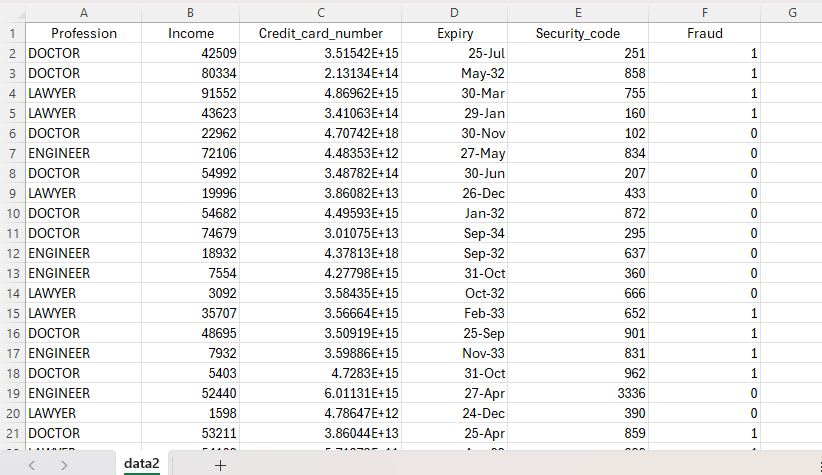
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | 18 Janvier, 2025 | |
| ***Data Mining*** | | | | |
|  | | | | |
| Mlle. Chaima Rouita  *Mlle. Latifa Didi Alaoui*  *Mlle. Nouhayla Elassouli*  *Mlle. Nihal Tizgha* | |  |  |  |

***Introduction du projet***:

Ce projet a pour objectif de détecter les transactions frauduleuses à l'aide de techniques avancées de data mining et d'apprentissage automatique. Le travail s’est articulé autour de l’exploration des données, du prétraitement, de l’extraction de caractéristiques textuelles avec TF-IDF, et de la classification à l’aide de modèles tels que SVM et KNN.

***Description du Dataset***:



1. **Structure des Données**

Le dataset contient les colonnes suivantes :

* **Profession** : Colonne catégorielle indiquant la profession de l'individu (DOCTOR, ENGINEER, LAWYER).
* **Income** : Revenu annuel de l'individu (valeurs numériques).
* **Credit\_card\_number** : Numéro de carte de crédit (valeur sensible non utilisée dans l’analyse).
* **Expiry** : Date d'expiration de la carte de crédit.
* **Security\_code** : Code de sécurité de la carte (numérique).
* **Fraud** : Colonne cible indiquant si une transaction est frauduleuse (1) ou légitime (0).

1. **Statistiques Descriptives**

* Taille : 10 000 lignes et 6 colonnes.
* Données présentes : Aucune valeur manquante.
* Types : Combinaison de colonnes catégoriques (Profession), numériques (Income, Security\_code) et sensibles.

***Etapes du Projet***:

1. **Prétraitement des Données**

**1.1 Nettoyage des Données**

* **Colonnes supprimées** : Les colonnes Credit\_card\_number et Expiry ont été supprimées car elles ne contribuent pas directement à l’analyse ou à la modélisation.
  + Exemple avant suppression :

|  |  |  |
| --- | --- | --- |
| **Credit\_card\_number** | **Expiry** | **Income** |
| 3515418493460774 | 07/25 | 42509 |
| 213134223583196 | 05/32 | 80334 |

* + Exemple après suppression:

|  |
| --- |
| Income |
| 42509 |
| 80334 |

**1.2 Encodage des Données Catégoriques**

* La colonne Profession a été encodée en valeurs numériques à l'aide de **LabelEncoder** :
  + DOCTOR → 0
  + ENGINEER → 1
  + LAWYER → 2

**1.3 Normalisation des Données Numériques**

* Les colonnes Income et Security\_code ont été normalisées à l'aide de **MinMaxScaler** pour les ramener dans une plage de valeurs entre 0 et 1.
* La normalisation a été nécessaire pour éviter que les échelles différentes des colonnes numériques n’influencent de manière disproportionnée les modèles d’apprentissage automatique, garantissant ainsi une meilleure convergence et une performance optimale.

**2. Extraction de Caractéristiques Textuelles (TF-IDF)**

**2.1 Ajout d’une Colonne Textuelle**

Pour appliquer l'extraction de caractéristiques TF-IDF, une colonne textuelle intitulée Description a été ajoutée au dataset. Voici quelques exemples de valeurs :

* "Doctor with high income"
* "Engineer specialized in software"
* "Lawyer focused on corporate law"

Cette colonne a été générée pour correspondre à la longueur totale du dataset.

A screenshot of a computer

Description automatically generated

**2.2 Application de TF-IDF**

* Extraction des 100 principales caractéristiques textuelles avec **TfidfVectorizer**.
* Les valeurs TF-IDF ont été converties en colonnes numériques et ajoutées au dataset.
* La colonne Description a été supprimée après l’extraction TF-IDF.

A screenshot of a computer

Description automatically generated

**3. Sélection des Caractéristiques**

* Utilisation de **SelectKBest** avec le test du **Chi2** pour sélectionner les caractéristiques les plus pertinentes.
* Les caractéristiques finales incluent :
  + Profession, Income, Security\_code (originales).
  + Plusieurs caractéristiques textuelles extraites par TF-IDF.