SCI 1402

Projet en science des données

Rapport de projet

(4 pages maximum)

* Remplissez soigneusement les informations ci-dessous.
* Commencez votre travail à la page suivante.
* Sauvegardez votre travail de cette façon :  
  **SCI1402-RapportProjet\_Votrenom.docx**
* Utilisez l’outil de dépôt des travaux (accessible par votre portail étudiant [MaTÉLUQ](https://ma.teluq.ca/)) pour acheminer votre travail à la personne responsable de votre encadrement

Tiokap Guemdjou

|  |  |
| --- | --- |
| Numéro d'étudiant : | 23315841 |

Trimestre d’été 2025

## Présentation générale du projet

Nom du projet sélectionné

**L’essor de l’IA : une étude sur l’adoption technologique à l’échelle mondiale**

Motivation lors de la sélection du projet (pourquoi ce projet)

Lors du choix du projet, je souhaitais traiter d’un sujet à la fois actuel et porteur d’avenir. L’intelligence artificielle (IA) est aujourd’hui un domaine en pleine expansion et son adoption par les pays et les organisations transforme profondément les sociétés. En tant qu’étudiante en science des données, analyser ces tendances me semblait une excellente occasion de mettre en pratique les méthodes vues en cours (collecte et traitement des données massives, visualisation, analyses statistiques) tout en approfondissant un thème qui me passionne et qui est très présent dans mon environnement professionnel et académique.

Définition du projet (en quoi il consiste)

Ce projet consiste à collecter, traiter et analyser des données sur l’adoption de l’IA dans le monde. L’objectif est d’identifier les tendances globales, les écarts entre régions, les principaux facteurs explicatifs et les perspectives d’évolution.  
Concrètement, j’ai travaillé à partir d’un jeu de données provenant de Kaggle et d’autres sources publiques, puis j’ai réalisé les analyses et les visualisations dans Google Colab. Le projet inclut aussi la création d’un tableau de bord et la mise en ligne du produit final sur GitHub*.*

**Rappel des objectifs visés lors de la sélection du projet**

* Mettre en œuvre les compétences acquises pendant la formation (collecte, préparation, analyse et visualisation de données massives).
* Comprendre les tendances d’adoption de l’IA à l’échelle mondiale.
* Identifier les facteurs (économiques…) pouvant influencer cette adoption.
* Produire un tableau de bord interactif permettant de visualiser les résultats.
* Déposer le produit final (code + analyses) sur GitHub.

**Planification du projet (Fournir une liste de tâches en lien avec les objectifs visés, un échéancier sous forme d’un** [**Gantt**](https://www.gantt.com/fr/) **incluant chaque tâche et une identification des risques.)**

Dans le cadre de la planification, j’ai élaboré un échéancier couvrant la période du 18 juin 2025 au 22 septembre 2025. Cet échéancier détaille les étapes de travail, les livrables attendus ainsi que les principaux risques identifiés et les mesures prévues pour y répondre. Cette planification m’a permis d’organiser mes tâches de façon réaliste tout en anticipant les imprévus.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Période** | **Tâche principale** | **Livrables / Commentaires** | **Risques identifiés** | **Mesures d’atténuation** |
| **18 – 24 juin 2025** | Recherche et sélection des données | Identification et téléchargement des datasets pertinents | Données incomplètes ou non disponibles | Prévoir plusieurs sources de données de rechange (Kaggle, Open Data, etc.) |
| **25 juin – 2 juillet 2025** | Importation et exploration sur Google Colab | Vérification du format, aperçu des variables | Problèmes de compatibilité avec Colab (librairies manquantes) | Installer manuellement les dépendances / utiliser alternatives |
| **3 – 10 juillet 2025** | Nettoyage et transformation des données | Traitement des valeurs manquantes, formats cohérents | Temps plus long que prévu pour le nettoyage | Répartir la tâche sur plusieurs jours, documenter chaque étape |
| **11 – 18 juillet 2025** | Analyses descriptives et premières visualisations | Statistiques descriptives et graphiques simples | Visualisations peu parlantes | Tester plusieurs types de graphiques et choisir les plus clairs |
| **19 – 26 juillet 2025** | Identification des tendances et facteurs explicatifs | Corrélations, anomalies, sous-groupes | Mauvaise interprétation des tendances | Valider les résultats avec le professeur |
| **27 juillet – 3 août 2025** | Création du tableau de bord (version initiale) | Première version interactive du dashboard | Problèmes techniques avec l’outil de dashboard | Choisir un outil stable (Streamlit/Plotly) et meme collab sauvegarder le code |
| **4 – 11 août 2025** | Tests et amélioration du tableau de bord | Ajustements ergonomiques et ajout de filtres | Retards dus aux bugs | Planifier des jours pour corriger les bugs |
| **12 – 19 août 2025** | Rédaction du rapport (version préliminaire) | Remplir le gabarit fourni avec les premières sections | Risque d’oubli d’informations ou incohérences | Utiliser un plan détaillé et relire chaque section |
| **20 – 27 août 2025** | Ajustements et relecture du rapport | Vérification de la cohérence et du français | Manque de temps pour la relecture complète | Réserver au moins une journée dédiée à la relecture |
| **28 août – 4 septembre 2025** | Finalisation du tableau de bord et du code | Préparer la version finale | Perte de données ou de code | Sauvegarde régulière sur GitHub et Google Drive |
| **5 – 12 septembre 2025** | Dépôt du projet sur GitHub et intégration du lien | Vérification du dépôt, des notebooks et du dashboard | Problème d’accès au dépôt | Créer le dépôt tôt et tester le lien avant remise |
| **13 – 20 septembre 2025** | Préparation de la remise finale et vérification des critères d’évaluation | Derniers ajustements et validation | Oublier un livrable demandé | Utiliser une checklist basée sur les critères du cours |
| **21 – 22 septembre 2025** | Remise officielle du projet et présentation finale | Livraison complète du rapport, du code et du tableau de bord | Problème technique de dernière minute (upload) | Faire la remise au moins 24h avant l’échéance officielle (ce qui n’a pas ete fait d’ailleurs) |

Évaluation de la performance de la réalisation (Pour établir des critères pour évaluer la performance de la réalisation de votre projet, vous pouvez consulter le lien suivant : <https://www.inspq.qc.ca/en/node/6352>.)

Pour évaluer la performance de mon projet, j’ai choisi plusieurs indicateurs inspirés des recommandations de l’INSPQ. Ces indicateurs couvrent l’avancement, l’utilisation des ressources, la qualité du partenariat et l’atteinte des résultats. Ils m’ont permis de suivre l’évolution du travail et d’ajuster mes actions en cours de route.

* Indicateurs d’avancement du projet
* Écart entre l’échéancier planifié et le calendrier réel de production (retards ou avances pour chaque tâche).
* Taux de réalisation des livrables (nombre de tâches complétées / nombre de tâches prévues).
* Indicateurs d’utilisation des ressources
* Écart entre le nombre d’heures planifiées et le nombre d’heures réellement travaillées par semaine.
* Respect de l’enveloppe de ressources informatiques (ex. quotas Google Colab, stockage GitHub).
* Indicateurs liés a la collaboration
* Satisfaction de l’encadrant vis-à-vis la progression (feedback lors des suivis).
* Degré de participation aux échanges (réponses aux questions, mises à jour régulières sur GitHub).
* Indicateurs de résultats
* Disponibilité et accessibilité du tableau de bord interactif (hébergement et lien GitHub fonctionnel).
* Taux de complétude du rapport final (sections du gabarit remplies et conformes aux attentes).
* Qualité du code (exécution sans erreur sur Google Colab).
* Satisfaction personnelle face aux objectifs initiaux (auto-évaluation qualitative).
* Indicateurs d’effets collatéraux
* Mesure du niveau de stress et d’organisation personnelle pendant le projet (auto-évaluation).

Ces indicateurs m’ont permis de comparer régulièrement ce qui était prévu à ce qui était effectivement réalisé et d’adopter des mesures correctives (par exemple, réorganiser mon temps, corriger des erreurs plus tôt ou demander de l’aide).

## Explication des écarts de la réalisation par rapport à la planification

Rappel des résultats réellement atteints avec les objectifs initiaux.

Lorsque j’ai planifié le projet (18 juin – 18 septembre 2025), j’avais établi un échéancier précis avec des étapes successives : collecte et nettoyage des données, analyse exploratoire, modélisation, construction du tableau de bord, puis rédaction du rapport final et dépôt sur GitHub.

Dans l’ensemble, le déroulement du projet a respecté l’échéancier global, mais j’ai rencontré plusieurs décalages ponctuels :

* Retards dans la préparation des données  
  Initialement prévue sur deux semaines, la phase de nettoyage a pris près de trois semaines. La cause principale a été l’hétérogénéité et l’incomplétude des données (valeurs manquantes, formats différents). J’ai dû tester plusieurs stratégies d’imputation avant de pouvoir obtenir un jeu de données exploitable.
* Modélisation moins avancée que prévu  
  Je prévoyais tester plusieurs algorithmes prédictifs (forêt aléatoire, XGBost), mais le manque de données complètes par pays a limité ces expérimentations. J’ai donc concentré mes efforts sur une régression linéaire interprétable et quelques visualisations exploratoires.
* Tableau de bord livré avec un léger décalage  
  La mise en ligne du tableau de bord interactif sur GitHub a été finalisée une semaine plus tard que prévu. Cela est dû à la nécessité de revoir l’organisation des fichiers et des scripts pour que tout s’exécute correctement dans Google Colab.

Malgré ces écarts ponctuels, les objectifs essentiels ont été atteints :

* Constitution d’un jeu de données exploitable combinant des indicateurs économiques, technologiques pour une vingtaine de pays.
* Analyse exploratoire complète avec statistiques descriptives et visualisations (histogrammes, boxplots, scatter plots).
* Première modélisation prédictive (régression linéaire) permettant d’estimer le score d’adoption de l’IA à partir des variables socio-économiques.
* Création et mise en ligne d’un tableau de bord interactif présentant les indicateurs clés et les premiers résultats (lien GitHub fourni dans le rapport).
* Rédaction progressive du rapport dans le gabarit officiel SCI 1402, intégrant les résultats et l’analyse critique.

En résumé, bien que certaines tâches aient pris plus de temps que prévu et que certains modèles avancés n’aient pas pu être testés, le cœur du projet analyse a été réalisé.

Justification des choix faits en cours de réalisation :

*Approche ou méthodes*

Pour ce projet sur l’adoption de l’intelligence artificielle à l’échelle mondiale, j’ai choisi une approche analytique combinant collecte de données quantitatives et analyse comparative. Cette méthode permet de comprendre les tendances générales, tout en identifiant les facteurs socio-économiques et technologiques influençant l’adoption de l’IA. L’approche adoptée est également progressive, c’est-à-dire que j’ai commencé par une exploration des données, puis une préparation et un nettoyage rigoureux avant de passer aux analyses et visualisations, ce qui garantit la fiabilité et la cohérence des résultats. Enfin, nous avons appliqué un clustering afin de regrouper les pays présentant des profils similaires en matière d’adoption de l’IA, ce qui a permis d’identifier des patterns et des segments pertinents pour l’analyse.

***Techniques appliquées***

Plusieurs techniques ont été appliquées pour traiter et analyser les données :

* Exploration des données (Data Exploration) : pour comprendre la structure des fichiers, détecter les valeurs manquantes et identifier les anomalies potentielles.
* Nettoyage et transformation des données (Data Cleaning) : indispensable pour garantir des analyses précises et éviter les biais.
* Visualisation des données (Data Visualization) : utilisation de graphiques et de diagrammes pour représenter clairement les tendances et faciliter l’interprétation des résultats.
* Analyse comparative et corrélationnelle : afin de mettre en évidence les relations entre différents indicateurs (ex. : PIB, niveau de développement technologique) et l’adoption de l’IA.

Ces techniques ont été choisies pour leur efficacité et leur capacité à rendre l’analyse compréhensible.

***Outils utilisés***

Pour la réalisation pratique du projet, j’ai utilisé les outils suivants :

* Google Colab : pour son accessibilité en ligne, sa gratuité et sa compatibilité avec Python.
* Python et bibliothèques spécialisées :
* pandas pour le traitement et la manipulation des données,
* matplotlib et seaborn pour la visualisation graphique,
* numpy pour les calculs statistiques.
* scikit-learn pour le clustering  
  Ces outils ont été sélectionnés pour leur fiabilité, leur large adoption dans le domaine de la data science.

Énoncé des problèmes rencontrés

Lors de la réalisation du projet, plusieurs problèmes ont été rencontrés :

* Problèmes de qualité des données : certaines données étaient incomplètes ou incohérentes, nécessitant un nettoyage minutieux.
* Gestion de gros volumes de données : certaines étapes de calcul étaient lentes, ce qui a nécessité l’optimisation du code Python.
* Choix des représentations graphiques : trouver la meilleure manière de visualiser les tendances pour qu’elles soient à la fois claires et pédagogiques a demandé plusieurs essais.
* Compatibilité des formats de fichiers : certains fichiers nécessitaient une conversion ou un ajustement pour être correctement importés dans Python.

Malgré ces difficultés, chaque problème a été traité avec méthode, garantissant ainsi la fiabilité et la clarté du projet final.

## Conclusion et recommandations

Apprentissage réalisé

Au cours de ce projet, j’ai acquis une compréhension approfondie des différentes étapes d’un projet d’analyse de données appliqué à l’adoption de l’intelligence artificielle. J’ai appris à collecter, nettoyer et explorer des données complexes, à utiliser des outils tels que Python et Google Colab, ainsi qu’à appliquer des techniques comme la visualisation, l’analyse comparative et le clustering. J’ai également renforcé mes compétences en analyse critique, en interprétation des résultats et en présentation graphique de tendances, ce qui m’a permis de transformer des données brutes en informations exploitables et compréhensibles.

Ce que vous feriez différemment si c’était à refaire

Si je devais refaire ce projet, je porterais une attention encore plus grande à la qualité initiale des données, en veillant à obtenir des jeux de données plus complets et standardisés dès le départ. Je consacrerais également plus de temps à l’optimisation du code pour gérer plus efficacement les volumes de données et réduire les temps de calcul. Enfin, j’expérimenterais davantage avec d’autres méthodes de clustering ou d’analyse statistique avancée afin de révéler des insights supplémentaires et enrichir l’interprétation des résultats.

*Décrire l’apprentissage le plus significatif*

L’apprentissage le plus significatif pour moi a été la capacité à transformer des données brutes et parfois imparfaites en informations claires et exploitables, notamment grâce à l’utilisation combinée du nettoyage de données, de la visualisation et du clustering. Cette expérience m’a permis de comprendre l’importance de la méthodologie et de la rigueur dans l’analyse de données, ainsi que la valeur de choisir des outils et techniques adaptés aux objectifs du projet. Ce savoir-faire est essentiel non seulement pour ce projet, mais aussi pour tout futur travail en science des données appliquée à des problématiques concrètes.

## Annexe(s)

Journal des activités (Il s’agit d’une liste des activités réalisées à chaque étape de votre projet. Chaque activité rapportée comprend une date, un nombre d’heures investi et une courte description de l’activité.)

|  |  |  |
| --- | --- | --- |
| Dates | Nombre d’heures investies | Activité réalisée |
| 18 juin 2025 | 2 | Lecture, compréhension du sujet, définition et planification initiale des objectifs |
| 20 juin 2025 | 3h | Collecte des jeux de données sur l’adoption de l’IA et vérification des sources. |
| 22 juin 2025 | 2h | Exploration préliminaire des données : identification des valeurs manquantes et anomalies. |
| 24 juin 2025 | 3h | Nettoyage et transformation des données pour préparation l’analyse. |
| 28 juin 2025 | 2h | Analyse descriptive et création de premières visualisations graphiques des tendances |
| 2 juillet 2025 | 3h | Analyse comparative entre différents indicateurs socio-économiques et adoption de l’IA. |
| 7 juillet 2025 | 2h | Application du clustering pour identifier des groupes de pays présentant des profils similaires. |
| 10 juillet 2025 | 3h | Optimisation du code Python pour accélérer le traitement des données volumineuses. |
| 15 juillet 2025 | 2h | Ajustement des visualisations graphiques pour rendre les résultats plus clairs et interprétables. |
| 20 juillet 2025 | 3h | Analyse des résultats du clustering et rédaction des interprétations. |
| 25 juillet | 2h | Compilation des résultats et rédaction des sections intermédiaires du rapport. |
| 30 juillet 2025 |  |  |
| 5 aout 2025 | 3h | Révision générale, corrections et intégration des recommandations. |
| 12 aout 2025 | 2h | Préparation des annexes et finalisation du journal des activités. |
| 19 aout 2025 | 2h | |  | | --- | |  |  |  | | --- | | Préparation des annexes et finalisation du journal des activités. | |
| 26 aout 2025 | 5h | Amélioration des visualisations et ajout de graphiques explicatifs et Rédaction de la section justification des choix méthodologiques et techniques. |
| 2 septembre 2025 | 2h | Relecture de l’ensemble du rapport et correction des incohérences. |
| 9 septembre 2025 | 2h | Finalisation des recommandations et conclusion du projet. |
| 16 sepetembre 2025 | 2h | Vérification finale des annexes, graphiques et du journal des activités |
| 22 septembre 2025 | 1h | Dépôt final du projet complet. |
|  |  |  |

Lien du projet GitHub :  
Le projet complet, incluant les notebooks Python, les scripts et le rapport final et le dashboard est disponible sur GitHub:

<https://github.com/sorel2030/projet-final-SCI1402>

### **Présentation du Tableau de bord (Dashboard)**

Dans le cadre de ce projet, j’ai conçu un tableau de bord interactif pour visualiser les indicateurs liés à l’adoption de l’intelligence artificielle (IA) à l’échelle mondiale.  
L’objectif de ce tableau de bord est de rendre accessibles et compréhensibles les données collectées et les résultats d’analyse, de manière simple et intuitive.

Contenu du tableau de bord :

* Vue d’ensemble par pays : carte ou graphique interactif présentant le score d’adoption de l’IA par pays.
* Indicateurs socio-économiques : PIB par habitant, productivite
* Comparaisons entre pays : possibilité de sélectionner plusieurs pays et de comparer leurs indicateurs sur la même période (2019 principalement).
* Résultats du modèle prédictif : estimation du score d’adoption de l’IA en fonction des variables explicatives retenues.
* Visualisations dynamiques : graphiques (histogrammes, boxplots, scatter plots) permettant d’explorer les relations entre indicateurs.

Technologies utilisées :

* Python et ses bibliothèques de visualisation (Matplotlib, Seaborn, Plotly pour l’interactivité).
* Google Colab comme environnement de développement et d’exécution.
* Dépôt GitHub pour la mise en ligne des notebooks et du tableau de bord, afin de le rendre accessible au professeur et aux autres étudiants.



