DEDICACE

À

**MA FAMILLE**

SOMMAIRE

**INTRODUCTION………………………………………………………………**

**CHAPITRE I : PRÉSENTATION DU PROJET…………………………………………………..**

**Introduction**

**I. Contexte et objectifs du projet**

- A) Contexte général

- B) Problématique abordée

- C) Objectifs visés

**II. Présentation de l’équipe et de l’organisation**

- a) Composition de l’équipe

- b) Répartition des rôles

- c) Méthodologie de travail adoptée

**III. Déroulement du projet**

1. Planning et étapes clés

2. Activités principales réalisées

3. Difficultés rencontrées

4. Choix des outils et technologies

**Conclusion**

**CHAPITRE II : ANALYSE DU BESOIN ET RECUEIL DES DONNÉES**

**Introduction**

**I. Étude du contexte d’analyse**

- Spécificités du domaine applicatif

- Justification du choix d’un projet de data science/notebook

**II. Recueil et description des données**

1. Source et description du dataset utilisé

2. Présentation des variables et dimensions (âge, sexe, IMC, charges, etc.)

3. Exploration initiale : valeurs manquantes, doublons, nettoyage des données

**III. Cahier des charges fonctionnel et technique**

1. Définition des objectifs d’analyse

2. Périmètre de l’étude

3. Contraintes identifiées

- a) Contraintes techniques

- b) Contraintes de temps

- c) Contraintes humaines

- d) Contraintes liées aux données

4. Indicateurs de succès et de suivi du projet

**Conclusion**

**CHAPITRE III :** **ANALYSE EXPLORATOIRE DES DONNÉES**

**Introduction**

**I. Analyse descriptive des données**

1. Statistiques descriptives des variables principales
2. Visualisation et distribution des variables démographiques
3. Répartition des coûts ou autres variables d’intérêt

**II. Visualisation et exploration des relations**

1. Identification des corrélations
2. Visualisation des relations majeures (ex : charges vs IMC, âge, etc.)
3. Analyse de l’impact de certains paramètres (tabac, région, etc.)

**III. Synthèse des problématiques et axes d’analyse**

1. Facteurs influents identifiés
2. Questions majeures soulevées pour la modélisation

**Conclusion**

**CHAPITRE IV : MODÉLISATION ET APPROCHE PRÉDICTIVE**

**Introduction**

**I. Préparation des données pour la modélisation**

* 1. Pré-traitement et encodage des variables
  2. Séparation des données (train/test)

**II. Construction et évaluation des modèles**

1. Choix et justification des modèles (régression linéaire, etc.)
2. Implémentation et tests
3. Évaluation des performances (R², RMSE, autres métriques)

**III. Interprétation des résultats**

1- Facteurs les plus impactants mis en évidence par le modèle

2- Discussion sur la pertinence des résultats obtenus

**Conclusion**

**CHAPITRE V : VALORISATION DES RÉSULTATS ET PERSPECTIVES**

**Introduction**

**I. Visualisation et restitution des résultats**

1. Présentation des outils de visualisation (Matplotlib, Seaborn, etc.)

2. Tableau de bord ou rapports pour les parties prenantes

**II. Synthèse des apports du projet**

1. Insights principaux

2. Recommandations pour la prise de décision

**III. Limites et perspectives d’amélioration**

1. Limites rencontrées lors du projet

2. Suggestions pour approfondir ou améliorer l’analyse/modélisation

**Conclusion**

**CHAPITRE I : PRESENTATION DU PROJET**

**Introduction**

Ce chapitre vise à présenter le contexte général du projet, les objectifs poursuivis, l’organisation de l’équipe projet et le déroulement global des travaux réalisés. Il permet de comprendre la genèse du projet, les raisons de son lancement, ainsi que les moyens mis en œuvre pour atteindre les objectifs fixés.

**I. Contexte et objectifs du projet**

A) Contexte général

Le projet tutoré s’inscrit dans le cadre d’une formation universitaire ou professionnelle, visant à mettre en pratique les compétences acquises en analyse de données et en développement d’outils de data science. Il s’agit généralement d’un projet collectif, réalisé sur une période déterminée, dont l’objectif principal est de résoudre une problématique réelle à travers l’exploitation d’un jeu de données authentique.

B) Problématique abordée

La problématique au centre de ce projet concerne l’analyse et la prédiction de certaines variables à partir d’un jeu de données, à l’aide d’outils de data science. Elle vise à démontrer comment l’exploration et la modélisation de données peuvent éclairer la prise de décision dans un contexte précis (par exemple : analyse des coûts médicaux, prédiction de charges, etc.).

C) Objectifs visés

Les objectifs principaux du projet sont les suivants :

- Explorer et comprendre un jeu de données réel.

- Mettre en œuvre différentes techniques d’analyse exploratoire des données (EDA).

- Appliquer des méthodes de modélisation statistique ou de machine learning pour réaliser des prédictions.

- Interpréter les résultats obtenus et proposer des recommandations aux parties prenantes.

- Valoriser les résultats à travers des visualisations pertinentes et un rapport synthétique

**II. Présentation de l’équipe et de l’organisation**

a) Composition de l’équipe

L’équipe projet est composée de plusieurs membres (étudiants, stagiaires, collaborateurs), chacun apportant ses compétences spécifiques en analyse de données, en programmation (notamment en Jupyter Notebook), en gestion de projet, etc.

b) Répartition des rôles

Chaque membre s’est vu attribuer des responsabilités précises, telles que :

- Collecte et préparation des données

- Réalisation de l’analyse exploratoire

- Développement et validation des modèles de prédiction

- Mise en place des visualisations et rédaction du rapport

c) Méthodologie de travail adoptée

La conduite du projet a suivi une démarche itérative, structurée en plusieurs étapes :

- Définition du périmètre et des objectifs

- Découpage des tâches et affectation des rôles

- Réunions régulières pour faire le point sur l’avancement

- Ajustements en fonction des difficultés rencontrées et des résultats obtenus

**III. Déroulement du projet**

1. Planning et étapes clés

Le projet s’est déroulé sur une période de X semaines/mois, comprenant les étapes suivantes :

- Prise en main du sujet et cadrage de la problématique

- Collecte et nettoyage des données

- Analyse exploratoire et formulation des premières hypothèses

- Modélisation et validation des résultats

- Rédaction du rapport et préparation de la soutenance

2. Activités principales réalisées

- Recherche documentaire et compréhension du contexte

- Manipulation et visualisation des données dans Jupyter Notebook

- Implémentation des modèles de prédiction

- Création de tableaux de bord ou de visualisations interactives

3. Difficultés rencontrées

- Gestion de données manquantes ou de mauvaise qualité

- Choix et paramétrage des modèles statistiques/machine learning adaptés

- Limitations techniques liées aux outils utilisés

- Organisation du travail en équipe et gestion du temps

4. Choix des outils et technologies

- Jupyter Notebook pour l’exploration, l’analyse et la modélisation

- Bibliothèques Python : pandas, numpy, matplotlib, seaborn, scikit-learn, etc.

- Outils de gestion de version (GitHub) pour le travail collaboratif

- Autres outils selon les besoins (Google Drive, Slack, Trello, etc.)

**Conclusion**

Ce premier chapitre a permis de poser les bases du projet, en présentant le contexte, la problématique, les objectifs et l’organisation mise en place. Il constitue le socle sur lequel s’appuieront les chapitres suivants, qui détailleront l’analyse des données, la modélisation, la valorisation des résultats et les perspectives d’amélioration.

**CHAPITRE II : ANALYSE DU BESOIN ET RECUEIL DES DONNÉES**

**Introduction**

Ce chapitre détaille l’analyse du besoin à l’origine du projet ainsi que le processus de recueil, d’exploration et de préparation des données utilisées. Il précise également le cadre fonctionnel et technique dans lequel s’inscrit le travail réalisé.

**I. Étude du contexte d’analyse**

Spécificités du domaine applicatif

Le projet s’inscrit dans le domaine de la data science appliquée à l’analyse de jeux de données réels, typiquement dans un contexte où la compréhension des facteurs influençant certaines variables (par exemple, les charges médicales, les dépenses, etc.) est essentielle pour la prise de décision. Ce domaine requiert la maîtrise de techniques d’exploration, de visualisation et de modélisation des données, et s’appuie sur des outils adaptés comme Jupyter Notebook et les bibliothèques Python.

Justification du choix d’un projet de data science/notebook

Le choix de réaliser ce projet sous forme de notebook Jupyter permet de combiner analyses, visualisations et commentaires explicatifs dans un même support interactif. Cela favorise la reproductibilité, la collaboration et la transparence de la démarche, tout en facilitant l’itération rapide entre exploration, modélisation et interprétation des résultats.

**II. Recueil et description des données**

1. Source et description du dataset utilisé

Le jeu de données utilisé provient d’une source fiable (open data, plateforme pédagogique, etc.) et contient un ensemble d’observations décrivant différents individus à travers des variables telles que l’âge, le sexe, l’indice de masse corporelle (IMC), le statut tabagique, la région, et la charge (montant ou coût lié à une prestation ou à un service).

2. Présentation des variables et dimensions (âge, sexe, IMC, charges, etc.)

- Âge : âge de chaque individu du dataset.

- Sexe : genre de l’individu.

- IMC : indice de masse corporelle, indicateur de corpulence.

- Statut tabagique : fumeur/non-fumeur.

- Région : zone géographique d’appartenance.

- Charges : montant à prédire ou à expliquer (ex : coût médical annuel).

Le dataset comporte généralement plusieurs centaines à milliers d’observations, chaque ligne représentant un individu, et chaque colonne une variable.

3. Exploration initiale : valeurs manquantes, doublons, nettoyage des données

Une première analyse a permis d’identifier les éventuelles valeurs manquantes ou aberrantes, ainsi que la présence de doublons. Un nettoyage rigoureux a été appliqué :

- Suppression ou imputation des valeurs manquantes selon leur criticité.

- Élimination des doublons pour garantir l’unicité des observations.

- Conversion des types de variables si nécessaire (par exemple, encodage des variables catégorielles).

**III. Cahier des charges fonctionnel et technique**

1. Définition des objectifs d’analyse

L’objectif principal est d’identifier les facteurs expliquant la variable cible (ex : charges) et de construire un modèle prédictif fiable. L’analyse doit permettre de répondre à des questions concrètes, telles que : quels sont les paramètres les plus influents ? Peut-on prédire efficacement la charge à partir des variables disponibles ?

2. Périmètre de l’étude

L’étude porte exclusivement sur le dataset fourni, sans recueil de données supplémentaires. Elle se limite à la prédiction/exploration des variables présentes, dans le cadre d’une analyse descriptive et prédictive.

3 Contraintes identifiées.

- a) Contraintes techniques : utilisation exclusive de Python et de Jupyter Notebook ; choix des bibliothèques limité à celles compatibles avec l’environnement ; capacité de traitement dépendante des ressources matérielles disponibles.

- b) Contraintes de temps : respect du calendrier du projet et des livrables intermédiaires.

- c) Contraintes humaines : répartition des tâches au sein de l’équipe, niveau de compétence en data science, disponibilité des membres.

- d) Contraintes liées aux données : qualité et complétude du jeu de données, représentativité de l’échantillon, éventuels biais ou déséquilibres.

4. Indicateurs de succès et de suivi du projet

- Qualité des analyses produites et pertinence des visualisations.

- Performance des modèles (ex : coefficient de détermination R², RMSE).

- Clarté et utilité des recommandations issues des résultat

- Respect des délais et qualité globale du rapport final

Conclusion

Ce chapitre a permis de détailler le contexte et les objectifs du projet, de présenter le dataset analysé et d’exposer les principales contraintes et attentes. Il constitue le socle méthodologique sur lequel repose l’ensemble des analyses et modélisations développées dans les chapitres suivants.

**CHAPITRE III : ANALYSE EXPLORATOIRE DES DONNÉES**

Introduction

Ce chapitre présente l’analyse exploratoire des données (EDA) réalisée sur le dataset du projet. L’objectif est de comprendre la structure, la distribution et les relations entre les différentes variables afin d’identifier les premiers axes d’analyse et les facteurs potentiellement influents pour la modélisation prédictive.

**I. Analyse descriptive des données**

1. Statistiques descriptives des variables principales

L’exploration commence par le calcul des statistiques descriptives pour les variables clés du dataset:

- Âge : moyenne, médiane, écart-type, min, max.

- IMC : moyenne, quartiles, valeurs extrêmes.

- Charges (coût à prédire) : moyenne, dispersion, distribution.

- Répartition par sexe, statut tabagique, région, etc.

Ces statistiques permettent de détecter d’éventuelles anomalies ou valeurs atypiques, et de se faire une première idée de la population étudiée.

2. Visualisation et distribution des variables démographiques

La distribution des variables démographiques (âge, sexe, IMC) est visualisée à l’aide d’histogrammes et de diagrammes en barres. Cela permet d’identifier :

- La structure par tranche d’âge.

- Les déséquilibres éventuels entre les sexes ou les régions.

- Les profils d’IMC (répartition des individus selon la corpulence).

3. Répartition des coûts ou autres variables d’intérêt

La variable “charges” (coût) est analysée en détail :

- Distribution : histogrammes, boxplots pour détecter les outliers.

- Comparaison des moyennes/variances selon les groupes (sexe, fumeur/non-fumeur, région).

- Identification d’éventuelles asymétries ou dispersions importantes.

**II. Visualisation et exploration des relations**

1. Identification des corrélations

Le calcul des coefficients de corrélation (Pearson, Spearman) permet de repérer les relations linéaires ou monotones entre les variables numériques, notamment entre l’âge, l’IMC et les charges.

2. Visualisation des relations majeures

Des nuages de points (scatter plots) et des heatmaps sont utilisés pour mettre en évidence les relations significatives, par exemple :

- Charges vs IMC

- Charges vs âge

- Charges vs sexe ou statut tabagique (avec des couleurs/groupes)

3. Analyse de l’impact de certains paramètres (tabac, région, etc.)

L’effet de variables qualitatives telles que le statut tabagique ou la région est examiné, à travers :

- Comparaison des moyennes de charges entre fumeurs et non-fumeurs.

- Variation des charges selon les régions.

- Exploration de l’influence potentielle du sexe ou d’autres facteurs.

**III. Synthèse des problématiques et axes d’analyse**

1. Facteurs influents identifiés

L’EDA met en lumière certains facteurs qui semblent fortement associés à la variable cible (charges), tels que :

- L’âge (augmentation des charges avec l’âge)

- L’IMC (corrélation positive avec les charges)

- Le statut tabagique (charges plus élevées chez les fumeurs)

- Potentiellement, la région ou le sexe

2. Questions majeures soulevées pour la modélisation

Cette phase exploratoire soulève plusieurs questions à traiter lors de la modélisation :

- Quels sont les facteurs réellement déterminants pour prédire les charges ?

- Faut-il tenir compte des interactions entre variables (ex : âge et IMC) ?

- Les modèles simples suffisent-ils ou faut-il recourir à des modèles plus complexes (non-linéaires) ?

**Conclusion**

L’analyse exploratoire a permis de mieux comprendre la structure et la distribution du dataset, d’identifier les variables et facteurs importants, ainsi que de formuler les premières hypothèses pour la suite du projet. Ces éléments serviront de base à la phase de modélisation prédictive qui sera détaillée dans le chapitre suivant.

**CHAPITRE 4 : MODÉLISATION ET APPROCHE PRÉDICTIVE**

**Introduction**

Ce chapitre présente la démarche de modélisation prédictive appliquée au jeu de données du projet. Il détaille les étapes de préparation des données, le choix et l’implémentation des modèles, ainsi que l’évaluation et l’interprétation des résultats obtenus.

I. Préparation des données pour la modélisation

1. Pré-traitement et encodage des variables

Avant d’entamer la phase de modélisation, un pré-traitement rigoureux des données a été réalisé :

- Transformation des variables catégorielles (sexe, région, statut tabagique) en variables numériques via l’encodage (Label Encoding ou One-Hot Encoding selon les cas).

- Normalisation ou standardisation des variables quantitatives si nécessaire, afin d’assurer une comparabilité entre les différentes échelles de mesure.

- Vérification de l’absence de valeurs manquantes ou aberrantes, conformément au nettoyage effectué lors de l’EDA.

2. Séparation des données (train/test)

Le dataset a été scindé en deux parties :

- Un ensemble d’entraînement (train set), représentant généralement 70 à 80 % des données, utilisé pour l’apprentissage du ou des modèles.

- Un ensemble de test (test set), représentant 20 à 30 % des données, utilisé exclusivement pour évaluer la performance des modèles sur des observations inédites.

Cette séparation garantit l’objectivité de l’évaluation et permet de limiter les risques de surapprentissage (overfitting).

II. Construction et évaluation des modèles

1. Choix et justification des modèles (régression linéaire, etc.)

Le choix du modèle prédictif s’est porté sur la régression linéaire, adaptée à la nature continue de la variable à prédire (charges/coûts). Ce modèle a été privilégié pour sa simplicité, son interprétabilité et sa pertinence dans le contexte de la prédiction de montants.

D’autres modèles plus complexes (arbres de décision, forêts aléatoires, etc.) peuvent être envisagés en complément pour améliorer la performance ou explorer des relations non-linéaires.

2. Implémentation et tests

L’implémentation s’est faite à l’aide des bibliothèques Python dédiées (scikit-learn, pandas, etc.) :

- Ajustement du modèle sur l’ensemble d’entraînement.

- Prédiction des valeurs de charges sur l’ensemble de test.

- Comparaison avec d’autres approches si nécessaire (validation croisée, modèles alternatifs).

3. Évaluation des performances (R², RMSE, autres métriques)

La performance du ou des modèles a été évaluée à l’aide de métriques appropriées :

- Coefficient de détermination \( R^2 \) : mesure la proportion de la variance expliquée par le modèle.

- Erreur quadratique moyenne (RMSE) : quantifie l’écart-type moyen des erreurs de prédiction.

- Autres métriques selon la pertinence (MAE, etc.).

Les résultats de ces métriques sont discutés pour jauger la qualité de la prédiction.

III. Interprétation des résultats

1. Facteurs les plus impactants mis en évidence par le modèle

L’analyse des coefficients du modèle de régression linéaire permet d’identifier les variables ayant le plus d’impact sur la variable cible (charges). Par exemple :

- L’âge et l’IMC sont souvent des facteurs majeurs.

- Le statut tabagique ou la région peuvent également influencer significativement les charges.

- Les coefficients permettent de quantifier l’influence de chaque variable, toutes choses égales par ailleurs.

2. Discussion sur la pertinence des résultats obtenus

Les résultats sont interprétés en regard de la problématique initiale :

- Le modèle est-il suffisamment performant pour une utilisation opérationnelle ?

- Les facteurs identifiés sont-ils cohérents avec la littérature ou les attentes métier ?

- Les éventuelles limites (biais, variables manquantes, simplification du modèle) sont discutées.

Conclusion

Ce chapitre a exposé la démarche de modélisation prédictive, depuis la préparation des données jusqu’à l’évaluation des modèles et l’interprétation des résultats. Les enseignements tirés de cette phase permettent d’orienter les recommandations et perspectives présentées dans le chapitre suivant.

**CHAPITRE VI : VALORISATION DES RÉSULTATS ET PERSPECTIVES**

**Introduction**

Ce chapitre met en avant la valorisation des résultats obtenus au cours du projet, leur restitution auprès des parties prenantes, ainsi que les perspectives d’amélioration pour de futurs travaux. Il vise à démontrer l’utilité concrète des analyses réalisées et à proposer des pistes pour aller plus loin.

**I. Visualisation et restitution des résultats**

1. Présentation des outils de visualisation (Matplotlib, Seaborn, etc.)

Afin de rendre les résultats accessibles et compréhensibles, des outils de visualisation tels que Matplotlib et Seaborn ont été utilisés. Ces bibliothèques Python permettent de générer :

- Des graphiques descriptifs (histogrammes, boxplots, nuages de points) pour illustrer la distribution des données et les relations entre variables.

- Des heatmaps pour visualiser les corrélations.

- Des courbes de performance pour l’évaluation des modèles prédictifs.

L’utilisation de ces outils a permis de rendre l’analyse plus intuitive et de mettre en évidence les principaux résultats.

2. Tableau de bord ou rapports pour les parties prenantes

Les résultats les plus pertinents ont été synthétisés sous forme de graphiques et de tableaux, intégrés dans le notebook Jupyter et/ou dans un rapport final. Ce travail de restitution vise à :

- Permettre une prise de décision rapide par les parties prenantes.

- Mettre en avant les facteurs clés influençant la variable cible.

- Proposer des visualisations interactives (si besoin) pour explorer les résultats en profondeur.

**II. Synthèse des apports du projet**

1. Insights principaux

L’analyse a permis de dégager plusieurs enseignements majeurs :

- Identification des variables les plus influentes sur la variable cible (ex : âge, IMC, statut tabagique).

- Mise en évidence de relations significatives entre certaines variables et la charge prédite.

- Détection de profils à risque ou de segments particuliers au sein des données.

2. Recommandations pour la prise de décision

En s’appuyant sur les résultats des modèles et des analyses exploratoires, le projet propose des recommandations concrètes, telles que :

- Focaliser l’attention sur les groupes ayant les charges les plus élevées.

- Mettre en place des actions ciblées pour réduire les coûts ou améliorer la prévention (dans le cas d’un contexte médical ou assurantiel).

- Utiliser le modèle prédictif comme outil d’aide à la décision pour anticiper les évolutions futures.

**III. Limites et perspectives d’amélioration**

1. Limites rencontrées lors du projet

Plusieurs limites ont été identifiées :

- Qualité et complétude des données parfois insuffisantes.

- Hypothèses simplificatrices induites par le choix de modèles linéaires.

- Potentiel biais de sélection ou de mesure dans le dataset.

- Limites liées aux outils ou au temps imparti pour l’étude.

2. Suggestions pour approfondir ou améliorer l’analyse/modélisation

Pour aller plus loin, il serait intéressant de :

- Collecter ou intégrer de nouvelles données pour enrichir l’analyse.

- Tester des modèles plus complexes (forêts aléatoires, gradient boosting, réseaux de neurones) pour améliorer la performance prédictive.

- Explorer des analyses de sensibilité ou des méthodes d’explicabilité (SHAP, LIME) pour affiner l’interprétation des modèles.

- Déployer une application ou un tableau de bord interactif pour une utilisation en production.

Conclusion

Ce dernier chapitre a permis de valoriser les résultats du projet, d’en synthétiser les apports pour les parties prenantes et d’ouvrir des perspectives d’amélioration. L’ensemble du travail réalisé démontre la pertinence de l’approche data science pour répondre à des problématiques concrètes, tout en soulignant l’importance d’une démarche itérative et évolutive.