

# Laboratoire : IA générative pour l'interrogation des bases de données



Effort estimé : 30 minutes

## Introduction

Les données traitées enregistrées dans une table de base de données sont accessibles, en fonction de vos besoins, à l'aide de requêtes. Les requêtes étant un élément essentiel du flux de travail de tout professionnel des données, la rédaction de requêtes *efficaces* est un ensemble de compétences nécessaires. Dans ce laboratoire, vous apprendrez comment exploiter les plateformes d'IA générative pour créer des requêtes optimisées pour vos données, à condition de pouvoir donner suffisamment de contexte au modèle.

## Objectif(s)

À la fin de ce laboratoire, vous serez en mesure d'inciter un modèle d'IA générative à créer des requêtes efficaces pour votre ensemble de données.

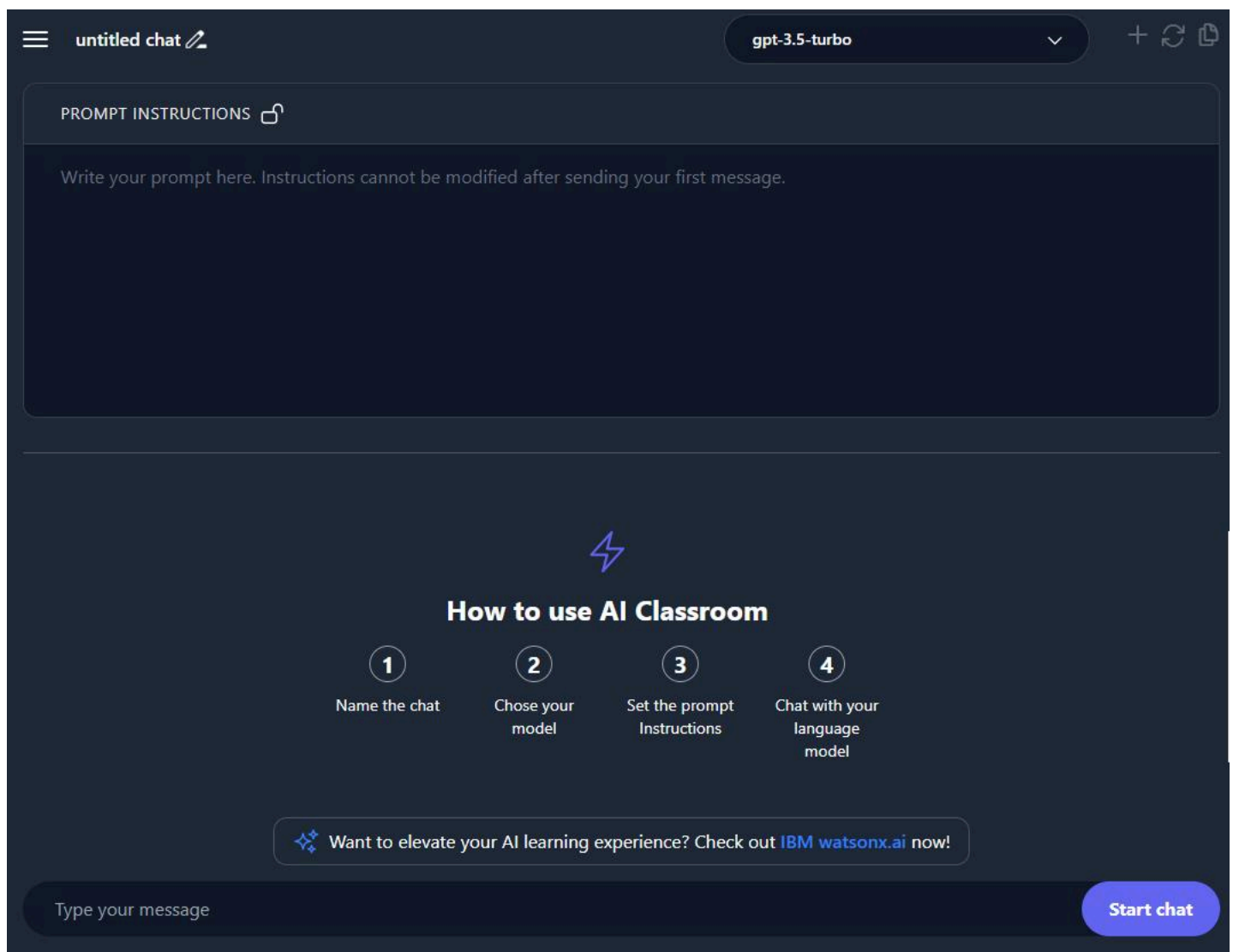
## L'ensemble de données

Dans le cadre de ce laboratoire, vous utilisez l' [ensemble de données sur les maladies cardiaques](#) de la bibliothèque UCI ML, disponible publiquement sous la licence [internationale CCA 4.0](#).

Vous pouvez télécharger l'ensemble de données et exécuter les requêtes générées dans ce laboratoire à l'aide de n'importe quel système de requête SQL.

## Donner le contexte

Vous remarquerez peut-être qu'il existe une section nommée Prompt Instructions sur l'interface Generative AI.



Si vous fournissez au modèle la description de vos données, vous pouvez générer des requêtes efficaces et facilement utilisables pour récupérer les données en fonction de vos besoins.

Collez le texte suivant dans les instructions de l'invite pour donner au modèle le contexte approprié pour les données.

We have a Heart Disease prediction dataset with a single table which has the following attributes.

1. age - age in years
2. gender- gender (1 = male; 0 = female)
3. cp - chest pain type
  - Value 1: typical angina
  - Value 2: atypical angina
  - Value 3: non-anginal pain
  - Value 4: asymptomatic
4. trestbps - resting blood pressure (in mm Hg on admission to the hospital)
5. chol - serum cholestoral in mg/dl
6. fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg - resting electrocardiographic results
  - Value 0: normal
  - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalach - maximum heart rate achieved
9. exang - exercise induced angina (1 = yes; 0 = no)
10. oldpeak - ST depression induced by exercise relative to rest
11. slope - the slope of the peak exercise ST segment
  - Value 1: upsloping
  - Value 2: flat
  - Value 3: downsloping
12. ca - number of major vessels (0-3) colored by flourosopy
13. thal - 3 = normal; 6 = fixed defect; 7 = reversable defect
14. num (the predicted attribute) - diagnosis of heart disease (angiographic disease status)
  - Value 0: < 50% diameter narrowing
  - Value 1: > 50% diameter narrowing

Le modèle disposera désormais de suffisamment de contexte pour générer des requêtes SQL pour vos invites.

Remarque : les instructions d'invite ne peuvent être modifiées que jusqu'à l'exécution de la première invite. Par la suite, vous ne pouvez plus modifier les instructions. Si une erreur se produit et que vous souhaitez recommencer, vous pouvez générer une nouvelle conversation et ajouter les instructions d'invite correctes avant de générer les requêtes SQL.

## Invites pour l'interrogation des données

### Répartition par âge

Considérez l'invite suivante.

Write an SQL query to find the minimum, maximum, and average age of patients in the dataset.

Vous pouvez vous attendre à voir la requête suivante en réponse.

```
SELECT
  MIN(age) AS min_age,
  MAX(age) AS max_age,
  AVG(age) AS avg_age
FROM
  heart_disease_prediction_dataset;
```

### Analyse de genre

Considérez l'invite suivante.

Write and SQL query to count the number of male and female patients in the dataset.

Vous pouvez vous attendre à voir la requête suivante en réponse.

```
SELECT
  gender,
  COUNT(*) AS patient_count
FROM
  heart_disease_prediction_dataset
GROUP BY
  gender;
```

### Type de douleur thoracique Fréquence

Considérez l'invite suivante.

Write an SQL query to determine the frequency of each type of chest pain (typical angina, atypical angina, non-anginal pain, asymptomatic).

Vous pouvez vous attendre à ce que la requête suivante soit générée.

```
SELECT
  cp,
  COUNT(*) AS pain_frequency
FROM
  heart_disease_prediction_dataset
GROUP BY
  cp;
```

### Analyse des groupes d'âge et des variables cibles

Considérez l'invite suivante.

Write an SQL query to investigate the distribution of the target variable (presence or absence of heart disease) within different age groups.

Vous pouvez vous attendre au résultat suivant.

```
SELECT
  CASE
```

```
        WHEN age BETWEEN 20 AND 30 THEN '20-30'
        WHEN age BETWEEN 31 AND 40 THEN '31-40'
        WHEN age BETWEEN 41 AND 50 THEN '41-50'
        WHEN age BETWEEN 51 AND 60 THEN '51-60'
        WHEN age BETWEEN 61 AND 70 THEN '61-70'
        ELSE 'Above 70'
    END AS age_group,
    SUM(CASE WHEN num = 1 THEN 1 ELSE 0 END) AS heart_disease_count,
    SUM(CASE WHEN num = 0 THEN 1 ELSE 0 END) AS no_heart_disease_count
FROM
    heart_disease_prediction_dataset
GROUP BY
    age_group
ORDER BY
    age_group;
```

## Questions pratiques

Essayez de générer des requêtes pour l'ensemble de données pour les invites suivantes :

1. Taux de cholestérol :

Find the range of cholesterol levels among patients (minimum, maximum).

2. Analyse par tranche d'âge et par sexe :

Determine the age range (youngest and oldest) for male and female patients separately.

3. Analyse des groupes d'âge et variable cible :

Investigate the distribution of the target variable (presence or absence of heart disease) within different age groups (e.g., 20-30,

4. Fréquence cardiaque maximale par groupe d'âge :

Find the maximum heart rate achieved during exercise for different age groups (e.g., 30-40, 40-50, etc.).

5. Pourcentage de patients présentant une hyperglycémie :

Calculate the percentage of patients with fasting blood sugar greater than 120 mg/dL.

6. Ratio de patients présentant une anomalie électrocardiographique au repos :

Find the ratio of patients with abnormal resting electrocardiographic results to those with normal results.

7. Nombre de patients atteints de thalassémie réversible :

Count the number of patients with reversible thalassemia detected by thallium stress testing.

8. Âge moyen des patients souffrant de douleurs thoraciques :

Calculate the average age of patients who experienced chest pain during diagnosis.

9. Répartition des patients par nombre de vaisseaux principaux :

Investigate the distribution of patients based on the number of major vessels colored by fluoroscopy (0-3).

## Conclusion

Congratulations on your successful completion of this lab!

You should now be able to use Generative AI to create efficient queries for retrieving relevant information from a database. Please note, that you need to provide the model with a detailed description of the attributes as context, to generate efficient and ready-to-use prompts.

### Author(s)

[Abhishek Gagneja](#)

© IBM Corporation. All rights reserved.