partie 1 :

* Titre : PyDiction
* le jeu de données météo sur pluriannuelle
* objectif : prédire la pluie du lendemain par ML
* deepnote (un mot)
* le plan :

charger en ligne

partie 2 (Nahim) :

*ouvrir le streamlit par Richard, partie 2*

bonjour à tous et merci Richard pour cette intro

**rappel objectif d’un modèle de prédiction** de la pluie demain **optimal entrainé par ML**

**maintenant : exploration des données brutes cad faire tout ce que nous pouvons à partir des données brutes pour obtenir un jeu de données d'entraînement optimal de notre modèle:**

* **critères des données optimisées :**

**-nb de données suffisant (critère minimal selon les algo)**

**-sans biais de récolte,**

**-sans valeurs évidemment aberrantes,**

**-sans manquantes ni doublons,**

**-variables encodées**

**-sans variables explicatives trop peu corréllées à la cible,**

**-variables au distribution normale,**

**- sans trop de valeurs**.

* Les modèles doivent être **présélectionnés** car on ne peut pas tous les évaluer
* **exploration des données BRUTES :** 1ere étape intuitive et classique qui permet d**’avancer rapidement sur certains de ces critères d’optimisation sans modifier les données**

*Richard : afficher le tableau de données*

* + permet de vérifier la **fiabilité de la rentrée des données** : c’est vérifié car : stations source fiable
  + repérer nombre suffisant pour entraîner un modèle de ML : **oui (critère déterminant)**
* **représentativité** des données : **oui** les 49 stations météo permettent cette représentativité (en théorie).
  + **sans valeurs aberrantes** : elles sont bien dans l’intervalle des données classiques en météo *montrer le tableau*
  + **Repérer les manquantes** **et doublons** concernant la question de leur élimination (obligatoire) : **il y en a beaucoup** donc on sait qu’il y aura du travail *montrer les positions case à cocher et idem cocher les % par variables.* on remarque certaines variables cependant à très fort taux de manquantes
* **Repérer les catégorielles** pour la question de l’encodage : les variables catégorielles à encoder sont : *montrer le tableau* **Localité - direction des vents forts (wind gust dir) ou non (WinDir) - pluie du jour et du lendemain. seront à encoder.**

30% de données catégorielles, donc travail à prévoir.

Les autres *montrer dans tableau* sont températures, pressions, vitessevents, humidité, enseleillements, etc sont toutes potentiellement orréllées selon les articles

*Richard : afficher les boxplots*

* **on a vérifié la normalité de la distribution des variables explicatives pour la question de la normalisation :**
  + on remarque des **échelles différentes** : 1 à 1000, ce qui est cohérent avec le type de données. → hypothèse de **normalisation à faire pour obtenir de bonnes performances**
* **cohérence de la distribution des catégorielles** : voir les vents, distribution cohérente
* **outlier** : sont normaux (exemple cet été! cela arrive!) mais seront conservés puisque le modèle doit prendre en compte le réchauffement climatique
* on a repéré ensuite les **déséquilibres de répartition** des valeurs dans les catégories de la **cible** pour la question de **l’intéret des méthodes de rééchantillonnage**, ici **oui** mais méthodes à tester (plusieurs possibles)
  + une **première sélection des modèles**: ici selon la méthode de scikit learn et les articles similaires notamment suite à la taille du jeu de données : KNN, log reg, DTC, RFC, à évaluer ensuite
* Tri des variables : nécessite un encodage pour être fait **tests stats PEARSON demande encodage nous voyons donc en temps suivant en même temps que les autres transformations pour arriver aux critères connus pour être optimaux.**

nous allons pouvoir créer aisément le pipeline à partir des hypothèses posées de travail posées lors de l’étape de l’exploration des données brutes

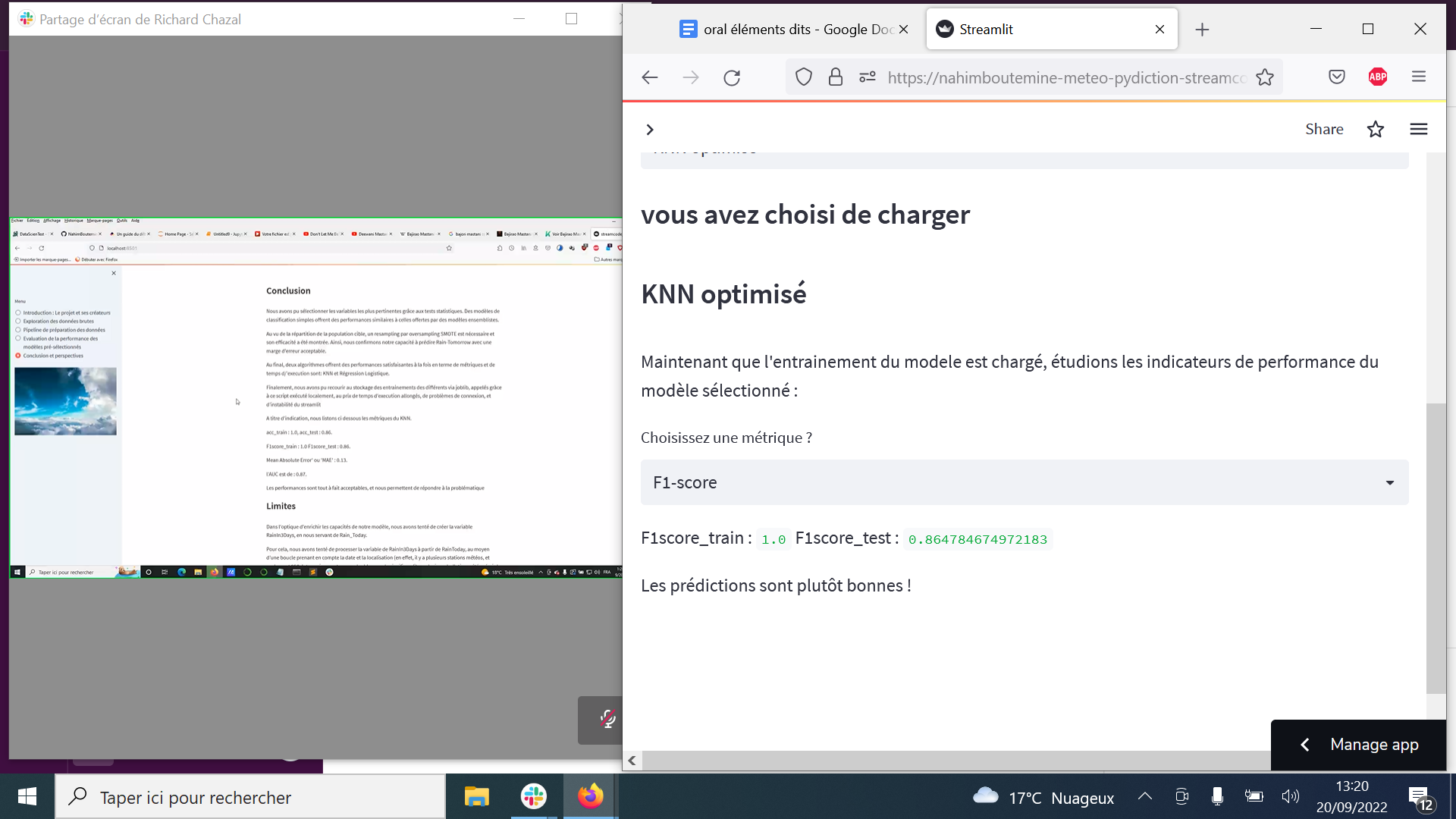
partie 3 construire le pipeline optimal : Richard

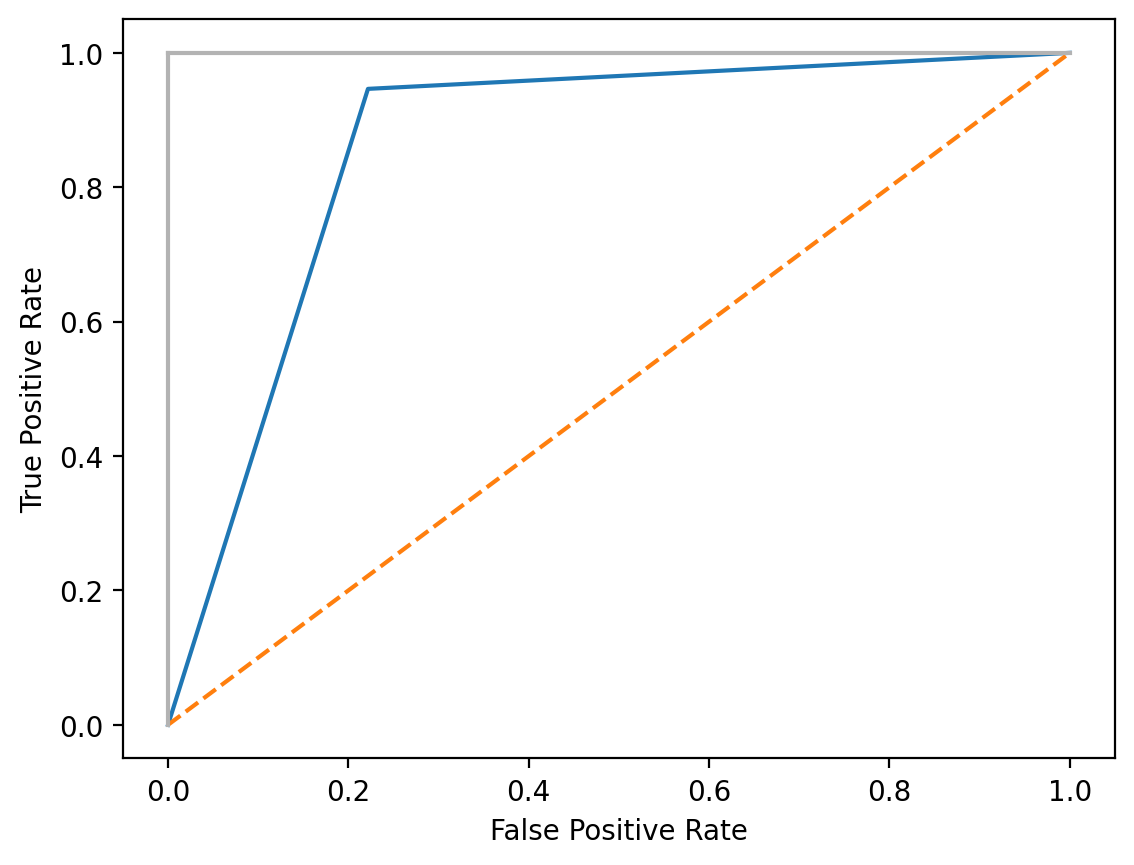
* fin de traitement des manquantes : dropna
* fin de traitement doublons donc pas de dropduplicate
* conclusions sur la section : méthode de rééchantillonnage oversampling smote produit meilleur résultat
* section normalisation : pas de normalisation pour meilleur résultat
* sélection des variables explicatives : fin par pearson, pas de variable assymétrique ou sans manquantes

partie 4 : modélisation :

* Evaluer les modèles préselectionnés : KNN, Log Reg, DTC et RFC
* Les indicateurs choisis : Accuracy, f1 score, AUC et MAE
* *ouvrir les affichages* : créé un affichage intéractif : le modèle puis les 4 indicateurs en liste déroulante
  + indicateurs sont tous satisfaisants globalement,:
    - 80 % accuracy, AUC acceptable, 20% erreur
    - raintomorrow = 1 : mal classé vu sur f1
  + le temps est le plus rapide pour KNN et log reg : 1mn30 knn et qq seconde pour log reg
  + erreur un peu moins forte pour KNN et log reg

*Richard ROC CURVE KNN*

**

**