

Plan du 19/03/2023

* Définir un sujet
* Se répartir les tâches et leur donner une durée de réalisation
* Faire l’agenda prévisionnel
* Ajouter des solutions en cas d’imprévus (quelqu’un qui tombe malade, un ordi qui plante,
* A partir du 21/03 reste 4 semaines mais partiels aussi donc pas se surestimer. Prévoir de le finir en 2 semaines au cas où trop de problèmes surviennent.
* Pertinence de parler en jour sachant qu’on a cours à côtés ? 1j => 3h ?
* Faire une réunion au bout d’une semaine pour que tout le monde se rende compte de l’avancé ?

# Récupération (1 p, 1j)

* Trouver un jeu de données avec des variables pertinentes
* Aucun datas set trouvé 🡪 prévoir d’adapter le sujet

# Nettoyage (1p, 1j)

* Avoir un nombre suffisant d’instances mais pas trop
* Garder que les attribues qui nous intéressent
* Enlever des valeurs aberrantes qui pourrait fausser le modèle
* Jeu de données trop compliqué à traiter, peu de données pertinentes 🡪 retour étape 1

# Exploration (2p, 1j)

* Comprendre le jeu de donnée et le lien entre les différentes variables
* Choisir un modèle de machine Learning pertinent (régression linéaire, classification, jeu de neurones ; supervisé ou non supervisé)
* Séparer les données entre training set (80%) et setting set (20%), faire attention à leur pertinence

# Modélisation (2p, 4j)

* Visualiser les données
* Coder le programme de ML
* Code trop compliqué à faire 🡪 revoir à la baisse le nombre de variables, changer de modèle de ML

# Evaluation et interprétation (2p, 4j)

* Vérifier que le modèle est en accord avec nos données
* Faire un programme pour calculer le taux d’erreurs (risque empirique, fonction de perte)
* Utiliser le testing set pour vérifier que ça fonctionne
* Modèle qui ne donne pas de bons résultats 🡪 trouve les causes, si le temps on essaye de rectifier

# Mise en production (1p, 1j)

* Test du modèle avec des valeurs extérieurs

21/03/2023

Sujet : Prédire la mode des prénoms en France, prédire comment évolue un prénom rentré en paramètre sur les prochaines années

# Récupération

* Jeu de données sur Inse

# Nettoyage (1p, 2h)🡪 Victor ( data architect)

* Mettre les prénoms en pourcentage de la population qui correspond à l’année
* Retirer la différence homme/femme
* Enlever les XXXX

# Exploration

* Réseau neuronale ? suggestion de la prof
* Training set (1900 – 2015), testing set (2015-2021)

# Modélisation (2p, 4j)

* Visualiser les graphes selon nos 4 prénoms, les prénoms rares (1p, 2h) 🡪Clara
* Faire le programme : (4p, 1j) Vendredi
* (a définir vendredi)

# Evaluation et interprétation (2p, 4j)

* Compare le résultat de notre programme pour les années 2015 – 2021
* Programme d’erreur

# Mise en production (1p, 1j)

* Tester pour après 2021

Présentation : utilisation d’un nouveau vocabulaire lié à la conduite de projet + utiliser des éléments du cours