**IFT 6390 Fundamentals of Machine Learning**
**Guillaume Rabusseau**

# Kaggle Competition 2

**November 24, 2023**

## 1 Introduction

Dans ce projet, vous participerez à une compétition Kaggle portant sur la classification d'expressions de la langue des signes basée sur des images de la main. L'objectif de cette compétition est de classer correctement deux images de langue des signes dans leur alphabet correspondant, puis de sommer leurs valeurs ASCII respectives pour fournir le caractère final correspondant à la valeur ASCII totale.

Cette compétition a comme but de vous donner de l'expérience pratique sur un problème réel, soit celui de comprendre la langue des signes, avec une composante additionnelle de manipulation ASCII.

Le jeu de données que nous utiliserons pour cette compétition est disponible sur Sign Language MNIST. Le format de ce jeu de données est très similaire au format classique MNIST. Chaque exemple d'entraînement a une étiquette (0-25) correspondant directement à une lettre alphabétique A-Z (et aucun cas pour 9=J ou 25=Z en raison des mouvements gestuels). Les données d'entraînement (27 455 exemplaires) et les données de test (3 000 exemplaires) sont plus ou moins la moitié de la taille du MNIST standard, mais ont un format similaire à MNIST. Chaque exemple est un vecteur pixel1, pixel2....pixel784, représentant une seule image de 28x28 pixels avec des valeurs d'intensité entre 0 et 255. Les données originales d'images de gestes de la main représentent plusieurs utilisateurs répétant le geste contre différents arrière-plans. Plus de détails sur le jeu de données et sa structure sont disponibles sur la page "Data" de la compétition Kaggle.

Le but de ce projet est d'implémenter et d'entraîner plusieurs algorithmes de classification, puis de développer la logique correcte pour la manipulation ASCII requise. Vous serez évalués sur la performance de votre meilleur algorithme sur l'ensemble de test retenu ainsi que sur un rapport écrit décrivant votre méthodologie et vos résultats

La compétition, y compris les données, est disponible ici :

https://www.kaggle.com/t/26a859d3790f4c7b8ed61dd724c378ee

## 2 Dates importantes

Portez attention aux dates limites suivantes :

- 27 novembre à 23h59 : Date limite pour s'inscrire à la compétition sur Kaggle.
- 5 décembre à 23h59 : Fin de la compétition. Aucune soumission supplémentaire sur Kaggle ne sera autorisée.
- 12 décembre à 23h59 : Remise des rapports et du code sur Gradescope.

**Note sur le partage et le plagiat** : Vous êtes autorisé à discuter des techniques générales avec d'autres équipes. Vous n'êtes PAS autorisés à partager votre code. Ce comportement constitue du plagiat, et il est très facile à détecter. Toutes les équipes impliquées dans le partage du code recevront une note de 0 dans la compétition.

## 3 S'inscrire à la compétition

Les étudiants de **IFT6390** peuvent travailler individuellement ou en équipe de deux. Les étudiants de **IFT3395** peuvent travailler en équipes de 2 ou 3 personnes. Créez un compte Kaggle si vous n'êtes pas déjà inscrit, et rejoignez la compétition en suivant ce lien : https://www.kaggle.com/t/26a859d3790f4c7b8ed61dd724c378ee

Veuillez enregistrer votre nom d'utilisateur Kaggle dans ce formulaire : https://forms.gle/uWCBysE9yPGmcarG7

**Note importante**: Cette compétition est optionnelle pour les étudiants de la section **IFT3395** et sera compté comme un bonus. Pour les modalités d'évaluation voir la section 8

## 4 Baselines

Vous devrez implémenter un classifieur de **forêts aléatoires** et utiliser réseau de neurone convolutif (CNN) et battre les baselines mises de l'avant dans le classement. Ces baselines sont les suivantes :

1. Un classificateur bidon initialisé avec des paramètres aléatoires.
2. Un classificateur de forêts aléatoires.
3. La meilleure baseline des TA.

Pour participer à la compétition, vous devez fournir au moins une liste de prédictions pour les instances données sur le site web Kaggle. Vous pouvez soumettre au plus **3 prédictions par jour** tout au long de la compétition. Nous vous suggérons donc de commencer tôt afin d'avoir le temps de soumettre plusieurs fois et d'avoir une idée de vos performances. Vous obtiendrez des points supplémentaires pour chaque baseline que vous battrez parmi les 3. Votre classificateur random forest doit être implémenté à partir de zéro. En dehors des librairies Python standard, seules les librairies numpy, pandas et les matrices creuses de scipy (scipy.sparse) sont autorisées. Pour le CNN vous pouvez utiliser n'importe quelle librairie.

## 5 Autres méthodes

Vous devez essayer d'utiliser au moins **un autre modèle** en plus du random forest et du CNN. Vous êtes encouragé à mettre en œuvre les techniques étudiées pendant le cours et à rechercher d'autres façons de résoudre cette tâche. Voici quelques idées :

1. Support Vector Machines
2. Adaboost
3. XGBoost

L'objectif ici est de concevoir la méthode la plus performante en ce qui concerne l'ensemble de test Kaggle. Votre performance finale sur Kaggle

sera prise en compte comme critère d'évaluation (voir la Section 6). Si un modèle testé ne donne pas de bons résultats, vous pouvez quand même l'inclure dans votre rapport et expliquer pourquoi vous pensez qu'il n'est pas adapté à la tâche. Ce type de discussion est un critère d'évaluation important pour votre rapport final.

Pour cette section, vous êtes libre d'utiliser les librairies Python de votre choix.

## 6 Rapport

En plus de vos méthodes, vous devrez rédiger un rapport détaillant les techniques de prétraitement, de validation et d'optimisation que vous avez utilisées ainsi que les algorithmes d'apprentissage que vous avez considérés. Vous devrez également fournir des résultats permettant de comparer différents modèles.

Le rapport doit contenir les sections et éléments suivants :

- Titre du projet
- Nom complet, matricule et nom d'utilisateur Kaggle.
- Introduction : décrivez brièvement le problème et résumez votre approche et vos résultats.
- Conception des caractéristiques : décrivez et justifiez vos méthodes de prétraitement, ainsi que la manière dont vous avez conçu et sélectionné vos attributs.
- Algorithmes : donnez un aperçu des algorithmes d'apprentissage utilisés sans entrer dans trop de détails, sauf si cela est nécessaire afin de comprendre d'autres aspects.
- Méthodologie : incluez les décisions relatives à la division entraînement/validation, les stratégies de régularisation, les astuces d'optimisation, le choix des hyperparamètres, etc.
- Résultats : présentez une analyse détaillée de vos résultats. Vous pouvez inclure graphiques et tableaux au besoin. Cette analyse doit être

plus large que simplement les résultats sur le leaderboard Kaggle : incluez une brève comparaison des hyperparamètres les plus importants et de toutes les méthodes (au moins 2) que vous avez implémentées.

- Discussion : discutez des avantages/inconvénients de votre approche et méthodologie, et suggérez des idées d'amélioration.
- Déclaration des contributions : ajoutez la déclaration suivante : "Je déclare par la présente que tout le travail présenté dans ce rapport est celui de l'auteur".
- Références : très importantes si vous utilisez des idées et des méthodes trouvées dans des articles ou en ligne ; c'est une question d'intégrité académique.
- Annexe (facultatif) : vous pouvez inclure des résultats supplémentaires, plus de détails sur les méthodes, etc.

Si vous ne suivez pas ces directives, vous pourrez perdre des points. **Le corps du rapport ne doit pas dépasser 6 pages**. Les références et l'annexe ne compteront pas vers la limite de 6 page du rapport. Votre rapport et votre code doivent être soumis sur Gradescope avant le 12 décembre à 23h59.

## 7 Instructions pour la soumission

- Vous devez soumettre le code développé pendant le projet. Le code doit être bien documenté. Vous devez inclure un fichier README contenant des instructions expliquant comment exécuter votre code.
- Le fichier contenant vos prédictions sur l'ensemble de test doit être soumis en ligne sur le site Kaggle.
- Le rapport au format PDF (rédigé selon la mise en page générale décrite ci-dessus) et le code doivent être téléchargés sur Gradescope.

## 8 Critères d'évaluation

**IFT6390** Vous serez évalués selon les critères suivants:

- Des points vous seront attribués pour chacun des 3 baseline que vous battrez.
- Vous recevrez des points en fonction de votre performance finale à la fin de la compétition. Le classement visible est calculé en utilisant 23% de l'ensemble de test. Les 77% restants seront utilisés pour calculer un classement privé (qui ne sera pas visible avant la fin). Par conséquent, vous pourriez être classé premier dans le classement public et non dans le classement privé. Votre note sera calculée en utilisant les deux classements.
- Des points vous seront attribués en fonction de la qualité et de la solidité technique de votre rapport final (voir ci-dessus).

**IFT3395** Cette compétition est optionnelle est sera comptée comme bonus. Vous n'êtes pas tenus d'implémenter ni le CNN ni random forest, et vous n'avez pas à rédiger un rapport. Vous pouvez simplement participer en soumettant des prédictions, et vous aurez des points bonus en fonction de votre classement.