# I. Présentation de la Mission

1. Titre du Projet: Prédiction des interventions vaines

2. Période de la Mission : 1er Janvier au 6 Septembre

# II. Contexte de la Mission

## L'entreprise

### Présentation de l’entreprise

Enedis est une entreprise de service public en charge de la gestion et de la distribution de l'électricité en France. Filiale à 100% d'EDF, Enedis est chargée de gérer le réseau public de distribution d'électricité pour environ 95% du territoire français, ce qui représente plus de 1,3 million de kilomètres de lignes électriques et près de 36 millions de compteurs électriques.

Enedis a pour mission de garantir la qualité et la continuité de la distribution d'électricité, de raccorder les clients au réseau, de réaliser les travaux d'extension et de renforcement du réseau, ainsi que de relever les compteurs et de réaliser les opérations de maintenance et de dépannage.

Enedis s'engage également dans la transition énergétique en favorisant le développement des énergies renouvelables et en accompagnant les clients dans la réduction de leur consommation d'énergie.

Enedis emploie plus de 39 000 collaborateurs et est présente sur l'ensemble du territoire français. L'entreprise est engagée dans une démarche de responsabilité sociétale et environnementale en favorisant notamment la prévention des risques professionnels et la réduction de son impact environnemental.

Le domaine GDI (Gestion des interventions) fait partie du département CESE (Conduite et exploitation du système électrique) au sein du pôle PRISME de la DSI d'ENEDIS. Son but est de fournir les applicatifs bureaux et mobilités pour les équipes de planification et de programmation, ainsi que de la réalisation des interventions terrains (Réseaux et clientèle).

CEVOL pilotage, est un projet de refonte développé par la GDI regroupant à la fois, la consultation de rapport Power BI mais également, l’accès à une base de données en libre-service, la Self-BI permettant au DR de créer leur propre rapport Power BI. Ce projet BI vise à corriger les defaults que pouvait avoir les précédente solution BI (GDI historique).

### Objectif de l’étude

L’objectif principal de cette étude est de développer un modèle prédictif capable d'anticiper les interventions vaines réalisées par les techniciens de l’entreprise ENEDIS. Une intervention vaine est définie comme une intervention où le technicien n’a pas pu réaliser son travail en raison de facteurs externes. La réduction du nombre d’interventions vaines présente plusieurs avantages opérationnels et économiques.

Les objectifs spécifiques de l’étude sont les suivants :

**1. Diminution du nombre d’interventions vaines :**

* Réduire les déplacements inutiles des techniciens, permettant ainsi d’économiser du temps et des ressources.
* Minimiser les interruptions et les frustrations pour les clients d’ENEDIS en améliorant la précision des rendez-vous.

**2. Anticipation et/ou évitement des reprogrammations :**

* Développer des prévisions précises afin d’éviter la nécessité de reprogrammer les interventions, améliorant ainsi la planification des techniciens et la satisfaction client.
* Optimiser l’agenda des techniciens en limitant les interventions annulées ou reprogrammées à la dernière minute.

**3. Réduction du nombre de kilomètres parcourus** :

* Limiter les trajets inutiles pour les interventions vaines, contribuant ainsi à une réduction des émissions de CO2 et à des économies de carburant.
* Améliorer l’efficacité logistique et réduire l’usure des véhicules de service.

**4. Amélioration du contexte de travail des techniciens et des programmateurs :**

* Fournir aux techniciens des informations précises et fiables sur les interventions planifiées, réduisant ainsi le stress et l’incertitude associés aux missions annulées.
* Faciliter le travail des programmateurs en leur offrant des outils prédictifs pour mieux gérer les plannings et anticiper les besoins de ressources.

L’étude vise donc à développer un modèle de machine learning robuste, en utilisant des données historiques et des techniques avancées de modélisation, afin de fournir des prédictions fiables et d’optimiser les opérations d’ENEDIS.

## 2. Le Service et les Interlocuteurs

### Service

La mission s'est déroulée au sein du service DSI Enedis, spécifiquement dans l'équipe FT-BI (équipe Business Intelligence de Enedis). Cette équipe a pour responsabilité de développer des rapports d’analyse avancés en utilisant PowerBI. Elle est également chargée de gérer et de maintenir l’alimentation de la base de données Self-BI, un Data Warehouse en libre-service. Ce système permet aux utilisateurs internes de créer et de personnaliser leurs propres rapports sous PowerBI, facilitant ainsi l'accès aux données et l'analyse autonome.

### Interlocuteurs

## Les principaux interlocuteurs pour cette mission ont été :

* **Kevin BURDIN et Rached THAMRI :** Tous deux sont des consultants en Business Intelligence au sein de l’équipe FT-BI. Ils ont pris en charge le projet après le départ du précédent Data Scientist. Leur rôle a été crucial pour assurer la continuité du projet, en apportant leur expertise et en s’assurant que les objectifs et les méthodes restent alignés avec les besoins de l'entreprise.

Kevin et Rached ont joué un rôle clé en fournissant les orientations nécessaires, en validant les approches méthodologiques, et en facilitant les échanges avec les autres parties prenantes.

## 3. L'environnement Matériel et Logiciel

### Matériel

Pour la réalisation de cette mission, nous avons utilisé plusieurs environnements distincts pour assurer un développement sécurisé et efficace. Ces environnements incluent :

* **Environnement de Développement** **:** Utilisé pour le développement initial et les tests des scripts et modèles.
* **Environnement de Recette :** Utilisé pour les tests plus avancés et les validations des fonctionnalités avant déploiement.
* **Environnement de Pré-Production :** Permet de simuler les conditions de production et de valider les performances et l’intégrité des solutions développées.
* **Environnement de Production :** Lieu final de déploiement des solutions validées pour une utilisation réelle et opérationnelle.

Ces environnements sont hébergés sur des machines virtuelles (VM) sous Windows Server, accessibles via une connexion Bureau à Distance, garantissant une flexibilité et une sécurité optimales pour le travail sur les différentes étapes du projet.

### Logiciel

L'ensemble du projet a été développé en utilisant un stack technologique moderne et efficace, principalement basé sur le langage Python. Voici les principaux outils et technologies utilisés :

* **Langage Python :** Langage principal utilisé pour le développement des scripts et des modèles de machine learning.
* **Anaconda :** Distribution de Python permettant de gérer les environnements virtuels et d'installer facilement les dépendances nécessaires.
* **Jupyter Notebook :** Outil utilisé pour le développement interactif et l'expérimentation des modèles, facilitant l'écriture et l'exécution de code en cellules.
* **Librairies Python :**
* pandas : Utilisée pour la manipulation et l'analyse des données.
* scikit-learn : Utilisée pour le développement des modèles de machine learning.
* XGBoost : Utilisée pour les algorithmes de gradient boosting afin d’améliorer les performances des modèles prédictifs.
* **Visual Studio Code :** Environnement de développement intégré (IDE) utilisé pour l’écriture du code avec les extensions Python et Jupyter, permettant une intégration fluide entre le développement de scripts et les notebooks interactifs.
* **GitLab :** Utilisé pour la gestion du code source et le contrôle de version. Grâce à GitLab, nous avons pu assurer un suivi rigoureux des modifications apportées au code, collaborer efficacement avec l'équipe, et gérer les branches et les merges de manière structurée.

Ces outils et environnements ont permis de mener à bien toutes les phases du projet, depuis l'exploration initiale des données jusqu'au déploiement des modèles prédictifs, tout en garantissant une collaboration efficace et une maintenance simplifiée.

# III. Déroulement Chronologique de la Mission

## 1. Planification Initiale et Réalisation

### Préparation des données (Extraction, filtrage, normalisation)

La préparation des données est une étape cruciale dans tout projet de Data Science. Pour le projet de prédiction des interventions vaines chez ENEDIS, cette étape comprend l'extraction des données, leur filtrage, et leur normalisation.

#### Extraction des données

Les données sont extraites de différentes sources internes à ENEDIS, notamment le Data Warehouse (DWH) et l'Operational Data Store (ODS). Voici les principales étapes d'extraction :

**1. Requêtes sur la DWH et l'ODS :**

* Une requête sur la DWH pour collecter la majorité des informations nécessaires.
* Trois requêtes sur l'ODS pour récupérer des informations complémentaires non disponibles dans la DWH.

**2. Fusion des données :**

* Les données extraites sont converties en DataFrames.
* Un processus de fusion (merge) est réalisé pour assembler toutes les informations en un unique DataFrame.

#### Filtrage des données

Après l'extraction, un filtrage rigoureux est appliqué pour s'assurer que seules les données pertinentes sont retenues pour l'analyse :

**1. Restriction temporelle :**

* Le jeu de données est restreint aux informations des deux premières années pour cibler l'analyse.

**2. Intégration des données externes :**

* Les données INSEE, provenant d'un fichier Excel, sont intégrées au DataFrame principal via un merge pour enrichir les informations disponibles.

**3. Critères spécifiques :**

* Les interventions sont filtrées pour exclure les cas où la présence du client n'est pas obligatoire et le compteur n'est pas accessible.
* Une nouvelle colonne `nb\_vain\_prev` est ajoutée pour indiquer le nombre d'interventions infructueuses précédentes par référence d'affaire et code\_uo.

**4. Épuration des données :**

* Certaines colonnes jugées inutiles sont supprimées pour alléger le jeu de données.
* Les sous-activités et catégories moins représentées sont regroupées sous la catégorie « Autres » pour simplifier l'analyse.

#### Normalisation des données

La normalisation est l'étape où les données sont transformées pour s'assurer qu'elles sont dans un format approprié pour la modélisation :

**1. Remplacement de valeurs :**

* Certaines valeurs du jeu de données sont remplacées en utilisant un dictionnaire spécifique pour améliorer la cohérence des données.

**2. Typage des données :**

* Les types de données sont ajustés pour s'assurer qu'ils correspondent aux besoins du modèle.

**3. Traitement des valeurs manquantes :**

* Les valeurs indiquées comme « à définir » sont remplacées par -1 pour maintenir l'intégrité numérique des données.
* Les interventions avec des valeurs manquantes significatives sont supprimées.

**4. Binarisation :**

* Les modalités des variables catégorielles sont remplacées par des valeurs numériques (0 ou 1) pour faciliter leur interprétation par les modèles de machine learning.

**5. Organisation et tri :**

* Les données sont triées par date pour faciliter les analyses temporelles.

Ces étapes de préparation des données permettent de transformer les données brutes en un jeu de données propre et structuré, prêt pour la modélisation et l'analyse prédictive.

### Modélisation (Sélection d'algorithmes, entraînement des modèles)

#### Modélisation (Sélection d'algorithmes, entraînement des modèles)

La phase de modélisation dans un projet de Data Science consiste à sélectionner les algorithmes appropriés, à entraîner les modèles sur les données préparées, et à évaluer leurs performances. Voici les étapes suivies pour la modélisation des interventions vaines chez ENEDIS :

#### Sélection d'algorithmes

Pour prédire les interventions vaines, plusieurs algorithmes de classification ont été envisagés et testés. Le choix final s'est porté sur l'algorithme XGBoost (eXtreme Gradient Boosting), en raison de sa robustesse et de ses performances élevées dans des contextes similaires. XGBoost est un algorithme de Gradient Boosting qui utilise une série d'arbres de décision pour corriger les erreurs des arbres précédents, améliorant ainsi les performances globales du modèle. Il inclut également des mécanismes de régularisation pour prévenir le surapprentissage, ce qui est crucial pour obtenir un modèle généralisable.

#### Entraînement des modèles

Une fois l'algorithme sélectionné, le modèle a été entraîné sur les données préparées. Les données ont d'abord été divisées en deux ensembles : l'ensemble d'apprentissage (80%) et l'ensemble de test (20%) pour évaluer les performances du modèle. Cette séparation permet de vérifier la capacité du modèle à généraliser sur des données non vues auparavant. Ensuite, le modèle XGBoost a été initialisé avec des paramètres par défaut adaptés au problème de classification et a été entraîné sur l'ensemble d'apprentissage, ajustant ses paramètres internes pour minimiser les erreurs de prédiction.

#### Évaluation des performances

Le modèle a ensuite été évalué sur l'ensemble de test pour vérifier sa performance en conditions réelles. Les principales métriques utilisées pour cette évaluation sont l'accuracy (précision), le recall (rappel), et la précision. En plus de cette séparation, une validation croisée a été mise en place. Cette technique consiste à diviser les données en plusieurs sous-ensembles et à entraîner le modèle plusieurs fois, en utilisant chaque sous-ensemble à tour de rôle comme ensemble de test. Cela permet d'obtenir une estimation plus robuste des performances du modèle.

#### Optimisation des hyperparamètres

Pour améliorer encore les performances du modèle, une recherche d'hyperparamètres a été effectuée. Cela inclut l'ajustement des paramètres de l'algorithme pour trouver la combinaison optimale qui maximise les performances. Une fois le modèle initial entraîné et évalué, plusieurs cycles d'amélioration ont été entrepris pour affiner les résultats. Différents hyperparamètres ont été ajustés pour optimiser les performances du modèle, et de nouveaux algorithmes ont été testés pour comparaison.

#### Segmentation des données

Pour améliorer la précision des prédictions, les données ont été segmentées en groupes homogènes partageant des caractéristiques similaires, par exemple, les interventions où la présence du client n'est pas obligatoire et le compteur est inaccessible. Cette segmentation permet de créer des modèles de prédiction plus précis car les modèles peuvent être ajustés pour refléter les spécificités de chaque segment plutôt que d'essayer d'appliquer une règle générale à un ensemble très varié.

### Evaluation des performances (Matrice de confusion, courbe ROC, etc.)

#### Évaluation des performances

L'évaluation des performances est une étape cruciale dans le processus de modélisation, car elle permet de mesurer l'efficacité du modèle à répondre au problème posé. Pour le projet de prédiction des interventions vaines chez ENEDIS, plusieurs métriques et techniques ont été utilisées pour évaluer les performances du modèle.

#### Séparation des données

Les données ont été divisées en deux ensembles : l'ensemble d'apprentissage (80%) et l'ensemble de test (20%). Cette séparation permet de vérifier la capacité du modèle à généraliser sur des données non vues auparavant. En plus de cette séparation, une validation croisée a été mise en place. La validation croisée consiste à diviser les données en plusieurs sous-ensembles et à entraîner le modèle plusieurs fois, en utilisant chaque sous-ensemble à tour de rôle comme ensemble de test. Cette technique permet d'obtenir une estimation plus robuste des performances du modèle.

#### Principales métriques utilisées

Les principales métriques utilisées pour évaluer les performances du modèle incluent l'accuracy (précision), le recall (rappel) et la précision. L'accuracy mesure le nombre de prédictions correctes faites par le modèle divisé par le nombre total de prédictions. Le recall indique la capacité du modèle à identifier correctement tous les cas positifs réels, tandis que la précision mesure la proportion de vrais positifs parmi les prédictions positives faites par le modèle.

* **Accuracy (Précision) :**

L'accuracy est la mesure la plus intuitive pour évaluer la performance d'un modèle. Elle est définie comme le nombre de prédictions correctes faites par le modèle divisé par le nombre total de prédictions. Cette métrique donne une vue d'ensemble de la performance du modèle, mais peut être trompeuse en cas de classes déséquilibrées.

* **Recall (Rappel) :**

Le recall est une mesure qui indique la capacité du modèle à identifier correctement tous les cas positifs réels. Autrement dit, parmi tous les cas réellement positifs, combien le modèle a-t-il réussi à identifier ? Le recall est particulièrement important dans les situations où il est crucial de détecter tous les cas positifs, même au risque de faire quelques erreurs positives (faux positifs).

* **Précision :**

La précision mesure la proportion de vrais positifs parmi les prédictions positives faites par le modèle. Elle est cruciale lorsque le coût des faux positifs est élevé. Une haute précision signifie que lorsqu'un modèle prédit un cas positif, il est très probablement correct.

#### Matrice de confusion

La matrice de confusion est un outil utile pour visualiser les performances du modèle. Elle compare les valeurs prédites par le modèle aux valeurs réelles, en décomposant les résultats en quatre catégories : vrais positifs, vrais négatifs, faux positifs et faux négatifs. Cela permet de comprendre non seulement la performance globale du modèle, mais aussi où il fait des erreurs.

#### Courbe ROC et AUC

La courbe ROC (Receiver Operating Characteristic) et la valeur AUC (Area Under the Curve) sont également utilisées pour évaluer les performances du modèle. La courbe ROC trace le taux de vrais positifs (recall) en fonction du taux de faux positifs, tandis que l'AUC quantifie cette courbe. Une AUC proche de 1 indique un modèle performant, tandis qu'une AUC proche de 0.5 indique une performance proche du hasard.

#### Précision moyenne et courbe de rappel-précision

La précision moyenne est calculée à partir des prédictions et permet d'évaluer la balance entre la précision et le rappel. La courbe de rappel-précision est utile pour comprendre comment ces deux métriques évoluent en fonction du seuil de classification choisi. Elle aide à trouver le bon équilibre entre détecter le plus de vrais positifs possible tout en minimisant les faux positifs.

#### Importance des caractéristiques

Enfin, l'importance des caractéristiques est évaluée pour comprendre quelles variables contribuent le plus à la prédiction du modèle. Cela permet de mettre en lumière les facteurs les plus influents et peut guider des améliorations ultérieures du modèle ou de la collecte de données.

### Améliorations et itérations du modèle

### Industrialisation et mise à disposition des résultats via PowerBI

L’industrialisation et la mise à disposition des résultats sont des étapes cruciales pour transformer les insights obtenus à partir des modèles de Data Science en actions concrètes. Pour le projet de prédiction des interventions vaines chez ENEDIS, ces étapes ont été soigneusement planifiées et exécutées pour garantir que les résultats soient accessibles et utiles aux parties prenantes.

#### Industrialisation

Le processus d’industrialisation implique deux scripts principaux, chacun ayant un rôle spécifique dans l'entraînement et l'exploitation du modèle.

**1. Script d’entraînement du modèle**

Ce script est conçu pour entraîner le modèle de prédiction en utilisant uniquement les interventions labellisées. L'entraînement du modèle est une opération qui n'est pas effectuée quotidiennement mais est relancée manuellement si le modèle montre des signes de perte de fiabilité au fil du temps. Cette approche permet de maintenir la robustesse et l'exactitude du modèle sans nécessiter un réentraînement constant.

**2. Script d’Alimentation des tables Self-BI**

Le deuxième script, écrit en Python, est exécuté quotidiennement pour alimenter plusieurs tables utilisées dans le rapport PowerBI. Ce script se lance automatiquement chaque matin à 9h via un scheduler Windows, assurant ainsi que les données sont toujours à jour. Il met à jour trois tables spécifiques :

* **Prédiction** **:** Cette table comprend les interventions labellisées et non labellisées avec les résultats de la prédiction du modèle.
* **Importance\_variable :** Cette table contient l'importance des variables au global pour toutes les interventions, fournissant une vue d'ensemble sur les facteurs influençant les prédictions.
* **Explication\_prediction :** Pour chaque intervention, cette table répertorie les cinq variables qui ont le plus impacté la prédiction du modèle vers la classe "vaine".

#### Mise à disposition des résultats via PowerBI

Pour rendre les résultats accessibles et exploitables, un rapport PowerBI a été créé et est rafraîchi quotidiennement sur la plateforme mutualisée PowerBI. Ce rapport se concentre sur le segment des interventions où la présence du client n'est pas obligatoire et le compteur est inaccessible.

Le rapport PowerBI présente divers onglets pour offrir une vue détaillée et segmentée des données :

* **Suivi de la réalisation sur le segment :** Cet onglet permet de suivre l'état d'avancement des interventions sur le segment spécifique.
* **Taux d'intervention vaines et réalisées par entité :** Ici, les taux d'interventions vaines et réalisées repérées par le modèle sont affichés, répartis par entité.
* **Poids des différentes variables et de leur modalité :** Cet onglet montre l'importance et l'impact des différentes variables utilisées par le modèle, permettant d'identifier les facteurs clés influençant les prédictions.
* **Tableau de détails des interventions :** Ce tableau répertorie les interventions avec les cinq variables ayant le plus impacté la prédiction vers la classe "vaine", offrant une transparence et une compréhension approfondie des raisons derrière chaque prédiction.

## 2. Comparaison avec les Prévisions

- Ajustements effectués par rapport aux prévisions initiales

- Problèmes rencontrés et solutions apportées (déséquilibre des classes, collecte des données, etc.)

# IV. Description du Sujet Traité

## 1. Présentation Générale du Projet

- Objectifs : Prédire les interventions vaines, comprendre les facteurs de risque, enrichir le rapport de réalisation.

## 2. Détails du Projet

- Données utilisées : Historique des interventions, variables explicatives

- Méthodologie : Compréhension du problème, préparation des données, modélisation, évaluation, et déploiement

## 3. Schémas et Diagrammes

- Inclure les schémas représentant le processus de modélisation, la segmentation des données, la matrice de confusion, etc.

# V. Résultats Obtenus et Conclusions

## 1. Résultats Obtenus

- Performances du modèle (Accuracy, Recall, Precision)

- Segmentation des données et impact sur les résultats

- Présentation des résultats par segments

## 2. Conclusions

- Efficacité du modèle dans la prédiction des interventions vaines

- Améliorations possibles et prochaines étapes (Nouvelle segmentation, intégration de nouvelles données)

- Bénéfices pour l'entreprise : Réduction des interventions vaines, amélioration de la planification

# VI. Annexes

1. Documentation Technique

- Détails des étapes de préparation des données, création et entraînement du modèle

- Codes sources principaux utilisés pour l'analyse et la modélisation

2. Schémas et Graphiques Complémentaires

- Graphiques supplémentaires illustrant les résultats des modèles, l’importance des caractéristiques, etc.