Prédiction de la maladie cardiaque dans un hôpital - Méthode CRISP-DM

Vous travaillez comme data scientist dans un hôpital, et votre tâche est de développer un modèle de machine learning pour prédire la maladie cardiaque. Ce modèle devra maximiser les bénéfices pour l'État en tenant compte des coûts liés aux erreurs de prédiction, tout en étant déployé dans un environnement accessible pour le personnel médical grâce à Streamlit. L'hôpital Saint-Pierre souhaite intégrer cette solution pour optimiser les ressources et améliorer la prise en charge des patients.

Cet exercice suit la méthodologie CRISP-DM (Cross-Industry Standard Process for Data Mining), un cadre structuré qui divise le processus d'analyse en six étapes clés. Voici comment cette approche est intégrée dans l'exercice.

Objectifs de l'exercice :

1. Compréhension des données : Nettoyer, préparer et analyser les données médicales en tenant compte du contexte métier.

2. Modélisation : Entraîner plusieurs modèles de machine learning (RandomForest, KNN, Regression Logistique).

3. Analyse des performances : Comparer les performances des modèles en termes de précision, rappel, F1-score, et bénéfices nets pour l'État.

4. Explication des prédictions : Expliquer pourquoi un patient est prédit comme "malade" ou "sain" en utilisant des techniques d'explicabilité.

5. Déploiement : Développer une interface avec Streamlit pour rendre le modèle accessible aux médecins de l'hôpital.

Méthodologie CRISP-DM appliquée à l'exercice :

Étape 1 : Business Understanding (Compréhension du problème)

L'hôpital Saint-Pierre a collecté des données sur ses patients pour améliorer le diagnostic des maladies cardiaques. Votre mission est de développer un modèle prédictif qui pourra aider à identifier rapidement les patients à haut risque de maladie cardiaque. Le modèle doit :

- Réduire les faux négatifs (patients malades non détectés) pour éviter des complications graves et coûteuses.

- Minimiser les faux positifs (patients diagnostiqués à tort) pour éviter des traitements inutiles et coûteux.

- Optimiser les bénéfices nets pour l'État.

Les coûts et bénéfices à prendre en compte sont :

1. Faux négatif (FN) : Ne pas diagnostiquer une personne malade. Coût = 50 000€

2. Faux positif (FP) : Diagnostiquer une personne comme malade alors qu'elle ne l'est pas. Coût = 10 000€

3. Vrai positif (VP) : Diagnostiquer correctement une personne malade. Bénéfice = 30 000€

4. Vrai négatif (VN) : Diagnostiquer correctement une personne non malade. Bénéfice = 0€

L'objectif final est de maximiser les bénéfices nets pour l'État en utilisant ce modèle.

Étape 2 : Data Understanding (Compréhension des données)

1. Chargement et compréhension des données :

- Le dataset contient des informations telles que l'âge, le sexe, le taux de cholestérol, la pression artérielle, etc.

- Vous devez analyser les données pour mieux comprendre la distribution des variables et les relations potentielles.

2. Visualisation des données :

- Utilisez des visualisations comme les histogrammes et les boîtes à moustaches (boxplots) pour explorer la distribution des données et identifier les anomalies.

Étape 3 : Data Preparation (Préparation des données)

1. Nettoyage des données :

- Remplacer les valeurs manquantes par la médiane des colonnes.

- Supprimer les lignes dupliquées.

2. Préparation pour le modèle :

- Séparer les caractéristiques (features) et la cible (target).

- Diviser les données en un ensemble d'entraînement (80%) et un ensemble de test (20%).

Étape 4 : Modeling (Modélisation)

1. Sélection des modèles :

- Entraîner trois modèles de machine learning : RandomForest, KNN, et Regression Logistique.

- Tester les modèles sur l'ensemble d'entraînement.

2. Évaluation des modèles :

- Utiliser les métriques suivantes : Accuracy, Recall, F1-score.

- Comparer les modèles en termes de performance.

3. Calcul des bénéfices nets :

- Pour chaque modèle, calculer les bénéfices nets avec la formule suivante :

Bénéfices nets = VP * 30 000 - FN *50 000) - FP * 10 000

- Sélectionner le modèle qui maximise les bénéfices nets.

Étape 5 : Evaluation (Évaluation du modèle)

1. Importance des variables :

- Utiliser un outil comme Yellowbrick pour visualiser l'importance des variables dans le modèle.

2. Explication des prédictions :

- Sélectionner un patient dans l'ensemble de test et expliquer pourquoi le modèle a prédit "malade" ou "non malade". Identifiez les variables les plus importantes qui ont influencé la décision.

Étape 6 : Deployment (Déploiement)

1. Création d'une interface avec Streamlit :

- Développer une application Streamlit qui permet aux médecins de saisir les informations médicales d'un patient (âge, pression artérielle, taux de cholestérol.) et d'obtenir une prédiction du modèle en temps réel.

- L'interface doit également afficher les variables importantes qui influencent le résultat de la prédiction.

2. Intérêt du déploiement pour l'hôpital :

- Le déploiement via Streamlit permet aux médecins de consulter rapidement les prédictions du modèle sans avoir besoin de connaissances techniques.

- L'interface rend le processus transparent, car elle montre les variables importantes qui influencent les décisions du modèle.

- Cette solution permet à l'hôpital de mieux gérer ses ressources en identifiant rapidement les patients à haut risque, réduisant ainsi les coûts liés aux soins inutiles ou aux complications dues à un manque de diagnostic.

Livrables attendus :

1. Rapport d'analyse :

- Résultats des différents modèles avec leurs scores et les bénéfices nets associés.

- Explication de l'importance des variables influentes.

2. Déploiement avec Streamlit :

- Une application qui permet aux médecins d'entrer des informations sur les patients et d'obtenir une prédiction avec des explications sur les variables importantes.

3. Code :

- Fournissez le code Python bien commenté qui montre l'ensemble du processus, de la préparation des données à l'explication des résultats via l'application Streamlit.

Compétences évaluées :

- Compréhension et préparation des données

- Modélisation et évaluation des modèles de machine learning

- Utilisation de la méthode CRISP-DM pour structurer un projet de data science

- Déploiement d'une application Streamlit pour un usage métier

- Analyse des résultats dans un contexte métier, en maximisant les bénéfices pour l'État

Cet exercice vous permettra de développer et déployer un modèle de machine learning orienté métier en suivant la méthode CRISP-DM. Vous devrez prendre en compte les enjeux techniques, économiques et médicaux, tout en rendant le modèle accessible et interprétable via Streamlit pour le personnel médical.

Bonne chance