Compte rendu – Mini projet – Machine Learning

Les données sont extraites des bases de caractéristiques et de lieux des accidents de la route en France pour 5 ans.

Les variables sélectionnées pour l'analyse incluent des caractéristiques telles que la luminosité (lum), l'agglomération (agg), les conditions atmosphériques (atm), la catégorie de route (catr), et la surface de la route (surf).

Installation librairies manquantes :

```
[ ]  !pip install ggplot
     !pip install pandas
     !pip install factoextra
     !pip install factoMineR
     !pip install caret
     !pip install scikit-learn
     !pip install skimpy
     !pip install tidyr
     !pip install scikit-learn
     !pip install xlrd
```

Recodage des Variables :

Certaines variables sont recodées pour simplifier l'analyse, par exemple, regroupement des modalités de certaines variables.

Les données sont ensuite nettoyées pour permettre l'application de l'ACP sur des variables qualitatives.

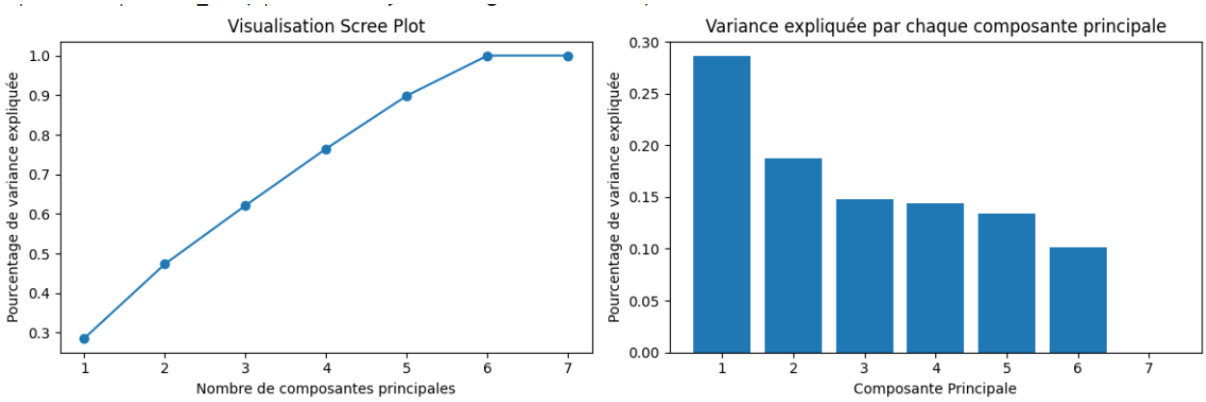
```
# Charger les données
caracteristics = pd.read_csv('caracteristics.csv', encoding='ISO-8859-1')
places = pd.read_csv('places.csv', encoding='ISO-8859-1')

# Sélectionner les colonnes pertinentes
ca_v1 = caracteristics[["Num_Acc", "an", "lum", "dep", "atm"]]
li_v1 = places[["Num_Acc", "catr", "surf"]]

ca_v1 = ca_v1[ca_v1["an"] <= 5]
```

Analyse en Composantes Principales (ACP) :

L'ACP est utilisée pour visualiser la structure des données dans un espace de dimension réduite. La contribution de chaque variable à la formation des axes est examinée, montrant quelles caractéristiques sont les plus importantes pour chaque axe.

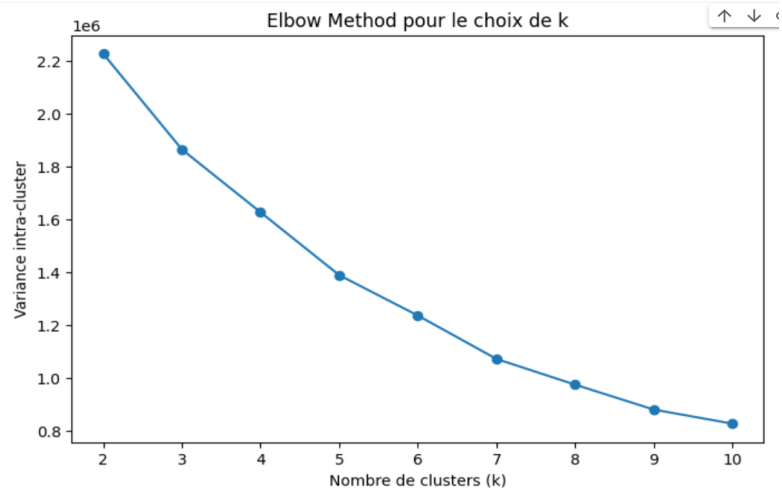
Choix du Nombre de Clusters (K) :

Une analyse est effectuée pour déterminer le nombre optimal de clusters en utilisant la méthode du coude, qui suggère que le choix de K devrait être basé sur une baisse significative de la variance intra-cluster.



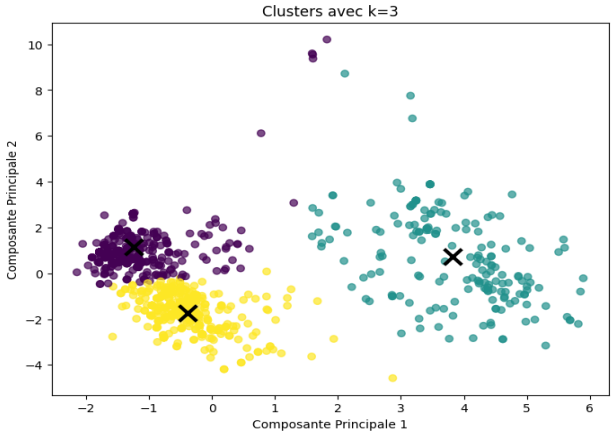
Clustering avec Kmeans :

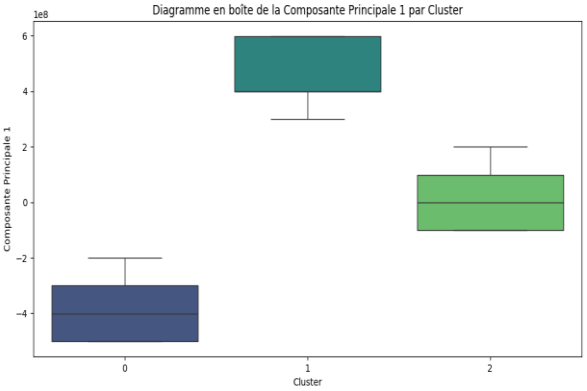
Les données sont divisées en clusters en utilisant l'algorithme Kmeans avec le nombre optimal de clusters obtenu précédemment.



Visualisation des Clusters :

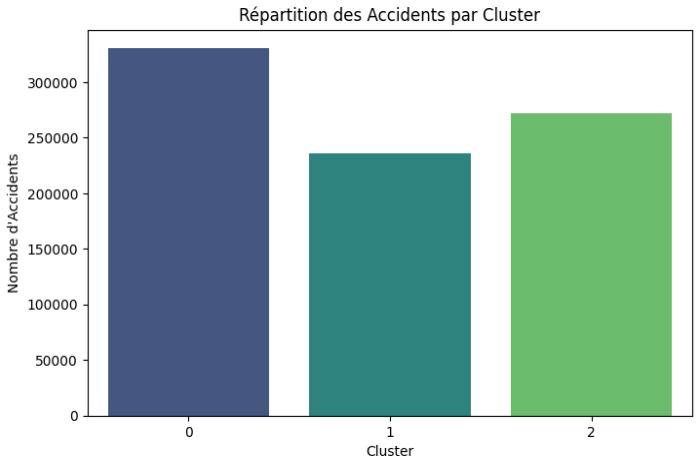
Les individus sont représentés dans un espace bidimensionnel en fonction des deux premiers axes factoriels de l'ACP.





Les clusters sont visuellement identifiés et les résultats sont interprétés.

Nombre d'Accidents par Cluster :



Les résultats finaux montrent que les accidents sont répartis en trois clusters, avec les effectifs suivants :

Cluster 0 (330,635 accidents), Cluster 1 (235,730 accidents), Cluster 2 (272,417 accidents).

Cluster 0 (330,635 accidents) :

Ce cluster représente la majorité des accidents, indiquant potentiellement une catégorie de situations de conduite courantes.

Des facteurs tels que l'agglomération, la catégorie de route et les conditions atmosphériques sont cruciaux ici. Par exemple, une concentration plus élevée d'accidents dans des zones urbaines (agglomération) pourrait suggérer des défis liés à la circulation, aux intersections, etc.

Il serait intéressant d'explorer les caractéristiques spécifiques de ce cluster pour identifier les points chauds et proposer des mesures préventives.

Cluster 1 (235,730 accidents) :

Ce cluster, bien que représentant un nombre significatif d'accidents, en compte moins que le Cluster 0. Cela pourrait indiquer des différences dans les circonstances entourant ces incidents.

Les caractéristiques spécifiques de ce cluster doivent être étudiées. Par exemple, une prévalence plus élevée d'accidents sur des routes spécifiques ou dans des conditions météorologiques particulières.

L'analyse approfondie des variables contributrices pourrait révéler des tendances intéressantes, comme une forte corrélation avec des conditions atmosphériques spécifiques.

Cluster 2 (272,417 accidents) :

Ce cluster représente également une part importante du nombre total d'accidents. Les caractéristiques qui distinguent ce cluster des autres méritent une attention particulière.

Les variables comme la catégorie de route, la surface de la route, etc., peuvent fournir des informations cruciales sur les conditions où ces accidents se produisent. Par exemple, une fréquence plus élevée d'accidents sur des routes spécifiques ou dans des zones géographiques spécifiques.

L'identification des caractéristiques distinctives de ce cluster peut aider à orienter les efforts de prévention.

Interprétation des Clusters :

L'interprétation des clusters est basée sur la visualisation des contributions des variables à la formation des axes et des cercles de corrélation.

Les différences majeures entre les clusters sont liées à des caractéristiques telles que l'agglomération, la catégorie de route, les conditions atmosphériques et la surface de la route.

Il est important de noter les similitudes et les différences entre les clusters pour formuler des recommandations et des stratégies de sécurité routière adaptées.

Les autorités de sécurité routière pourraient utiliser ces informations pour cibler des campagnes de sensibilisation spécifiques à chaque cluster, mettant l'accent sur les facteurs de risque particuliers.

Des analyses démographiques supplémentaires (par exemple, l'âge du conducteur, le type de véhicule) pourraient enrichir davantage la compréhension des modèles d'accidents.

En résumé, l'approche de clustering permet de dégager des tendances distinctes entre les groupes d'accidents, offrant ainsi une base solide pour des initiatives ciblées en matière de sécurité routière. Les résultats peuvent orienter les politiques et les programmes de prévention des accidents, contribuant ainsi à une réduction globale des incidents sur les routes.

Lumière (lum1, lum2, lum3):

Cluster 0: Présence significative de lumière (0.693) par rapport aux autres clusters.

Cluster 1: Lumière modérée (0.225).

Cluster 2: Lumière faible (0.0817).

Agglomération (agg1, agg2):

Cluster 0: Présence significative dans les zones agglomérées (agg1 = 0.303).

Cluster 1: Présence modérée dans les zones agglomérées (agg2 = 0.697).

Cluster 2: Présence relativement équilibrée entre aggloméré et non aggloméré.

Conditions atmosphériques (atm1, atm2, atm3):

Cluster 0: Présence significative de conditions atmosphériques (atm1 = 0.954).

Cluster 1: Présence modérée de conditions atmosphériques (atm2 = 0.0263).

Cluster 2: Présence faible de conditions atmosphériques (atm3 = 0.0193).

Catégorie de route (catr1, catr2, catr3):

Cluster 0: Présence significative sur catégorie de route (catr1 = 0.459).

Cluster 1: Présence modérée sur catégorie de route (catr2 = 0.520).

Cluster 2: Présence faible sur catégorie de route (catr3 = 0.0212).

Surface de la route (surf1, surf2, surf3, surf4):

Cluster 0: Présence significative sur la surface de la route (surf1 = 0.810).

Cluster 1: Présence modérée sur la surface de la route (surf2 = 0.182).

Cluster 2: Présence faible sur la surface de la route (surf3 = 0.00194, surf4 = 0.00615).

Conclusion :

Le clustering permet de regrouper les accidents en fonction de caractéristiques similaires, fournissant ainsi des informations utiles pour comprendre les tendances et les patterns sous-jacents.

Le choix de trois clusters semble approprié, offrant une segmentation significative tout en maintenant une interprétabilité.