



ML SUR DONNEES DE PANEL

Prédiction des retards de vols aux USA

Master 2 : Ingénierie et évaluation économique | Université d'Angers | 2020-2021

**Projet final JEDHA**

**Réalisé par :** BARRY Mamadou Yaya

FOTSO Galus

SOMMAIRE

[Introduction 1](#_Toc131506085)

[I. Définition du cas d’usage 2](#_Toc131506086)

[I.1. Concept fonctionnel 2](#_Toc131506087)

[I.2. Définition des utilisateurs cibles 2](#_Toc131506088)

[I.3. Ecriture de 3 users stories 2](#_Toc131506089)

[I.4. Définition des interfaces du produit (entrée/sortie) 3](#_Toc131506090)

[I.5. Définition du format des interfaces 3](#_Toc131506091)

[II. Étude Économétrique 3](#_Toc131506092)

[II.1. Présentation des données 4](#_Toc131506093)

[II.2. Présentation des variables 4](#_Toc131506094)

[II.3. Spécification du modèle 5](#_Toc131506095)

[II.4. Statistiques descriptives 6](#_Toc131506096)

[II.5. Modèle à conserver (sous TSP) 8](#_Toc131506097)

[II.6. Estimation par la méthode *within* (sous STATA) 12](#_Toc131506098)

[II.6.1. Modèle à effet fixe individuel (One-way FE) 12](#_Toc131506099)

[II.6.2. Modèle within avec effet temporel 13](#_Toc131506100)

[II.6.3. Autocorrélation intra-individus (Test de Wald modifié) 14](#_Toc131506101)

[II.6.4. Test d’hétéroscédasticité inter-individus (Test de Wald modifié) 15](#_Toc131506102)

[II.7. Interprétations économiques du modèle final 17](#_Toc131506103)

[Conclusion 19](#_Toc131506104)

[Bibliographie 20](#_Toc131506105)

# Introduction

Le transport aérien est devenu l'un des moyens de transport les plus populaires au monde, offrant rapidité, commodité et confort pour les voyageurs. Cependant, malgré les progrès technologiques et l'amélioration des infrastructures aéroportuaires, les retards de vols demeurent un problème majeur pour l'industrie.

Selon les données de la Federal Aviation Administration (FAA), environ un quart des vols commerciaux aux États-Unis ont été retardés en 2019, avec un temps moyen de retard de 54 minutes. Ces retards peuvent avoir un impact important sur les voyageurs, qui peuvent manquer des correspondances, perdre des réservations d'hôtel ou subir des coûts supplémentaires.

De plus, les retards peuvent coûter très cher aux compagnies aériennes, avec des coûts opérationnels supplémentaires tels que la rémunération des membres d'équipage, le carburant, le temps d'occupation de l'aéronef et la perte de revenus due à la réduction de la capacité de vol. Selon une étude réalisée par le Bureau of Transportation Statistics (BTS), les retards de vols ont coûté environ 1,5 milliard de dollars aux compagnies aériennes américaines en 2019.

Historiquement, les retards de vols ont été un problème récurrent dans l'industrie du transport aérien. En 2007, les compagnies aériennes américaines ont connu l'une des pires années en termes de retards, avec un taux de ponctualité moyen de 73,4%. Les causes des retards sont diverses, allant de la congestion aérienne aux intempéries, en passant par les problèmes techniques de l'avion ou les retards dus à l'embarquement et au débarquement des passagers.

C'est pourquoi, dans le cadre de ce projet, nous allons nous concentrer sur la prédiction des retards de vols aux États-Unis. Nous allons développer un modèle de Machine Learning simple à déployer, une API et une application pour aider les compagnies aériennes à mieux anticiper les retards de vols et prendre les mesures nécessaires pour minimiser leur impact. Nous allons également explorer différentes techniques de nettoyage et de prétraitement des données, ainsi que différents modèles d'apprentissage automatique pour obtenir la meilleure précision possible.

En somme, ce projet ambitieux a pour objectif d'améliorer l'expérience de voyage des passagers, d'optimiser les opérations des compagnies aériennes et de faciliter la gestion des retards de vols pour l'ensemble de l'industrie du transport aérien.

Haut du formulaire

## I. Définition du cas d’usage

* L'objectif est de prédire les retards de vols pour aider les compagnies aériennes à mieux planifier les vols et à réduire les désagréments pour les passagers.
* Les données à disposition seront les données historiques de vols, les conditions météorologiques, les données de trafic aérien et les informations sur les avions.

### I.1. Concept fonctionnel

Le concept fonctionnel de ce projet consiste à développer une solution de prédiction des retards de vols aux États-Unis, basée sur l'analyse de données historiques de vols, météorologiques, de trafic aérien et d'autres facteurs pertinents. L'objectif est de fournir aux compagnies aériennes, aux passagers et aux gestionnaires d'aéroports une solution fiable pour anticiper les retards de vols, ce qui permettrait d'améliorer la satisfaction des clients, de réduire les coûts et d'optimiser l'utilisation des ressources.

### I.2. Définition des utilisateurs cibles

Les utilisateurs cibles de ce projet sont les compagnies aériennes, les voyageurs et les professionnels du transport aérien. Les compagnies aériennes peuvent utiliser le modèle pour planifier les horaires des vols, prévoir les retards et minimiser les perturbations pour les passagers. Les voyageurs peuvent utiliser l'application pour obtenir des informations en temps réel sur les retards de vols et planifier leurs itinéraires en conséquence. Les gestionnaires du transport aérien, tels que les régulateurs et les opérateurs d'aéroports, peuvent utiliser le modèle pour évaluer les performances des compagnies aériennes et les inciter à améliorer leurs services.

I.3. Ecriture de 3 users stories

1. En tant que gestionnaires d'une compagnie aérienne, nous souhaitons avoir accès à une solution de prédiction des retards de vols, afin de pouvoir planifier les opérations de manière plus efficace, réduire les coûts liés aux retards et améliorer la satisfaction des clients.
2. En tant que voyageur : être informé en temps réel des retards potentiels de mon vol, afin de pouvoir mieux planifier mon voyage et éviter les temps d'attente inutiles à l'aéroport.
3. En tant que professionnel du trafic aérien : avoir accès à une solution de prédiction des retards de vols, afin de pouvoir optimiser l'utilisation des ressources de l'aéroport, améliorer la ponctualité des vols et améliorer la satisfaction des passagers.

### I.4. Définition des interfaces du produit (entrée/sortie)

Le produit sera constitué d'une API (Application Programming Interface) qui permettra aux utilisateurs d'interagir avec le modèle de prédiction de retard de vol. L'utilisateur devra fournir les données d'entrée via une requête HTTP à l'API dont les interfaces seront les suivantes :

Entrée :

* Données historiques de vols (compagnie aérienne, date de départ, heure de départ, aéroport de départ, aéroport d'arrivée, retard éventuel)
* Données météorologiques (température, précipitations, vent, etc.)
* Données de trafic aérien (nombre de vols dans la région, heure de pointe, etc.)
* Autres facteurs pertinents (événements spéciaux, grèves, etc.)

Sortie :

* Prédictions des retards de vols (heure estimée d'arrivée, retard potentiel, etc.) pour une compagnie aérienne donnée.

### I.5. Définition du format des interfaces

Les données historiques (entrée) de vols seront stockées dans une base de données relationnelle (par exemple, MySQL) et fournies au format JSON, qui est un format de données couramment utilisé dans les applications Web pendant que les autres données (météorologiques, de trafic aérien, etc.) également au format JSON, seront collectées via des API publiques. Les prédictions des retards de vols seront disponibles via une API REST, qui pourra être utilisée par les applications front-end (ce qui permettra une intégration facile avec d'autres applications et systèmes.).

## II. Conception de l'architecture du produit

Pour la conception de l'architecture de ce projet de prédiction de retards de vols aux USA, plusieurs choix doivent être faits afin de garantir une conception optimale et performante.

Tout d'abord, il convient de choisir un langage de programmation adapté au traitement des données et à la mise en place d'un modèle de prédiction de qualité. Python est un choix populaire pour le développement de projets de science des données et de Machine Learning, grâce à ses bibliothèques puissantes telles que Pandas, NumPy et Scikit-learn.

Ensuite, il faut choisir une plateforme de développement qui permettra de concevoir, déployer et tester facilement les différents composants du projet. Le choix de la plateforme dépendra également des compétences et des préférences de l'équipe de développement. Par exemple, l'utilisation d’un Notebook pour l'exploration des données et la mise en place du modèle de prédiction peut être une option pratique et facile à utiliser pour les scientifiques des données.

Le produit doit être découpé en différents composants logiciels, chacun ayant un rôle spécifique dans le processus de prédiction de retards de vols. Les différents composants peuvent inclure :

* Le composant de collecte de données : qui récupère les données sur les vols à partir de différentes sources telles que les sites web des compagnies aériennes, les sites de suivi de vols, etc.
* Le composant de traitement de données : qui nettoie, prépare et transforme les données pour les rendre compatibles avec le modèle de prédiction.
* Le composant de modélisation : qui inclut l'entraînement des modèles de prédiction et le choix des meilleurs modèles.
* Le composant de déploiement de l'API : qui expose les prévisions de retard de vols sous forme d'API pour être utilisées par les applications front-end ou les systèmes back-end.

Chaque composant doit être conçu de manière à pouvoir fonctionner de manière autonome, avec des interfaces clairement définies pour permettre une communication fluide entre eux. Un schéma de synthèse de l'architecture peut être créé pour visualiser et communiquer l'ensemble du système de manière claire et concise.

En résumé, la conception de l'architecture est une étape cruciale dans la mise en place d'un projet de prédiction de retards de vols aux USA. Les choix du langage de programmation, de la plateforme de développement et de la structure de composants doivent être faits avec soin pour garantir la performance, la flexibilité et la scalabilité du produit final.

# 

# Conclusion

En somme, nous retenons que l’aboutissement à un bon modèle se trouve au préalable dans la construction et la gestion des données. Cette étape nous permet d’avoir, notamment avec un nuage de point ou une matrice V de cramer, une idée sur le choix des variables visant à expliquer le phénomène d’étude.

En effet, la méthodologie de travail utilisée dans ce projet, visait au-delà d’une simple analyse économétrique. L’objectif était de réaliser des relations de causes à effets afin de confirmer ou infirmer certaines littératures qui le plus souvent des cas, sont basées sur des cas factuels et non sur la recherche de causalités via des méthodes qui pourraient exposer des phénomènes non apparents.

Les niveaux de pauvreté et de chômage galopants sont, de nos jours, des problèmes de sociétés majeurs devant impliquer toutes les différentes couches sociales.

Des politiques de relance économiques et budgétaires sont vivement recommandées, notamment en cette période de crise sanitaire afin d’éviter une implosion des classes sociales déficitaires qui se soldera par des troubles d’ordre sociaux où personne ne sera épargné.

# Bibliographie

BOSSIO ROTONDO, J. C. « Amérique latine: pauvreté, inégalités, chômage et orientations politiques. », *Vie économique* 1.3 (2010).

ÉPAULARD, Anne. « Croissance et réduction de la pauvreté dans les pays en développement et les pays en transition », *Reflets et perspectives de la vie économique*, vol. tome xlii, no. 2, 2003, pp. 9-20.

VALDES, A., and MISTIAEN J. A. « Pauvreté rurale en Amérique Latine: tendances récentes et nouveaux enjeux. », (2001).

Webographie

<https://www.banquemondiale.org/fr/region/lac/overview>, consulté le 12/01/2021

https://www.lepoint.fr/monde/l-extreme-pauvrete-touche-plus-10-de-la-population-en-amerique-latine-15-01-2019-2286104\_24.php, consulté le 12/01/2021