evolution de la précision du modèle

Le graphique ci-dessus illustre les valeurs de perte pour l'entraînement et la validation au fil des époques. Une perte décroissante indique que le modèle apprend efficacement et réduit les erreurs au fil du temps.



**FIGURE 4.8 :** Perte pour l'entraînement et la validation au fil des époques

Le tableau des scores pour Accuracy, Précision, Rappel, et F1-Score est présenté sous forme de carte thermique pour une visualisation claire des performances globales du modèle.

**Note :** Les graphiques et tableaux seront ajoutés ultérieurement en utilisant les outils appropriés pour leur génération et leur insertion.

## Conclusion

Après avoir évalué plusieurs modèles pour l'autocorrection des montants écrits en lettres, nous avons sélectionné le modèle Transformers comme solution optimale. Ce modèle a démontré une performance supérieure dans la correction des erreurs grâce à sa capacité à comprendre et à corriger les incohérences dans les montants manuscrits. En optant pour les Transformers, nous assurons une correction précise et fiable des montants sur les chèques, en tirant parti de leur puissance pour traiter et ajuster efficacement les erreurs dans les textes manuscrits. Ce choix garantit que notre solution est à la fois robuste et efficace pour les besoins de correction automatique.

# DÉPLOIEMENT

---

## Plan

<b>Introduction . . . . .</b>	<b>39</b>
<b>1 Les technologies utilisés . . . . .</b>	<b>39</b>
1.1 Python . . . . .	39
1.2 Django . . . . .	39
1.3 SQLite . . . . .	40
1.4 PowerBi . . . . .	40
<b>2 Pipeline de projet . . . . .</b>	<b>41</b>
2.1 Architecture du projet . . . . .	41
2.2 Sitemap du site web . . . . .	42
<b>3 Les interfaces web de l'application . . . . .</b>	<b>42</b>
3.1 Page d'accueil . . . . .	42
3.2 Page de connexion . . . . .	43
3.3 Page d'inscription . . . . .	43
3.4 Dashboard Administrateur . . . . .	44
3.5 Interface de téléchargement des chèques . . . . .	45
3.6 Interface de résultats du traitement des chèques . . . . .	45
3.7 Interface d'informations supplémentaires . . . . .	46
3.8 Table des Clients . . . . .	47
3.9 Table des Chèques . . . . .	47
3.10 Table des Employés . . . . .	48
3.11 Données des clients . . . . .	49
3.12 Données des chèques . . . . .	49
3.13 Données des employés . . . . .	50
3.14 Profil utilisateur . . . . .	51

Conclusion . . . . .	51
----------------------	----

## Introduction

Après avoir développé et testé notre modèle d'autocorrection des montants écrits en lettres, nous entamons la phase de déploiement, une étape cruciale pour la mise en production de notre application. Cette phase consiste à rendre notre solution pleinement opérationnelle, permettant son utilisation dans un environnement réel par les employés bancaires. Le déploiement implique l'utilisation de diverses technologies, la structuration d'un pipeline de projet efficace, et la création d'interfaces utilisateur intuitives pour assurer une expérience utilisateur fluide et performante.

### 5.1 Les technologies utilisées

#### 5.1.1 Python

Python est un langage de programmation polyvalent et puissant. C'est un excellent premier langage puisque le code Python est concis et facile à lire. Python a permis de gérer efficacement la logique de traitement des chèques, d'intégrer des modèles de traitement du langage naturel (NLP), et de faciliter l'interaction avec la base de données et les interfaces web via le framework Django.[6]



#### 5.1.2 Django

Django est un framework Web avancé écrit en Python qui utilise le modèle architectural du contrôleur de vue de modèle (MVC). Django a été créé dans un environnement de salle de presse en évolution rapide, et son objectif principal est de faciliter le développement de sites Web complexes et basés sur des bases de données. Ce framework Web a été initialement développé pour The World

Company pour la gestion de certains de leurs sites orientés actualités.[7]



### 5.1.3 SQLite

SQLite est un système de gestion de base de données relationnelle qui a été utilisé pour stocker les données de l'application, telles que les informations des utilisateurs, les détails des chèques, et les résultats des corrections automatiques. Sa simplicité et sa portabilité en font un choix idéal pour les projets de taille moyenne où une base de données légère et embarquée est nécessaire. SQLite s'intègre parfaitement avec Django, permettant une gestion transparente des données.[8]



### 5.1.4 PowerBi

Microsoft Power BI est une solution d'analyse de données développée par Microsoft. Elle offre la possibilité de créer des visualisations de données personnalisées et interactives.

Power BI fournit une interface conviviale qui permet aux utilisateurs de concevoir leurs propres rapports et tableaux de bord sans nécessiter de compétences techniques avancées. Ces tableaux de bord permettent aux utilisateurs d'analyser les données de traitement des chèques de manière visuelle et intuitive.[9]



## 5.2 Pipeline de projet

La mise en place d'une pipeline de projet bien structurée est essentielle pour garantir un développement fluide et un déploiement sans heurts. La pipeline de projet inclut l'architecture globale de l'application

### 5.2.1 Architecture du projet

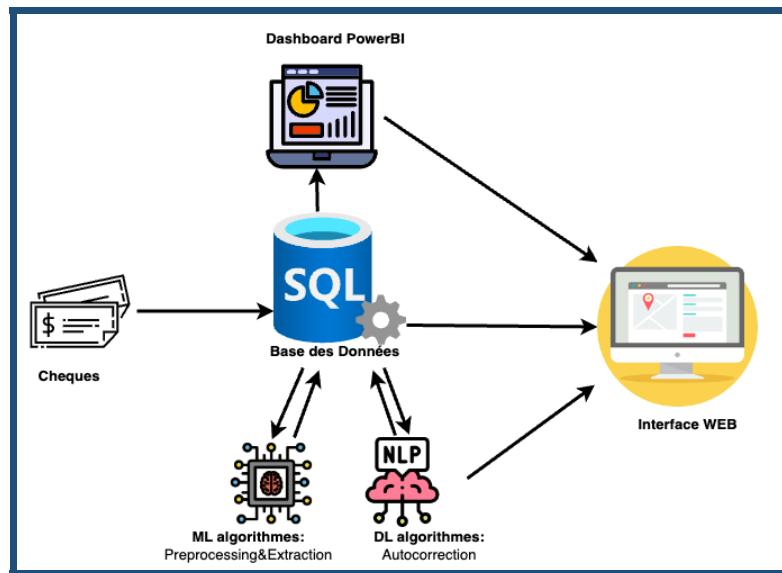
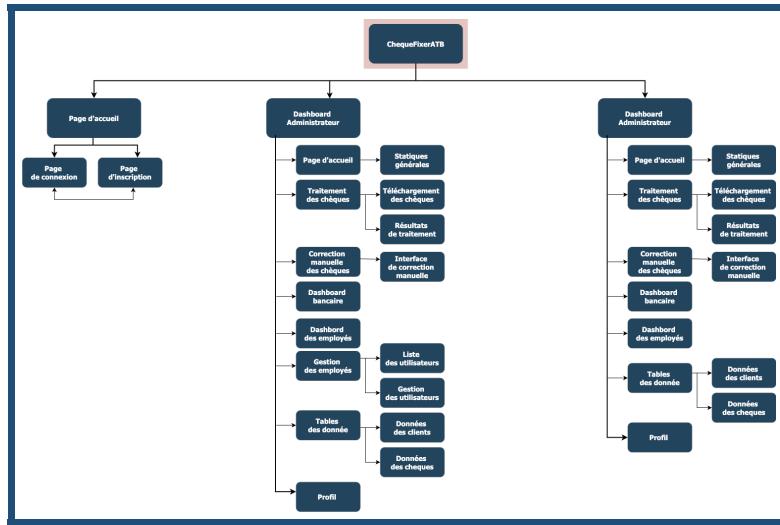


FIGURE 5.1 : Architecture du projet

La figure présente l'architecture modulaire de notre projet, où chaque composant interagit de manière cohérente pour assurer une gestion efficace des données. Les chèques sont collectés et stockés dans une base de données SQL, où des algorithmes de machine learning sont utilisés pour le prétraitement et l'extraction des données. Parallèlement, des techniques de deep learning, notamment en traitement du langage naturel (NLP), sont appliquées pour l'autocorrection des données textuelles. Ces informations sont ensuite accessibles via une interface web et visualisées dans un tableau de bord.

### 5.2.2 Sitemap du site web



**FIGURE 5.2 : Sitemap du site web**

La figure présente le sitemap du site web "ChequeFixerATB", illustrant la structure hiérarchique des différentes pages et leur organisation. Ce schéma met en évidence les principales sections, telles que la page d'accueil, les dashboards administrateurs, la gestion des chèques et des employés, ainsi que les interfaces de traitement et de correction manuelle. Cette vue d'ensemble permet de planifier efficacement la navigation sur le site, assurant ainsi que toutes les fonctionnalités essentielles soient facilement accessibles pour les utilisateurs. Ces tables de jointure sont utilisées pour le processus de matching, afin de relier les CVs des candidats aux offres d'emploi correspondantes.

## 5.3 Les interfaces web de l'application

### 5.3.1 Page d'accueil

La figure 5.3 présente l'interface de la page d'accueil, où les utilisateurs peuvent naviguer vers la page de connexion ou la page d'inscription.



FIGURE 5.3 : Page d'accueil

### 5.3.2 Page de connexion

La figure 5.4 présente l'interface de connexion, où l'utilisateur est invité à saisir son nom d'utilisateur et son mot de passe pour accéder à l'application.

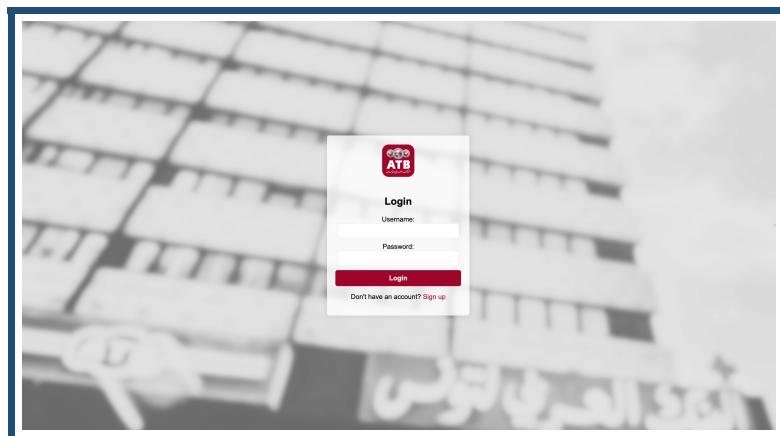
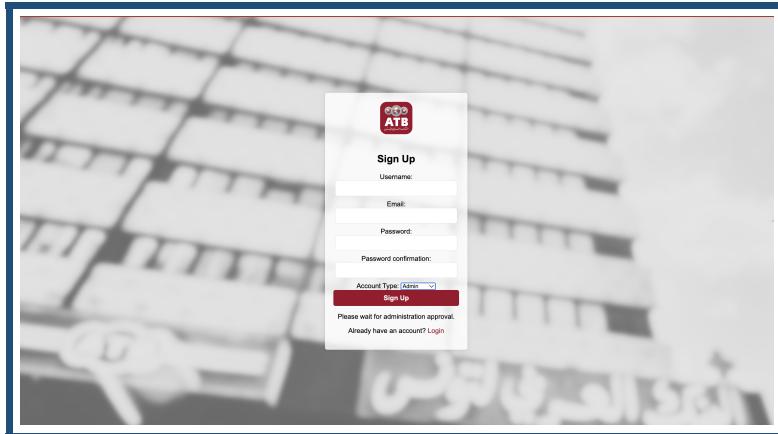


FIGURE 5.4 : Page de connexion

### 5.3.3 Page d'inscription

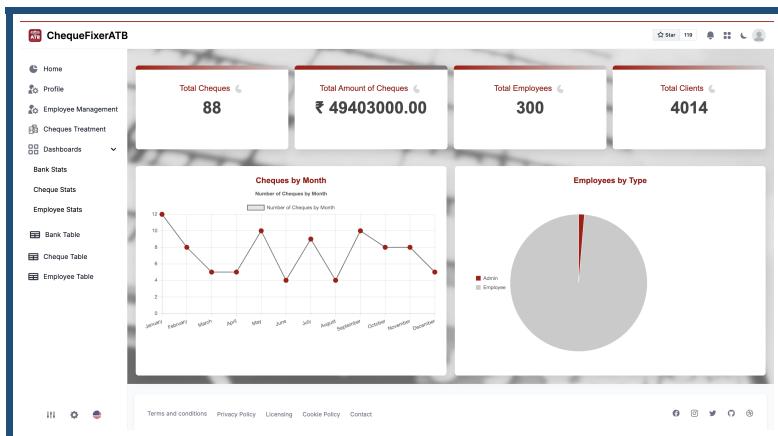
La figure 5.5 illustre l'interface du formulaire d'inscription, où l'utilisateur doit fournir son nom, son adresse e-mail, et créer un mot de passe.



**FIGURE 5.5 :** Page d'inscription

### 5.3.4 Dashboard Administrateur

La figure montre le tableau de bord de l'administrateur, où diverses fonctionnalités telles que la gestion des chèques, la visualisation des statistiques, et la correction manuelle des chèques sont accessibles.



**FIGURE 5.6 :** Dashboard Administrateur

La page d'accueil (home page) comporte plusieurs éléments clés, dont des indicateurs et des graphiques fournissant des informations récapitulatives sur les données importantes : le nombre total de chèques (88), le montant total des chèques ( 494,030,000.00), le nombre total d'employés (300) et le nombre total de clients (4014). Sous ces indicateurs principaux, un graphique linéaire montre l'évolution du nombre de chèques par mois tout au long de l'année, tandis qu'un graphique en secteurs illustre la répartition des employés entre administrateurs et employés ordinaires. Ces

éléments permettent à l'administrateur d'avoir une vue d'ensemble rapide et efficace des données essentielles pour la gestion quotidienne.

### 5.3.5 Interface de téléchargement des chèques

La figure montre l'interface de téléchargement des chèques. Cette interface permet aux utilisateurs de sélectionner et de télécharger les fichiers de chèques qu'ils souhaitent traiter. Les utilisateurs peuvent choisir des fichiers directement depuis leur appareil en cliquant sur le bouton "Upload".

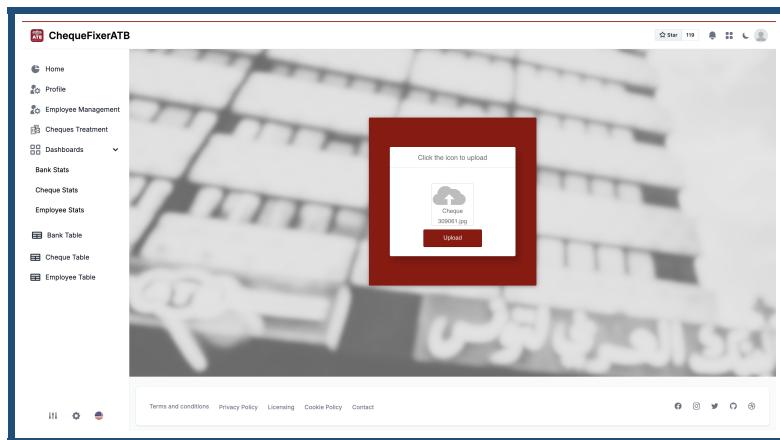
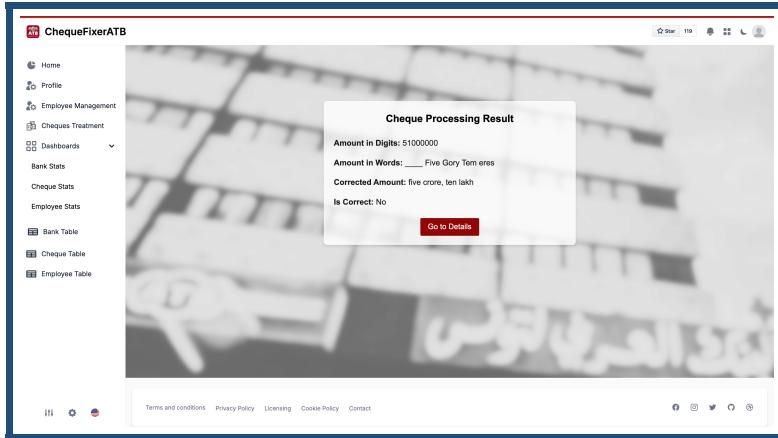


FIGURE 5.7 : Interface de téléchargement des chèques

### 5.3.6 Interface de résultats du traitement des chèques

La figure présente l'interface des résultats après traitement des chèques.

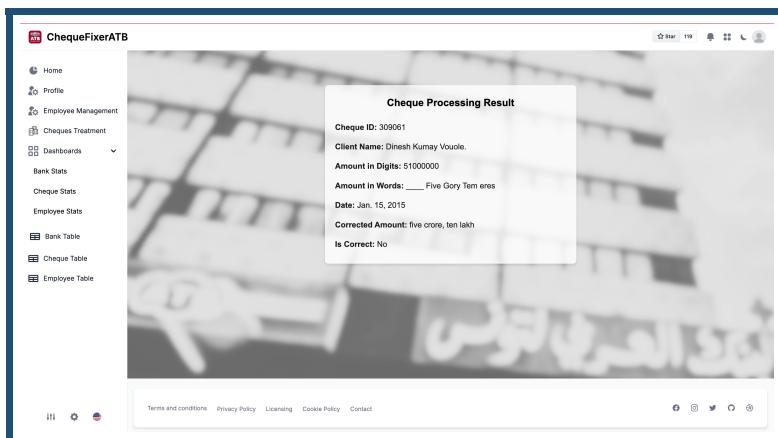


**FIGURE 5.8 :** Interface de résultats du traitement des chèques

Cette interface affiche les détails du chèque traité, notamment le montant en chiffres, le montant en lettres (avec une éventuelle correction), ainsi qu'une indication sur la correction nécessaire. Si une erreur est détectée, le montant corrigé est également affiché, et l'utilisateur peut cliquer sur "Go to Details" pour obtenir des informations supplémentaires.

### 5.3.7 Interface d'informations supplémentaires

La figure illustre l'interface des informations supplémentaires concernant les chèques traités.



**FIGURE 5.9 :** Interface d'informations supplémentaires

Cette interface fournit des détails plus spécifiques, notamment l'ID du chèque, le nom du client, la date du chèque, ainsi que les montants en chiffres et en lettres. Elle inclut également une version corrigée du montant si nécessaire et une indication de l'exactitude des informations.

### 5.3.8 Table des Clients

La figure montre la table des clients de la banque, où sont stockées les informations sur les clients de la banque.

CUSTOMER ID	NAME	SURNAME	GENDER	AGE	REGION	JOB CLASSIFICATION	DATE JOINED	BALANCE
100000001	Simon	Walsh	♂ Male	21	England	White Collar	Jan. 5, 2015	11881015 ↑
400000000	Jasmine	Miller	♀ Female	34	Northern Ireland	Blue Collar	Jan. 6, 2015	36919.73 ↑
100000003	Liam	Brown	♂ Male	46	England	White Collar	Jan. 7, 2015	50159.83 ↑
300000004	Trevor	Parr	♂ Male	32	Wales	Blue Collar	Jan. 8, 2015	1421.52 ↑
100000005	Denise	Pullman	♀ Female	38	England	White Collar	Jan. 9, 2015	35639.79 ↑
300000006	Ava	Coleman	♀ Female	30	Wales	Blue Collar	Jan. 9, 2015	12443.77 ↑
100000007	Dorothy	Thomson	♀ Female	34	England	Blue Collar	Jan. 11, 2015	4287.84 ↑
200000008	Lisa	Knox	♀ Female	48	Scotland	Other	Jan. 11, 2015	36680.17 ↑
300000009	Ruth	Campbell	♀ Female	33	Wales	White Collar	Jan. 11, 2015	74284.35 ↑
100000010	Dominic	Parr	♂ Male	42	England	White Collar	Jan. 12, 2015	10912.45 ↑
100000011	Dominic	Lewis	♂ Male	40	England	White Collar	Jan. 12, 2015	39661.83 ↑
100000012	Benjamin	Grant	♂ Male	39	England	White Collar	Jan. 12, 2015	32281.62 ↑
100000013	Lynn	MacDonald	♂ Male	24	England	White Collar	Jan. 12, 2015	40781.63 ↑
200000014	Thomas	Lawrence	♂ Male	46	Scotland	Other	Jan. 12, 2015	48791.46 ↑
300000015	Madeleine	Marshall	♀ Female	36	Wales	Other	Jan. 12, 2015	28461.03 ↑

FIGURE 5.10 : Table des Banques

Cette table contient des colonnes telles que l'identifiant unique du client, le prénom, le nom de famille, le sexe, l'âge, la région d'origine, la classification du travail (comme "White Collar" ou "Blue Collar"), la date d'inscription à la banque, et le solde du compte bancaire avec une indication de tendance. À gauche, un menu de navigation permet d'accéder à diverses fonctionnalités de l'application, y compris la gestion des employés, le traitement des chèques, et des tableaux de bord pour les statistiques des banques, des chèques et des employés.

### 5.3.9 Table des Chèques

La figure illustre la table des chèques, où sont enregistrées les données relatives aux chèques traités par l'application.

AMOUNT	DATE	AMOUNT IN WORDS	CLIENT NAME	CHEQUE ID
₹827000.00	Jan. 11, 2016	six lakh, twenty-seven thousand	Hysmo Salih Kumar.	309135
₹827000.00	Jan. 11, 2016	six lakh, twenty-seven thousand	Hysmo Salih Kumar.	309135
₹785000.00	Dec. 2, 2016	seven lakh, eighty-five thousand	Jean Kamat Rale	309136
₹499000.00	May 8, 2015	four lakh, thirty-nine thousand	Jean he. Kapesh Kumar.	309137
₹729000.00	May 20, 2015	seven lakh, thirty thousand	Mithun Adhel Kumar	309138
₹877000.00	March 11, 2016	eight lakh, seventy-seven thousand	Kumar Vinay Singh	309139
₹604000.00	Oct. 22, 2016	six lakh, four thousand	Anil sharma	309141
₹205000.00	Feb. 27, 2016	two lakh, five thousand	Bomaborts'Yahud Singh	309142
₹67000.00	Jan. 25, 2016	one lakh, sixty-seven thousand		309144
₹765000.00	March 21, 2016	seven lakh, sixty thousand	Naroth Th Shyam Kao	309145
₹746000.00	Jan. 6, 2016	seven lakh, forty thousand	Dath Karna Michra	309147
₹378000.00	July 26, 2015	three lakh, seventy-eight thousand	Ramini Kepila Kumar	309148
₹163000.00	July 16, 2015	one lakh, sixty-three thousand	SHRI RAD SINGH	309149
₹931000.00	March 7, 2016	nine lakh, thirty-one thousand	bowl stay?	309150
₹251000.00	Oct. 30, 2015	two lakh, fifty-three thousand	ew bh born kum	309151

FIGURE 5.11 : Table des Chèques

Cette table comprend plusieurs colonnes, telles que le montant du chèque, la date de traitement, le montant écrit en toutes lettres, le nom du client, et l'identifiant du chèque. Ces informations permettent de suivre et de vérifier les détails des chèques dans le cadre du processus de traitement automatisé par l'application.

### 5.3.10 Table des Employés

La figure présente la table des employés, où sont consignées les informations sur les membres du personnel.

EMPLOYEE ID	NAME	GENDER	USER TYPE	SIGN-IN DATE	SIGN-IN TIME	SIGN-OUT TIME	IP ADDRESS	CONNECTION STATUS	JOB POSITION	PHONE
65893	Joseph Ortiz	Male	Employee	Aug. 11, 2024	5:53 a.m.	1:47 p.m.	218.3.6.795	Failed	Manager	+216-35400041
6848	Rachel Wright	Female	Admin	Feb. 22, 2024	9:50 a.m.	9:57 a.m.	155.20.99.61	Failed	Analyst	+216-49545877
62563	James Allen	Male	Employee	Jan. 14, 2024	6:56 p.m.	5:38 p.m.	123.11.184.154	Failed	Developer	+216-58038374
92952	Stephen Warren	Male	Employee	July 19, 2024	6:20 a.m.	12:29 p.m.	211.4.231.228	Success	HR	+216-28480236
76207	Melanie McGee	Female	Employee	July 23, 2024	10:36 p.m.	1:21 p.m.	2014.150.252	Success	HR	+216-26863688
10714	Adam Burgess	Male	Admin	April 11, 2024	10:10 p.m.	11:19 a.m.	177.181.149.164	Success	Analyst	+216-64129310
59062	Michele Prince	Female	Employee	March 9, 2024	12:38 a.m.	5:21 a.m.	76.198.94.58	Failed	Developer	+216-25548482
18261	Karen Phillips	Female	Employee	April 27, 2024	3:09 p.m.	4:01 p.m.	121.88.219.83	Success	HR	+216-68493363
69695	Paula Hunter	Female	Employee	July 7, 2024	8:53 p.m.	3:27 a.m.	144.104.25.47	Success	Developer	+216-80347103
28187	Todd Lopez	Male	Employee	June 6, 2024	12:19 p.m.	6:01 p.m.	210.20.780.17	Success	Analyst	+216-09036075
22569	Adrian Schmidt	Male	Employee	July 23, 2024	3 a.m.	1:15 p.m.	142.130.30.126	Failed	Analyst	+216-27299869
4109	Matthew Wong	Male	Employee	April 14, 2024	12:42 p.m.	4:19 a.m.	93.169.66.114	Success	HR	+216-23012547
98013	Christopher Rodriguez	Male	Admin	June 5, 2024	11:35 p.m.	6:19 a.m.	101.10.208.194	Failed	Analyst	+216-29621912
55806	Deborah Glover	Female	Admin	May 12, 2024	11:01 a.m.	10:53 p.m.	40.251.249.107	Failed	Analyst	+216-9026670
29506	Amber Gutierrez	Female	Employee	Feb. 8, 2024	8:30 p.m.	6:03 p.m.	213.0.42.34	Failed	HR	+216-46680782

FIGURE 5.12 : Table des Employés

Cette table inclut diverses colonnes telles que l'ID de l'employé, le nom, le genre, le type d'utilisateur, les heures de connexion et de déconnexion, l'adresse IP, le statut de connexion, le poste occupé, ainsi que le numéro de téléphone. Ces informations permettent de suivre les activités

et les détails des employés au sein de l'application, facilitant ainsi la gestion des ressources humaines.

### 5.3.11 Données des clients

La figure illustre les dashboards Power BI relatifs aux données des clients.



**FIGURE 5.13 :** Données des clients

Ce tableau de bord, intitulé "ChequeFixerATB", regroupe plusieurs indicateurs clés de performance, notamment le nombre total de clients, ventilé par sexe, ainsi que des visualisations détaillées sur la classification des clients selon leur emploi, leur répartition géographique et par groupe d'âge. Un graphique linéaire montre l'évolution du nombre de clients au fil du temps, tandis qu'une carte du Royaume-Uni illustre leur distribution par région. La section dédiée aux cinq principaux clients par solde bancaire est mise en évidence, ainsi qu'un graphique à barres qui analyse le solde total en fonction du groupe d'âge et du sexe. Ce tableau de bord interactif permet une analyse approfondie des données clients, facilitant ainsi la prise de décisions stratégiques pour l'entreprise.

### 5.3.12 Données des chèques

La figure 5.14 illustre les dashboards Power BI relatifs aux données des chèques traités par l'application.



FIGURE 5.14 : Données des chèques

Ce tableau de bord affiche plusieurs indicateurs essentiels, tels que le nombre total de chèques, la somme totale des montants, et le nombre de clients concernés. Des graphiques interactifs permettent de visualiser la répartition des montants moyens par année, ainsi que le montant total par client. Un graphique linéaire montre l'évolution des montants moyens, maximums, et minimums par mois, tandis qu'un autre graphique illustre le nombre de chèques émis par mois. La liste des clients avec leurs noms, année et mois de traitement des chèques est également mise en évidence, permettant une analyse fine des transactions par période. Ce tableau de bord facilite la compréhension des tendances liées aux chèques, offrant un outil puissant pour la gestion et l'analyse des données bancaires.

### 5.3.13 Données des employés

La figure 5.15 illustre les dashboards Power BI relatifs aux données des employés.

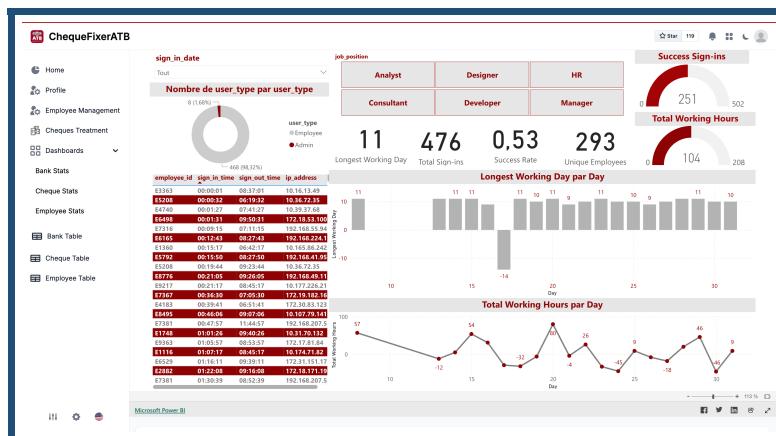


FIGURE 5.15 : Données des employés

Ce tableau de bord fournit des informations détaillées sur les connexions des employés, y compris le nombre total de connexions réussies, le taux de succès, le nombre unique d'employés, et les heures de travail totales. Des graphiques interactifs montrent la répartition des types d'utilisateurs, les postes occupés, et les jours de travail les plus longs. Une liste détaillée des employés, incluant l'identifiant de l'employé, l'heure de connexion et de déconnexion, ainsi que l'adresse IP, est également affichée, facilitant ainsi le suivi des activités des employés et l'analyse des heures de travail. Ce tableau de bord permet une gestion efficace et une surveillance des performances des employés au sein de l'organisation.

### 5.3.14 Profil utilisateur

La figure illustre l'interface de gestion du profil utilisateur, où chaque utilisateur peut modifier ses informations personnelles et ses préférences.

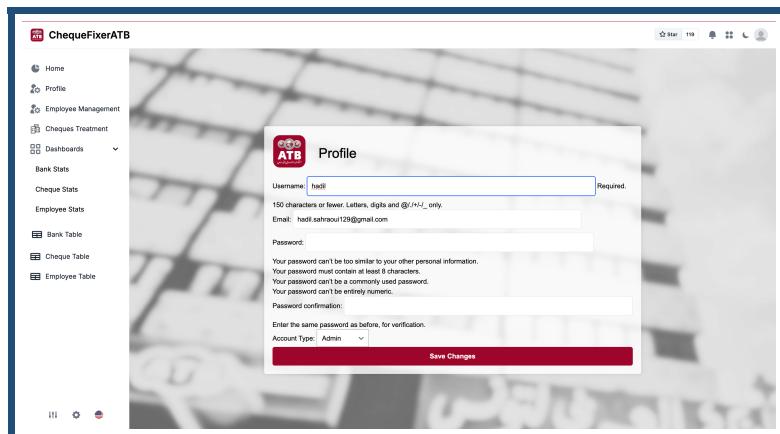


FIGURE 5.16 : Profil utilisateur

## Conclusion

Dans ce chapitre, nous avons abordé la procédure de déploiement. Nous avons commencé par décrire les technologies utilisées. Nous avons ensuite abordé le pipeline de déploiement du modèle. Enfin, nous avons présenté les interfaces de l'application web.

# Conclusion générale et perspectives

Au cours des dernières années, le secteur bancaire a connu une transformation numérique notable, avec une attention croissante à l'automatisation des processus pour améliorer l'efficacité et réduire les erreurs humaines. Le traitement manuel des chèques, longtemps en usage, est particulièrement concerné par ces avancées technologiques.

Notre projet vise à développer un système d'autocorrection des montants inscrits sur les chèques bancaires pour la Banque Arabe Tunisienne (ATB). En utilisant la reconnaissance optique de caractères (OCR) et des modèles d'apprentissage automatique, ce système détecte et corrige automatiquement les erreurs, améliorant ainsi la précision et la rapidité du traitement des chèques.

Les résultats montrent que cette solution accroît l'efficacité opérationnelle et réduit le risque d'erreurs, tout en améliorant la satisfaction des clients.

Pour l'avenir, l'intégration de technologies plus avancées, comme le deep learning, pourrait encore augmenter la précision du système. De plus, une plateforme centralisée de gestion des chèques renforcerait la supervision des opérations.

En conclusion, l'automatisation du traitement des chèques représente un progrès majeur pour ATB, offrant des gains en efficacité tout en ouvrant la voie à de futures améliorations pour répondre aux besoins évolutifs du secteur bancaire.

