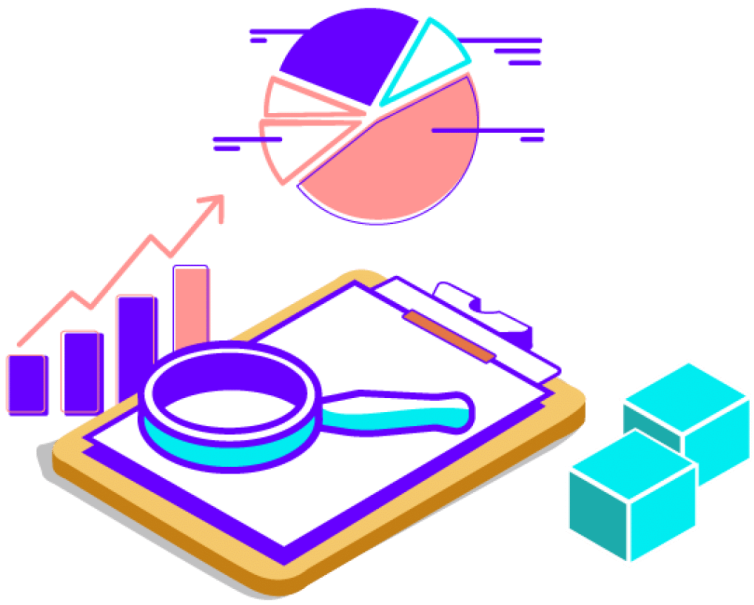
Université de Lille — FACULTÉ DES SCIENCES ET TECHNOLOGIES

Master Mathématiques et informatiques appliquées aux sciences humaines et sociales

Rapport du projet :

Développement d'un modèle prédictif d'orientation académique intégré dans une application web intuitive



Réaliser par :

Cheffi Abdelaziz

Keroudine Belladjo

Sommaire

I. Introduction

1. Contexte

Le contexte de ce projet, ancré au sein du Master MIASHS, émerge de la complexité croissante des choix éducatifs auxquels font face les bacheliers et étudiants. Face à cette réalité, notre initiative s'inscrit dans la volonté d'apporter une réponse tangible et innovante aux défis de l'orientation académique. En intégrant les enseignements des sciences humaines, des mathématiques et de l'informatique offerts par notre formation, nous avons entrepris le développement d'un modèle d'orientation. Ce modèle vise à exploiter de manière efficace les connaissances acquises au cours de notre formation pour offrir des conseils éclairés aux individus cherchant à définir leur trajectoire éducative. L'élaboration d'un questionnaire méthodologique, la création d'une base de données robuste, et l'implémentation d'un processus d'apprentissage supervisé représentent les étapes clés de notre démarche.

2. Objectif

L'objectif principal de notre projet est de concevoir un outil d'orientation précis et personnalisé, capable de guider de manière pertinente les bacheliers et étudiants dans leurs choix académiques. À travers le questionnaire méthodologique, nous visons à créer une base de données exhaustive qui alimente un processus d'apprentissage supervisé. Ce dernier doit permettre au modèle d'affiner ses prédictions en fonction des réponses des utilisateurs, offrant ainsi des conseils d'orientation adaptés et fondés sur des données solides. La création d'un site web interactif vient compléter notre objectif en offrant aux utilisateurs un accès convivial à ces prédictions, facilitant ainsi le processus de prise de décision éducative. Notre ambition est de démontrer que les compétences acquises au sein du Master MIASHS peuvent être mises en œuvre de manière concrète pour répondre aux besoins concrets de l'orientation académique.

II. Processus de collecte de données

1. Élaboration du questionnaire

L'élaboration du questionnaire a été réalisée en utilisant la platform Google Forms, cette plateforme nous a permis de créer un questionnaire anonyme, combinant habilement divers types de questions pour obtenir des réponses riches et variées. Structuré avec des questions à choix multiples, des questions libres, des questions obligatoires et optionnelles, des questions à choix unique, ainsi que des sélections, le questionnaire était conçu pour refléter la diversité des facteurs influençant les choix académiques. *(Figure 1)*

La description du questionnaire a été soigneusement rédigée pour expliquer son objectif, invitant les participants à contribuer de manière éclairée. La diffusion a été effectuée sur

différentes plateformes, notamment sur LinkedIn, Reddit r/éducation, et le hub universitaire sur Discord**.** Cette stratégie de diffusion sur des réseaux sociaux variés a favorisé une participation diversifiée, garantissant une représentation étendue des expériences éducatives. Ainsi, cette approche méthodologique réfléchie a non seulement facilité la collecte de données, mais a également contribué à l'enrichissement de notre base d'informations pour le modèle d'orientation.

2. Critères de sélection des données

La sélection judicieuse des données a constitué une étape essentielle dans notre méthodologie, intervenant après l'élaboration du questionnaire. Cette phase était cruciale pour s'assurer de la qualité, de la cohérence, et de la représentativité des informations analysées. Les critères de sélection suivants ont été appliqués avec rigueur :

- **Complétude des réponses :** Les données incomplètes ou partielles, décelées lors de l'examen des réponses au questionnaire, ont été exclues de l'analyse. Ceci visait à garantir la fiabilité des informations retenues.

- **Consistance des réponses :** Les réponses incohérentes ou contradictoires, malgré la structuration minutieuse du questionnaire, ont fait l'objet d'une vérification attentive et ont été écartées si nécessaire.

- **Anonymat des réponses :** Les réponses susceptibles de compromettre l'anonymat des participants ont été exclues de la sélection, conformément aux principes éthiques de la collecte de données.

- **Représentativité :** Les données sélectionnées ont été évaluées pour garantir une représentation équilibrée de divers profils éducatifs. Cela visait à assurer une analyse juste et exhaustive, reflétant la diversité des expériences des participants.

- **Considérations éthiques :** Les réponses présentant des aspects sensibles ou enfreignant les normes éthiques ont été traitées avec précaution, respectant ainsi les principes déontologiques.

L'application méticuleuse de ces critères a permis de constituer un ensemble de données cohérent et représentatif, jetant ainsi les bases solides nécessaires à notre analyse ultérieure et au développement du modèle d'orientation.

III. Traitement des données

1. Nettoyage et prétraitement

La phase de nettoyage et de transformation des données a été une étape cruciale dans notre démarche, intervenant après la sélection méticuleuse des informations provenant du questionnaire. Nous avons entrepris plusieurs actions visant à garantir la qualité et la cohérence des données en vue de l'analyse ultérieure *(Figure 2)*. Voici les principales étapes du nettoyage et de la transformation des données :

- **Modification des noms des colonnes :** Nous avons uniformisé les noms des colonnes pour assurer une cohérence dans la structuration des données, facilitant ainsi la compréhension et l'analyse.

- **Suppression des colonnes supplémentaires :** Les colonnes redondantes telles que 'Horodateur', 'Score', et 'Adresse e-mail' ont été éliminées pour alléger le jeu de données tout en préservant la pertinence des informations restantes.

- **Suppression des espaces :** Nous avons éliminé les espaces superflus en début et fin des éléments textuels des colonnes, garantissant une uniformité dans la présentation des données.

- **Remplacement des valeurs manquantes :** Les valeurs manquantes ont été remplacées par la mention 'Non Renseigné' pour maintenir l'intégrité des informations tout en évitant les distorsions potentielles dans l'analyse.

- **Mise en majuscules :** Nous avons converti en majuscules tous les éléments textuels des colonnes pour standardiser la présentation des données, facilitant ainsi la recherche et l'analyse.

Ces actions, menées avec précision et discernement, ont contribué à la création d'un jeu de données propre, homogène et prêt à être exploité dans le cadre de notre projet d'orientation académique. Ce processus de nettoyage et de transformation a jeté les bases solides nécessaires pour l'analyse statistique et le développement du modèle prédictif.

2. Structuration des réponses

Structuration des Réponses impliqué une série d'actions visant à corriger, organiser et structurer les informations pour une analyse plus approfondie.

Regroupement des valeurs redondantes et des valeurs avec moins de réponses : Nous avons regroupé les valeurs redondantes et celles avec un nombre limité de réponses pour simplifier le jeu de données. Cela a permis de concentrer l'analyse sur des catégories significatives et représentatives.

Traitement des réponses à choix multiple : Pour les réponses à choix multiple, nous avons mis en place un processus de séparation des résultats par virgule en différentes colonnes. Cette approche a facilité la manipulation des données et a permis une analyse détaillée des préférences multiples des participants. ***(Figure 3)***

Traitement des réponses optionnelles : La méthode adoptée pour le traitement des réponses optionnelles visant à identifier des motifs et à regrouper les réponses similaires. Une fois cette phase accomplie, une catégorisation a été effectuée pour organiser ces réponses en groupes spécifiques en fonction de leurs caractéristiques communes. Cette approche analytique a permis de structurer efficacement les données

IV. Modèle prédictif d'orientation académique

La sélection des variables pour notre modèle d'orientation s'est concentrée sur des facteurs pertinents tels que le baccalauréat, la mention, la filière actuelle, le type d'étudiant, et d'autres. Notre objectif principal était de prédire la filière choisie par les participants. Pour valider nos choix, nous avons utilisé des tests statistiques, notamment le test de Cramér, le test du Chi-deux, et le test de Tchuprow, afin d'évaluer les relations significatives au sein de nos données, contribuant à la construction d'un modèle prédictif robuste.

La filière a été choisie comme variable cible pour mes algorithmes supervisés en raison de son rôle central dans la prise de décision académique. Cette approche vise à modéliser et prédire les choix de filière en fonction de différentes caractéristiques, offrant ainsi des insights clés pour les orientations éducatives.

Pour améliorer la manipulation des données catégorielles, nous avons utilisé l'encodage one-hot. Cette approche a été appliquée à deux catégories de variables : les colonnes ordinales, telles que la mention, le type d'étudiant, la tranche d'âge et le niveau d'études, et les colonnes nominales comprenant des attributs tels que le baccalauréat, la filière actuelle, le type de lieu d'études, le genre, la satisfaction envers la formation, les spécialités, et les matières préférées.

Les colonnes ordinales ont été encodées de manière à préserver l'ordre hiérarchique des catégories, tandis que les colonnes nominales ont été transformées en variables binaires distinctes à l'aide de l'encodage one-hot.

1. Algorithmes d'apprentissage employés

Dans le cadre de notre étude sur l'orientation académique, notre approche s'est concentrée sur l'implémentation pratique de divers algorithmes d'apprentissage pour extraire des informations significatives de nos données. Trois méthodes distinctes ont été déployées, chacune étant adaptée à des aspects spécifiques de notre analyse.

Random Forest :

L'implémentation du modèle Random Forest a débuté par la création et l'entraînement du modèle. Les prédictions ont ensuite été générées sur l'ensemble de test, suivi de l'évaluation de la performance du modèle. Dans une perspective d'optimisation, nous avons défini les paramètres à ajuster, initialisé le classifieur Random Forest, et utilisé la validation croisée pour déterminer les meilleurs hyperparamètres. Enfin, le modèle a été entraîné avec ces hyperparamètres optimaux, consolidant ainsi sa capacité prédictive.

Régression Logistique :

La mise en œuvre de la régression logistique a suivi une séquence standard. Après la division des données en ensembles d'entraînement et de test, le modèle a été initialisé et entraîné sur l'ensemble d'entraînement. Les prédictions ont été générées sur l'ensemble de test, permettant une évaluation rigoureuse de la performance du modèle.

K plus proches voisins :

Pour l'algorithme KNN, nous avons optimisé le nombre de voisins (k) via la validation croisée, configuré le modèle, et intégré les étiquettes des voisins les plus proches dans le dataframe initial. Cette démarche a permis d'offrir des perspectives détaillées sur les relations entre les observations, renforçant ainsi notre analyse.

2. Évaluation des performances des modèles

Suite à l'application des algorithmes d'apprentissage, nous avons évalué leurs performances, les résultats des tests de précision, présentés ci-dessous avec des valeurs, offrent une perspective détaillée sur les capacités prédictives de chaque modèle.

	Précision (%)	Erreur Empirique	Précision en Généralisation (%)	Erreur en Généralisation
Random Forest	65.85	2.52	74.39	2.68
Régression Logistique	83.17	1.87	81.95	1.35
KNN	58.78	3.11	62.73	3.47

La Régression Logistique émerge comme le modèle le plus performant avec une précision de 83.17% sur l'ensemble de test et une excellente généralisation avec une précision de 81.95%. Le Random Forest affiche également une performance solide, tandis que le KNN montre une précision plus modérée.

V. Intégration du modèle dans une application web

Dans le cadre de notre projet, nous avons mis en œuvre une application web interactive intégrant notre modèle d'orientation. Notre objectif est de rendre les résultats de notre travail accessibles pour le grand public.

1. Conception de l'Interface utilisateur

Le concept central de cette mise en œuvre est de permettre aux utilisateurs de recevoir une recommandation personnalisée pour leur choix académique. Les variables sélectionnées pour le site web ont été délibérément choisies pour capturer des aspects clés des utilisateurs, assurant ainsi la pertinence des conseils fournis par notre modèle prédictif.

Le site est structuré autour de deux pages distinctes, chacune jouant un rôle essentiel dans la facilitation des choix académiques des utilisateurs.

Page du formulaire :

La première page du site est dédiée à la collecte d'informations pertinentes à travers un formulaire interactif. L'utilisateur est guidé à travers une série de questions soigneusement élaborées, présentées sous forme de menus déroulants pour simplifier le processus de réponse. Ces questions sont spécifiquement conçues pour capturer les variables essentielles, telles que le type de baccalauréat, la mention obtenue, la filière actuelle, et d'autres aspects cruciaux. *(Figure 4)*

Page des résultats :

Une fois le formulaire complété, les utilisateurs accèdent à la deuxième page, où La filière recommandée par notre modèle prédictif est affichée en évidence, offrant ainsi une réponse directe à la requête de l'utilisateur et deux graphiques informatifs, le premier graphique met en évidence les choix les plus fréquemment de filières d'orientation, le second graphique illustre les préférences générales des utilisateurs en matière de domaines d'étude. ***(Figure 5) et (Figure 6)***

2. Implémentation du modèle dans l'architecture backend

L'intégration réussie de notre modèle d'orientation avec le Framework Flask et la page HTML a impliqué plusieurs étapes clés, chacune jouant un rôle essentiel dans la création d'une interface utilisateur fonctionnelle et réactive. Voici une explication détaillée des différentes étapes de ce processus :

Exportation du modèle : Tout d'abord, nous avons exporté notre modèle d'orientation, entraîné avec soin, pour le préparer à l'intégration dans l'environnement Flask. Cette exportation incluait non seulement le modèle lui-même mais également les encodeurs, décodeurs.

Création de l'application Flask : La première étape côté backend a été la mise en place d'une application Flask. Cela a impliqué la définition des routes nécessaires pour gérer les différentes étapes de l'interaction utilisateur. Ces routes ont été conçues pour recevoir les réponses du formulaire et renvoyer les résultats de l'orientation.

Formulaire HTML dynamique : Sur le frontend, la page du formulaire a été créée en HTML avec des inputs select. Les options de ces inputs ont été générées dynamiquement à partir des données préétablies, assurant ainsi une synchronisation avec le modèle. Le formulaire a été conçu pour collecter des informations spécifiques, alignées avec les variables clés nécessaires au modèle.

Traitement des réponses : Lorsqu'un utilisateur soumet le formulaire, les réponses sont transmises au backend Flask. Ces réponses sont ensuite prétraitées conformément aux encodeurs et décodeurs exportés, et utilisées comme données d'entrée pour le modèle prédictif.

Génération de la Recommandation : Le modèle prédictif, une fois alimenté avec les réponses de l'utilisateur, génère une recommandation d'orientation académique. Cette recommandation est renvoyée au frontend via Flask pour être présentée à l'utilisateur.

Cette démarche d'intégration complète entre Flask, HTML, et le modèle d'orientation a abouti à la création d'une interface utilisateur interactive et fonctionnelle, offrant une expérience transparente pour les utilisateurs cherchant des conseils d'orientation académique.

VI. Conclusion

Ce projet de modélisation et d'orientation académique a été une exploration approfondie des synergies entre la collecte de données, la modélisation prédictive et l'intégration technologique, visant à offrir une solution interactive et personnalisée pour guider les choix éducatifs des utilisateurs.

La phase de collecte de données a révélé la complexité des facteurs influençant les décisions académiques, mettant en lumière la nécessité d'une approche holistique. À travers l'implémentation de techniques d'apprentissage supervisé, nous avons développé un modèle capable de fournir des recommandations précises en prenant en compte cette variété de paramètres.

L'intégration de ce modèle dans une interface web avec Flask et HTML a marqué une étape cruciale, transformant notre concept en une ressource tangible et accessible. Cependant, ce processus n'a pas été sans ses défis, nécessitant des solutions ingénieuses pour assurer une expérience utilisateur fluide et réactive.

Au-delà des compétences techniques acquises, ce projet a renforcé notre compréhension des nuances inhérentes aux décisions académiques, mettant en lumière l'importance de l'approche collaborative au sein de notre équipe.

En anticipant l'avenir, il est évident que ce projet ne constitue qu'une première étape. L'amélioration continue de la précision du modèle, l'optimisation de l'interface utilisateur pour une expérience utilisateur optimale et les perspectives de recherche future dans le domaine de l'intelligence artificielle et de l'orientation académique ouvrent la voie à des opportunités passionnantes.

En résumé, ce projet académique témoigne de notre engagement envers l'innovation et l'application pratique des connaissances acquises. Nous sommes animés par la conviction que cette trajectoire d'apprentissage nous positionne pour contribuer de manière significative aux solutions intelligentes dans le domaine complexe de l'orientation académique.

VII. Annexe

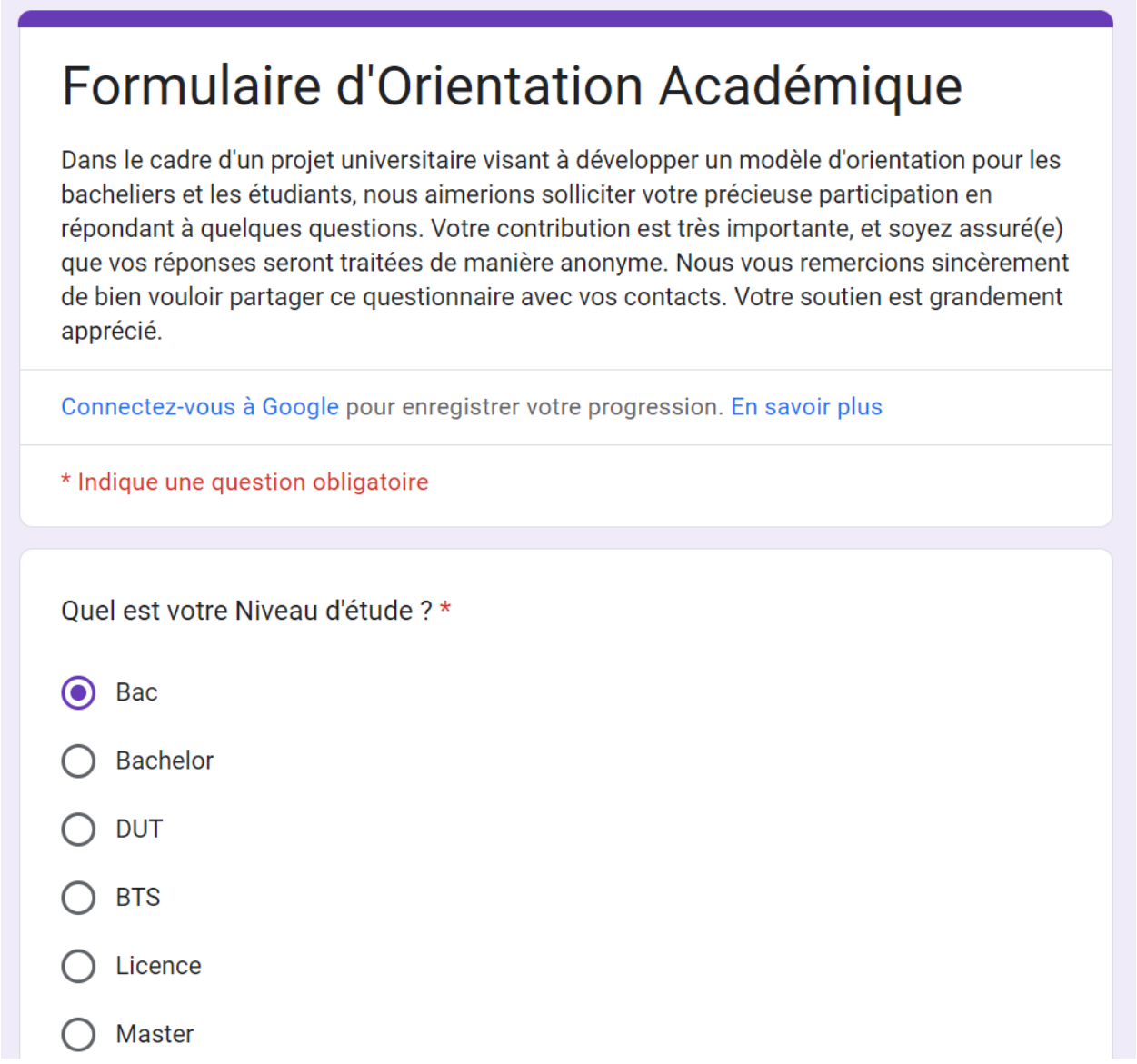


FIGURE 1: QUESTIONNAIRE D'ORIENTATION ACADEMIQUE

```python
#les colonnes supplementaires
data = data.drop(['Horodateur','Score', 'Adresse e-mail'], axis=1)


# Suppression des espaces au début et à la fin à tous les éléments textuels des colonnes
data = data.applymap(lambda x: x.strip() if isinstance(x, str) else x)


#remplacer les valeurs manquantes par non renseigné
data.fillna("non renseigné", inplace=True)


# Mettre en majuscule tous les éléments textuels des colonnes
data = data.applymap(lambda x: x.upper() if isinstance(x, str) else x)
```

Figure 2: NETTOYAGE et prétraitement des données

```python
#Fonction pour divisisé une colonne presentant des element enumerés  en plusieurs colonnes
def division_colonne(df,nom_col):
    data = df.copy()
    # Division de la colonne 'nom_col' en listes en utilisant la virgule comme séparateur
    data[nom_col] = data[nom_col].str.split(',')

    # Trouver le nombre maximum d'éléments dans une cellule
    max_elements = data[nom_col].apply(len).max()

    # Création de nouvelles colonnes à partir des listes
    df_new = pd.concat([data[nom_col].apply(lambda x: x[i] if len(x) > i else None).rename(f'Col_{nom_col+str(i+1)}')
                        for i in range(max_elements)], axis=1)

    # Concaténation des nouvelles colonnes avec le DataFrame d'origine
    df = pd.concat([data, df_new], axis=1)

    return df
```

Figure 3: STRUCTURATION des réponses à choix multiple

FIGURE 4: INTERFACE UTILISATEUR (PAGE DU FORMULAIRE)

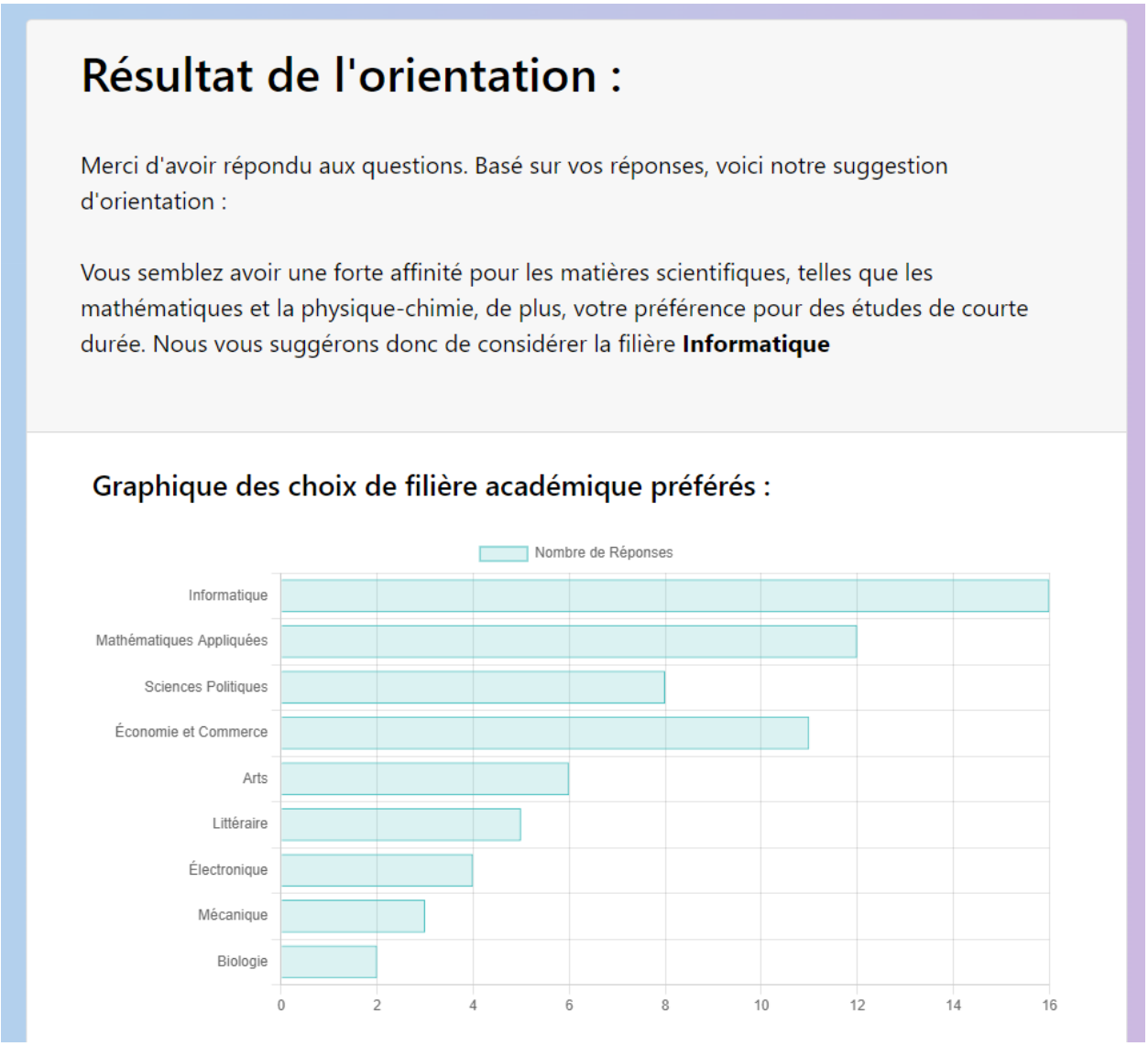


FIGURE 5: INTERFACE UTILISATEUR (PAGE DES RESULTATS 1)

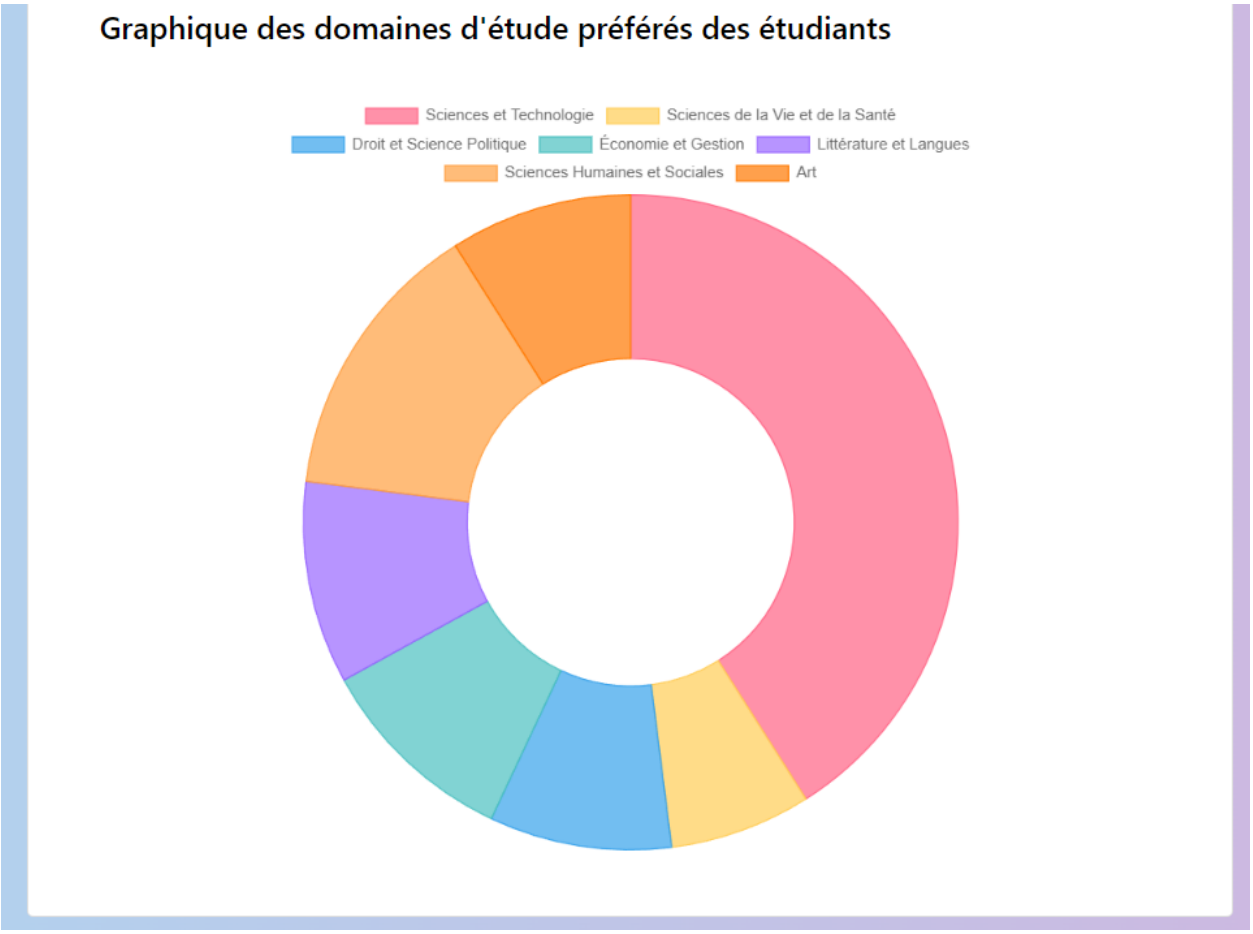


FIGURE 6: INTERFACE UTILISATEUR (PAGE DES RESULTATS 2)