**Objectifs**

* Visualisation des données;
* Prétraitement des données.
* Analyse de données.
* Réalisation de tableaux de bord.
* Création de modèles.
* Prédiction de l'économie africaine.
* Analyse et prise de décision pour la croissance de l'économie africaine dans les différents secteurs .
* Tout savoir sur Rapidminer

**Plan**

**Introduction**

**Analyse Exploratoire**

* Identification des types de données dans chacune des colonnes;
* Vérification de l'exactitude des données(la source, les valeurs);
* Identification de la nature des colonnes(informations apportés, unité de mesure)

**Prétraitement des données**

* Detection de valeurs aberrantes(nulles, anormales, doublons) dans le jeu de données;
* Traitement (remplacement / suppression des valeurs manquantes et gestion des valeurs anormales).

**Analyse des données**

* Identification d'une série de questions à des fin d'analyse et de prises de décisions pertinentes;
* Identification des colonnes à utiliser;
* Identifications d'outils mathématiques et statistiques permettant la réalisation de l'analyse des données;

**Visualisation des données**

* Identification de l'outils de visualisation;
* Choix des types de graphes;

**Conclusion**

LOTHAIRE

**I°) Analyse exploratoire des données**

**I.1) Définition**

L'analyse exploratoire des données est un moyen puissant d'explorer un jeu de données. Même lorsque l'objectif est d'effectuer des analyses planifiées, l'analyse exploratoire des données peut être utilisée pour le nettoyage de données, l'analyse des sous-groupes ou simplement pour mieux comprendre vos données.

L'analyse exploratoire des données s'apparente à un travail d'investigation : vous recherchez des indices et des informations qui peuvent conduire à l'identification des causes potentielles du problème que vous essayez de résoudre.

**I.2) Pourquoi l'analyse exploratoire des données est-elle importante dans la science des données ?**

L'objectif principal de l'AED est d'aider à examiner les données avant de formuler des hypothèses. Elle peut permettre d'identifier les erreurs évidentes, mais aussi de mieux comprendre les modèles (patterns) au sein des données, de détecter les valeurs aberrantes ou les événements anormaux, de trouver des relations intéressantes entre les variables.

L’analyse exploratoire des données consiste en :

-Identification des types de données dans chacune des colonnes ;

-Vérification de l'exactitude des données (la source, les valeurs) ;

- Identification de la nature des colonnes (informations apportés, unité de mesure).

**I.3.) Vérification de l'exactitude des données(la source, les valeurs);**

La source des données est [Kaggle](https://www.kaggle.com/datasets/mahmoudsaeed99/african-economy-from-1980-to-2022). Kaggle est une plateforme web qui accueille la plus grande communauté de Data Science au monde, avec plus de 536 000 membres actifs dans 194 pays et reçoit près de 150 000 soumissions par mois, et qui lui fournit des outils et des ressources puissants pour aider à atteindre tous les progrès de science des données. Kaggle, de la même manière que Datascientest, offre un environnement Jupyter Notebooks personnalisable et sans configuration. Sont accessibles gratuitement des GPU et une grande quantité de données et de codes publiés par la communauté. À l’intérieur de Kaggle, vous trouverez tout le code et les données dont vous avez besoin pour réaliser vos projets de science des données. Il y a plus de 50 000 jeux de données publics et 400 000 notebooks publics disponibles pour tous.

Le dataset contient 31 colonnes relatives à des indicateurs de croissance économique ; elle possède aussi 2322 lignes.

**II°) Prétraitement des données**

Le prétraitement des données est une technique d’exploration de données utilisée pour transformer les données brutes dans un format utile et efficace.

**II.1) Étapes impliquées dans le prétraitement des données**

**II.1.1) Nettoyage des données :**

Les données peuvent comporter de nombreuses parties non pertinentes et manquantes. Pour gérer cette partie, un nettoyage des données est effectué. Cela implique la gestion des données manquantes, des données bruitées, etc.

**a°) Données manquantes :**

Cette situation survient lorsque certaines données manquent dans le dataset. Il peut être manipulé de différentes manières.

Certains d’entre eux sont :

§ Ignorer les tuples :

Cette approche ne convient que lorsque l’ensemble de données dont nous disposons est assez volumineux et qu’il manque plusieurs valeurs dans un tuple.

§ Remplir les valeurs manquantes :

Il existe différentes manières d’effectuer cette tâche. Vous pouvez choisir de remplir les valeurs manquantes manuellement, par la moyenne d’attribut ou la valeur la plus probable.

**a°) Données bruyantes :**

Les données bruyantes sont des données sans signification qui ne peuvent pas être interprétées par des machines. Elles peuvent être générées en raison d’une mauvaise collecte de données, d’erreurs de saisie de données, etc.

**II.1.2) Transformation des données :**

Cette étape est effectuée afin de transformer les données sous des formes appropriées adaptées au processus d’exploration de données. Cela implique les moyens suivants :

a°) Normalisation :

Elle est effectuée afin de mettre à l’échelle les valeurs de données dans une plage spécifiée (-1,0 à 1,0 ou 0,0 à 1,0).

b°) **Sélection d’attributs :**

C’est une partie importante de l’apprentissage automatique. Elle désigne le processus visant à réduire les entrées à traiter et à analyser, ou à identifier les entrées les plus pertinentes.

La sélection d’attributs est essentielle à la création d’un modèle performant pour plusieurs raisons. L’une est que la sélection de caractéristiques implique un certain degré de réduction de la cardinalité, pour imposer un seuil sur le nombre d’attributs qui peuvent être pris en compte lors de la création d’un modèle. Les données contiennent presque toujours plus d’informations que nécessaire pour générer le modèle ou elles contiennent un type d’informations inapproprié. Par exemple, vous pouvez avoir un dataset de 500 colonnes qui décrit les caractéristiques des clients. Toutefois, si certaines colonnes contiennent des données éparses, cela n’est pas très utile de les ajouter au modèle et, si certaines colonnes sont en double, leur utilisation peut rendre le modèle inexact.

La sélection d’attributs améliore la qualité du modèle, tout en optimisant le processus de modélisation. Si vous incluez des colonnes inutiles pour générer un modèle et de mémoire sont consommées pendant le processus d’apprentissage, et il faut plus d’espace de stockage pour le modèle terminé. Indépendamment du problème des ressources, vous avez intérêt à effectuer la sélection des fonctionnalités et à identifier les colonnes les plus pertinentes, car l’utilisation de colonnes inutiles peut diminuer la qualité du modèle de plusieurs façons :

- Des données parasites ou redondantes rendent plus difficile la découverte de séquences significatives.

**II.1.3) Réduction des données :**

Étant donné que l’exploration de données est une technique utilisée pour gérer une énorme quantité de données. Tout en travaillant avec un énorme volume de données, l’analyse est devenue plus difficile dans de tels cas. Afin de s’en débarrasser, nous utilisons une technique de réduction des données. Il vise à augmenter l’efficacité du stockage et à réduire les coûts de stockage et d’analyse des données.

**III°) Analyse de données**

L'analyse des données est le processus qui consiste à examiner et à interpréter des données afin d'élaborer des réponses à des questions. Les principales étapes du processus d'analyse consistent à cerner les sujets d'analyse, à déterminer la disponibilité de données appropriées, à décider des méthodes qu'il y a lieu d'utiliser pour répondre aux questions d'intérêt, à appliquer les méthodes et à évaluer, résumer et communiquer les résultats.

L'analyse des données est essentielle pour comprendre les résultats des enquêtes, des sources administratives et des études pilotes, pour obtenir des renseignements sur les lacunes en matière de données, pour concevoir et remanier les enquêtes, pour planifier de nouvelles activités statistiques et pour formuler des objectifs en matière de qualité.

L’analyse des données est un sous-domaine des statistiques qui se préoccupe de la description de données conjointes. On cherche par ces méthodes à donner les liens pouvant exister entre les différentes données et à en tirer une information statistique qui permet de décrire de façon plus succincte les principales informations contenues dans ces données.

Nous distinguons l’analyse des données en trois étapes à savoir :

· Identification d'une série de questions à des fin d'analyse et de prises de décisions pertinentes ;

· Identification des colonnes à utiliser ;

· Identifications d'outils mathématiques et statistiques permettant la réalisation de l'analyse des données ;

**III.1°) Identification des colonnes utilisée**

Il s’agit là des colonnes que nous avons choisies pour la réalisation de notre analyse. Notons que nous avons exploité 13 colonnes du jeu de données; ci-dessous, les détails sur ces colonnes.

***III.1.a) Real per Capita GDP Growth Rate (annual %)***

Il s’agit du **Taux de croissance du PIB réel par habitant** (% annuel). Le PIB réel par habitant est une mesure de la **production économique totale** d’un pays **divisé** par le **nombre de personnes** et **ajusté en fonction de l’inflation**. Il est utilisé pour **comparer le niveau** **de vie** entre les pays et dans le temps.

Cet indicateur économique comprend les trois concepts suivants. Vous devez d’abord les comprendre si vous voulez comprendre le PIB par habitant.

Le premier concept est le « **produit intérieur brut** ». Cela mesure tout ce qu’un pays produit en un an. Les composantes du PIB sont la consommation personnelle, l’investissement des entreprises, les dépenses publiques et les exportations moins les importations.

Le second est le « **PIB réel** », c’est-à-dire le PIB sans l’effet des variations de prix. L’inflation augmente le PIB « nominal » régulier, de sorte que le PIB réel est une mesure plus précise lorsque vous voulez comparer une économie au fil du temps.

Le troisième est « **par habitant** », ce qui ne signifie « par personne ». Le PIB réel est divisé par la population d’un pays pour calculer le PIB réel par habitant. C’est le meilleur moyen de comparer des indicateurs économiques comme le PIB pour des pays ayant des tailles de population très différentes.

***III.1.b) Real GDP growth (annual %)***

C’est **la croissance du PIB réel** (% annuel). Dans la section 1.3 on y retrouve la définition du PIB réel.

***III.1.c) Gross domestic product (constant prices US$)***

**Le produit intérieur brut (PIB) réel** est une mesure corrigée de l’inflation qui reflète la valeur de tous les biens et services produits par une économie au cours d’une année donnée. **Le PIB réel est exprimé en prix de l’année de base**. **Il est souvent appelé PIB à prix constants**, **PIB corrigé de l’inflation ou PIB en dollars constants**. En termes simples, le PIB réel mesure la production économique totale d’un pays et est ajusté en fonction des variations de prix.

Principaux points à retenir

* Le produit intérieur brut réel (PIB réel) est une mesure corrigée de l’inflation qui reflète la valeur de tous les biens et services produits par une économie au cours d’une année donnée ;
* Le PIB réel est exprimé en prix de l’année de base et est souvent appelé PIB à prix constants, corrigé de l’inflation ou PIB en dollars constants ;
* Le PIB réel rend la comparaison du PIB d’une année à l’autre et de différentes années plus significatives parce qu’elle montre des comparaisons pour la quantité et la valeur des biens et services ;
* Le PIB réel est calculé en divisant le PIB nominal par un déflateur du PIB.

**PIB réel = où : PIB =** Produit intérieur brut ; **R =** Déflateur du PIB

***III.1.d) Gross domestic product current prices (current US$)***

**Il s’agit du PIB nominal**. **Les gouvernements utilisent à la fois le PIB nominal et le PIB réel comme paramètres pour analyser la croissance économique et le pouvoir d’achat au fil du temps**. Cela se fait à l’aide du déflateur des prix du PIB (également appelé déflateur implicite des prix), qui mesure les variations des prix de tous les biens et services produits dans une économie.

Étant donné que le PIB est l’une des mesures les plus importantes pour évaluer l’activité économique, la stabilité et la croissance des biens et services dans une économie, il est généralement examiné sous deux angles : nominal et réel. Le PIB nominal est une évaluation macroéconomique de la valeur des biens et services utilisant les prix courants dans sa mesure ; on l’appelle aussi le PIB en dollars courants.

**La principale différence entre le PIB nominal et le PIB réel est la prise en compte de l’inflation**. Étant donné que le PIB nominal est calculé à l’aide des prix courants, il n’est pas nécessaire de procéder à des ajustements en fonction de l’inflation. Cela rend les comparaisons d’un trimestre à l’autre et d’une année à l’autre beaucoup plus simples, bien que moins pertinentes, à calculer et à analyser.

***III.1.e) Gross capital formation (current US$)***

La FBCF, ou formation brute de capital fixe, **est l'agrégat qui mesure, en comptabilité nationale, l'investissement (acquisition de biens de production) en capital fixe des différents agents économiques résidents.**

Le capital fixe **est l'ensemble des actifs corporels ou incorporels destinés à être utilisés dans le processus de production pendant au moins un an** (ce sont des biens durables).

La FBCF des entreprises et des administrations publiques (État et autres collectivités), consiste en l'achat :

Depuis 1999, la mesure de la FBCF inclut également certains investissements immatériels comme les dépenses de logiciels, auparavant considérées comme des consommations intermédiaires. Ces investissements immatériels sont ceux qui connaissent la croissance la plus rapide de nos jours, et ne sont pourtant que peu pris en compte dans le calcul.

Une partie de la FBCF inclut le renouvellement des équipements usés ou obsolètes. Si l'on retire de la formation brute de capital fixe cet investissement de remplacement qui sert à pallier l'usure et l'obsolescence (ou amortissement), on obtient la Formation nette de capital fixe (FNCF, ou investissement nouveau), une mesure assez peu utilisée.

***III.1.f) Gross capital formation Private sector (current US$)***

**Formation brute de capital fixe dans le secteur privé** (% du PIB). L’investissement privé couvre les dépenses brutes du secteur privé (y compris les organismes privés à but non lucratif) pour les ajouts à ses actifs nationaux immobilisés.

***III.1.g) Gross capital formation Public sector (current US$)***

Il s’agit de la **FBCF des entreprises et des administrations publiques** qui consiste en l'achat :

* D’actifs fixes (biens d'équipement, logements, bâtiments) acquis pour être utilisés au moins un an dans le processus de production1 ;
* De biens et services incorporés au capital acquis, aux terrains, et aux actifs incorporels.

***III.1.h) Exports of goods and services (current US$)***

**Les exportations sont des biens et des services qui sont produits dans un pays et vendus à des acheteurs dans un autre.** Les exportations, ainsi que les importations, constituent le commerce international. Au lieu de se confiner à l’intérieur de ses frontières géographiques, les pays recherchent souvent intentionnellement des marchés extérieurs dans le monde entier pour le commerce, ce qui permet d’accroître les revenus et les possibilités transactionnelles.

**Les entreprises mesurent souvent leurs exportations nettes**, c’est-à-dire leurs **exportations totales moins leurs importations totales**. **Les exportations nettes sont une composante de la mesure du produit intérieur brut (PIB) d’un pays**, de sorte que les exportations jouent un rôle dans la détermination du bien-être financier et économique d’un pays.

***III.1.i) Imports of goods and services (current US$)***

**Une importation est un bien ou un service acheté dans un pays et qui a été produit dans un autre**. Les importations et les exportations sont les composantes du commerce international. Si la valeur des importations d’un pays dépasse la valeur de ses exportations, **le pays a une balance commerciale négative**, **également connue sous le nom de déficit commercial.**

Les pays sont plus susceptibles d’importer des biens ou des services que leurs industries nationales ne peuvent pas produire aussi efficacement ou à moindre coût que le pays exportateur. Les pays peuvent également importer des matières premières ou des produits qui ne sont pas disponibles à l’intérieur de leurs frontières. Par exemple, de nombreux pays importent du pétrole parce qu’ils ne peuvent pas le produire sur le marché intérieur ou parce qu’ils ne peuvent pas produire suffisamment pour répondre à la demande.

***III.1.j) Household final consumption expéditeur (% of GDP)***

**La dépense de consommation finale des ménages par habitant** (consommation privée par habitant) est calculée à l’aide de la consommation privée en prix constants de 2015 et des estimations démographiques de la Banque mondiale. **La dépense de consommation finale des ménages est la valeur marchande de tous les biens et services, y compris les produits durables (tels que les voitures, les machines à laver et les ordinateurs personnels), achetés par les ménages**. Il exclut les achats de logements, mais comprend le loyer imputé des logements occupés par leur propriétaire. Il comprend également les paiements et les frais aux gouvernements pour obtenir des permis et des licences. Ici, les dépenses de consommation des ménages comprennent les dépenses des institutions sans but lucratif au service des ménages, même lorsqu’elles sont déclarées séparément par le pays. Les données sont exprimées en dollars américains constants de 2015.

***III.1.k) Inflation consumer prices (annual %)***

L’inflation, telle que mesurée par l’indice des prix à la consommation, reflète la variation annuelle en pourcentage du coût pour le consommateur moyen de l’acquisition d’un panier de biens et de services qui peut être fixé ou modifié à des intervalles précis, par exemple annuellement. La formule de **Laspeyres** est généralement utilisée.

***III.1.l) Country***

Cette colonne désigne le pays pour lequel les analyses sont observées.

***III.1.m) Year***

Il s’agit de l’année d’observation pour les mesures effectuées.

**III.2°) Stratégies d'analyse à des fins d’aide à la prise de décisions**

Dans cette partie, il s’agit d’établir un plan d’analyse dans le domaine de l’économie pour les pays africains.

1) Etude générale sur l'évolution économique des cinq grandes régions de l'Afrique(est, ouest, nord et sud, centre) s'étalant sur la période 1980-2022

Dans cette partie d’analyse, il est question de se faire une idée de la croissance économique des pays de l’Afrique de l’Ouest, de l’Est, du Nord, du Sud et du Centre sur la période allant de 1980 à 2022; De surcroît, les indicateurs qui serviront d’évaluation sont: Le taux de croissance du PIB réel par habitants, la croissance du PIB réel, l’importation et l’exportation des biens et services.

2.) Etude comparative de l'évolution économique des régions de l'Afrique (est, ouest, nord, sud et centre) s'étalant sur la période [les années 2000, ou les années 1900, à partir de 2009 ] sur un aspect bien précis (PIB réel).

Pour cette partie, il s’agit de comparer la croissance de l’économie entre l’Afrique de l’Ouest et de l’Afrique centrale en ce qui concerne le PIB réel car le PIB réel est une mesure plus précise lorsque nous voulons comparer une économie au fil du temps, ainsi que la formation brute du capital fixe dans le secteur privé et la formation brute du capital fixe dans le secteur public, tout ceci, analysé avant les années 2000 puis après les années 2000.

3.) Comparaison de l'évolution économique entre les pays francophones et les pays anglophones sur les 5 dernières années.

Les aspects considérés dans ce cas sont : le taux de croissance du PIB réel, le PIB nominal, la formation brute du capital fixe dans le secteur privé et la formation brute du capital fixe dans le secteur public

4.) étude de la croissance économique pour le cas du Sénégal (pour les années 2000 et pour les années 1900).

Plus concrètement, cette partie fait état d’une approche de l’économie du Sénégal à travers la formation brute du capital fixe dans le secteur privé, la formation brute du capital fixe dans le secteur public, la dépense de consommation finale des ménages par habitant, le tout sur la période d’avant la guerre en Ukraine et celle d’après la guerre en Ukraine.

5) étude comparative sur l'inflation pré-covid, pendant covid et post-covid pour les pays formant l’UEMOA.

pré-covid → 2015-2019

pendant covid→ 2020

post-covid→ 2021 et 2022

**III.3°) Identifications d'outils mathématiques et statistiques permettant la réalisation de l'analyse des données**

**IV°) Visualisation des données**

La visualisation des données (ou data visualization ou dataviz en anglais) désigne le fait de représenter visuellement ses data pour pouvoir déceler et comprendre des informations, les données brutes étant difficilement interprétables et exploitables. Ce processus se fait par des outils analytiques spécifiques et se matérialise par des tableaux (type Excel), des graphiques, des cartes visuelles ou même des infographies regroupées dans des dashboards (tableaux de bord).

En classant, segmentant et scénarisant les données visuellement, une entreprise peut déceler des informations auparavant inaccessibles en un simple coup d’œil. La visualisation des données permet donc à toute organisation de piloter son activité plus efficacement en adoptant une stratégie data-driven et agile.

**IV.1) Les outils de visualisation de données**

v Power BI : l’outil dataviz de référence par Microsoft

Le premier outil que nous avons à présenter est évidemment Microsoft Power BI. Il s’agit sûrement l’un des outils de Dataviz le plus connu du marché. Il couvre l’ensemble des fonctionnalités que l’on peut attendre d’un tel logiciel, pour analyser tes données et améliorer ton processus décisionnel.

Son intégration avec les fichiers Excel permet de gérer l’importation de données très facilement.

v Tableau, l’analytique pour tous

Aux côtés de Power BI dans le classement des solutions Dataviz les plus connues, on trouve Tableau. Tableau propose des solutions pour toutes les équipes, quelles que soient leurs tailles, leurs besoins et leurs niveaux de compétences.

Tableau Public est une plateforme gratuite où n'importe qui peut créer une visualisation de données. Les données peuvent être importées à partir d'une feuille de calcul ou saisies manuellement.

v Qlik, exploite la Data au profit de ton entreprise

Qlik est une solution d’analytique très complète. Grâce à cet outil, tu vas pouvoir mettre en place des visualisations en libre-service, des tableaux de bord interactifs, faire du reporting, de l’analytique personnalisée et embarquée, de l’analytique avancée à l’aide de l’IA et même de la recherche en langage naturel et de l’analytique conversationnelle.

Cette solution se prête à de nombreux secteurs d’utilisation : la finance, les ressources humaines, le marketing, l’IT, le supply chain, le commerce, etc…

v Visme

Visme est un outil de datavisualisation qui va au-delà des données.

À l'aide d'un ensemble complet d'outils, y compris un créateur de graphiques, des modèles de graphiques et une bibliothèque de widgets de données, vous pouvez facilement créer des visualisations de données de différents types.

La meilleure partie est que vous pouvez ajouter ces visualisations à n'importe quel projet sur lequel vous travaillez dans Visme, comme une proposition, une présentation, un rapport ou une infographie.

v Google Charts

Il s'agit du seul outil de la liste qui nécessite des connaissances en codage pour créer des data visualisation. Si vous ne savez pas comment coder, comprendre comment utiliser Google Charts peut nécessiter une courbe d'apprentissage élevée.

Il est facile d'intégrer n'importe quel graphique Google dans un site internet ou une plateforme en ligne. Les données peuvent être statiques ou dynamiques selon la source.

La meilleure partie de Google Charts est qu'il est entièrement gratuit et disponible pour des projets personnels et commerciaux.

v Datapine

Datapine est un outil tout-en-un de veille économique et de data visualisation conçu avec une interface conviviale qui permet à chacun de participer au processus d'analyse, sans avoir besoin de compétences techniques.

Datapine sert d'outil collaboratif, car il comprend plusieurs options de partage pour que chaque membre de l'équipe ou partie prenante concernée reste engagé avec les données de manière sécurisée.

v Rapid Miner

A la base, RapidMiner est une plateforme de machine learning intégrée. Open source et très orientée no code, elle va de la création des modèles et applications d'IA jusqu'à leur déploiement. Il fait partie des outils de visualisation de données car il contient une partie où on est à mesure de représenter les données sous des formes graphiques.

**IV.2) Les types de graphes**

Que ce soit de manière continue ou discontinue, les diagrammes peuvent visualiser n'importe quel type de données pour les analyser et les interpréter efficacement. Il existe de nombreux types de graphiques comme les graphiques en aires, en courbes, à barres, empilés, à secteurs, les nuages de points, les graphiques combinés, en entonnoir, Web, etc. Chaque graphique offre un large éventail d'options d'interactions utilisateur pour une analyse plus approfondie et contextuelle.

Parmi les types de graphiques connus et largement utilisés se trouvent les types suivants :

v Graphique en courbes

Il peut être utilisé pour visualiser les tendances sur n'importe quelle période.

v Graphique en aires

Les graphiques en aires ombrent l'aire située sous les lignes et vous aident ainsi à comparer plus facilement les magnitudes des données.

v Graphique à barres

Les données sont affichées sous forme de barres individuelles dont la hauteur est proportionnelle à la valeur indiquée. Le graphique à barres est utile pour comparer les données classées dans des groupes distincts.

v Graphique à barres empilées

Effectuez une analyse comparative des parties dans un ensemble de données à l'aide d'un graphique empilé. Il indique la valeur de chaque valeur par rapport au total dans les différentes catégories.

v Graphique en secteurs

Ce graphique est plus adapté pour représenter des données en pourcentage ou proportionnelles. Il peut être utile pour analyser et déterminer la contribution des parties dans un ensemble.

v Graphique en anneau

Autrement connu sous le nom de diagramme « donut », il est également utile pour présenter les données sous forme de somme de ses parties, en pourcentage.

v Nuage de points

L'utilisation type du nuage de points consiste à représenter des données sporadiques avec des intervalles irréguliers. Il permet de comparer deux axes numériques, contrairement au graphique en courbes dont l'axe n'est jamais numérique.

v Graphique en bulles

Il est utile pour visualiser les données en mettant en évidence l'ampleur de la valeur des données avec des tailles proportionnelles. De plus, le graphique en bulles peut également être utilisé de manière interchangeable avec le nuage de points.

v Graphique en entonnoir

Ce graphique est utilisé pour représenter la réduction ou le flux progressif d'une mesure d'activité à différentes phases.

v Carte thermique

Ce graphique quadrillé en deux dimensions est utile pour afficher les valeurs avec des couleurs d'intensités variables, afin de faciliter la visualisation, la catégorisation et le filtrage des données.

v Cartes géographiques

Ces graphiques sont utiles pour visualiser les données géographiques par le biais de visualisations interactives au sein et entre les villes, États, pays et continents de l'ensemble du paysage mondial.

**V°) Création de modèle**

Il existe plusieurs algorithmes en machine Learning ; En ce qui concerne le modèle de régression, nous avons :

§ Linear Regression

§ Decision Tree Regression

§ Random Forest

§ Polynomial Regression

**1.)** **Linear Regression**

C'est l'un des algorithmes de régression les plus utilisés en Machine Learning. Une variable significative de l'ensemble de données est choisie pour prédire les variables de sortie (valeurs futures). L'algorithme de régression linéaire est utilisé si les étiquettes sont continues, comme le nombre de vols quotidiens depuis un aéroport, etc. La représentation de la régression linéaire est y = b\*x + c.

Dans la représentation ci-dessus, 'y' est la variable indépendante, tandis que 'x' est la variable dépendante. Lorsque vous tracez la régression linéaire, la pente de la ligne qui nous fournit les variables de sortie est appelée « b », et « c » est son interception. Les algorithmes de régression linéaire supposent qu'il existe une relation linéaire entre l'entrée et la sortie. Si les variables dépendantes et indépendantes ne sont pas tracées sur la même ligne en régression linéaire, il y aura une perte de sortie. La perte de sortie en régression linéaire peut être calculée comme suit :

Loss function : (Predicted output – actual output).

**2.)** **Decision Tree Regression**

La régression non linéaire dans Machine Learning peut être effectuée à l'aide de la régression par arbre de décision. La fonction principale de l'algorithme de régression de l'arbre de décision est de diviser l'ensemble de données en ensembles plus petits. Les sous-ensembles de l'ensemble de données sont créés pour tracer la valeur de tout point de données qui se connecte à l'énoncé du problème. Le fractionnement de l'ensemble de données par cet algorithme aboutit à un arbre de décision qui a des nœuds de décision et des nœuds feuilles. Les experts en ML préfèrent ce modèle dans les cas où il n'y a pas suffisamment de changement dans l'ensemble de données.

3.) **Random Forest**

La forêt aléatoire est également un algorithme largement utilisé pour la régression non linéaire dans l'apprentissage automatique. Contrairement à la régression par arbre de décision (arbre unique), une forêt aléatoire utilise plusieurs arbres de décision pour prédire la sortie. Des points de données aléatoires sont sélectionnés à partir de l'ensemble de données donné (par exemple, k points de données sont sélectionnés) et un arbre de décision est construit avec eux via cet algorithme. Plusieurs arbres de décision sont ensuite modélisés qui prédisent la valeur de tout nouveau point de données.

Puisqu'il existe plusieurs arbres de décision, plusieurs valeurs de sortie seront prédites via un algorithme de forêt aléatoire. Vous devez trouver la moyenne de toutes les valeurs prédites pour un nouveau point de données pour calculer la sortie finale. Le seul inconvénient de l'utilisation d'un algorithme de forêt aléatoire est qu'il nécessite plus d'entrées en termes de formation. Cela se produit en raison du grand nombre d'arbres de décision cartographiés sous cet algorithme, car il nécessite plus de puissance de calcul.

## **4.) Polynomial Regression**

La régression polynomiale est un algorithme de régression qui modélise la relation entre une variable indépendante (x) et une variable dépendante (y) sous la forme d'un polynôme de degré n. L'équation de la régression polynomiale est la suivante :

y= b 0 b 1 x 1 b 2 x 1 2 b 2 x 1 3 …… b n x 1 n

Il est également connu sous le nom de scénario spécial de régression linéaire multiple dans l'apprentissage automatique. Parce que pour en faire une régression polynomiale, certains termes polynomiaux sont ajoutés à l'équation de régression linéaire multiple. C'est un modèle linéaire qui a été modifié pour améliorer la précision. L'ensemble de données utilisé pour l'entraînement à la régression polynomiale est non linéaire. Pour s'adapter aux fonctions et aux ensembles de données non linéaires et compliqués. Les caractéristiques d'origine sont transformées en caractéristiques polynomiales du degré requis (2, 3, …, n) puis modélisées à l'aide d'un modèle linéaire.

**VI°) PREDICTION**

La prédiction signifie essentiellement prédire un résultat futur, de la même manière que ce qui est accompli dans l'apprentissage automatique. Il fait référence à la sortie d'un algorithme post-formation sur un ensemble de données historiques. Après avoir appris l'ensemble de données historiques, la prédiction permet de prévoir le résultat d'un nouvel ensemble de données. L'algorithme génère des valeurs probables pour les variables inconnues.

**VII°) TOUT SAVOIR SUR RAPID MINER**

***RapidMiner : la plateforme de machine Learning "tout-en-un"***

RapidMiner est une plateforme de machine learning intégrée. Open source et très orientée no code, elle va de la création des modèles et applications d'IA jusqu'à leur déploiement.

RapidMiner recouvre l'ensemble du cycle de vie du machine learning : intégration et préparation des données d'apprentissage, création et déploiement des modèles de machine learning, management et monitoring des modèles en production (MLOps), développement sans code d'applications d'IA, explicabilité des résultats... RapidMiner intègre plus de 1500 algorithmes et fonctions.

****

**VII.1) Installation**

§ L'installation de RapidMiner Studio est très facile sur Windows (testé avec Windows 7 x64, Windows 8 x64 et Windows 10 technical preview x64), à partir de l'exécutable téléchargé sur la page de votre compte RapidMiner (l'inscription est obligatoire).

§ L'installation de RapidMiner Studio sur Mac OS X s'est révélée être compliquée, du fait de ses dépendances en Java. Le système intègre Java 1.8 (OS X 10.10.1) et RapidMiner demande la version 1.7 ne détectant que la version 1.6.

**VII.2) Un jeu d'outils complet**

Tout d'abord, il me semble important de dire que RapidMiner Studio et RapidMiner Server qui le complète est un jeu d'outils complet, plus qu'un logiciel spécifique. Le site web de RapidMiner nous dit que RapidMiner vous laisse parcourir et exécuter avec facilité parmi plus de 1500 opérations.

Du fait de cette complexité, je n'explorerais et décrirais que certaines des fonctionnalités de RapidMiner Studio. Je m'intéresserais donc tout d'abord au "mining" de texte classique, à partir de documents locaux ; j'explorerais ensuite comment il est possible avec cet outil, d'importer, de transformer et d'analyser des tweets (tweets mining).

Les principales forces de RapidMiner selon ses éditeurs sont :

§ Un environnement visuel « code-free » que l'on peut utiliser sans maitriser de langage de programmation

§ Sa disponibilité sur les principaux systèmes d'exploitation et plateformes

§ Sa principale fonction est le "design" de processus d'analyses

§ Analyses prédictives

§ Chargement de données

§ Transformation de données

§ Modélisation de données

§ Visualisation de données

§ Extension par API

§ De nombreuses sources de données : Excel, Access, Oracle, IBM DB2, Microsoft SQL, Sybase, Ingres, MySQL, Postgres, SPSS, dBase, Text files, et d'autres encore ...

§ RapidMiner vous permet de travailler avec différents types et tailles de sources de données.

**VII.3) Type de données**

Avec RapidMiner, il est possible d’importer tout type de données que ce soit des données enregistrées sur des fichiers ou des données hébergées sur des bases de données externes SQL et no SQL. D’après le site de RapidMiner, il est notamment possible d’intégrer des entiers (Integer), des réels (real), des données numériques, du binomial, du polynomial, des données nominales, ainsi que des dates.

De cette manière, on peut intégrer des données opérationnelles ou transactionnelles (données de coûts par exemple), des données non opérationnelles (données prévisionnelles par exemple) et des métadonnées (données concernant les données elles-mêmes).

RapidMiner offre une vaste gamme d'outils pour le traitement de données que ce soit pour le prétraitement ou pour le traitement. Ces outils, ou opérateurs, permettent d’effectuer un grand nombre d’actions qui vont de l'importation et l’exportation des bases de données, au filtrage et la réorganisation des données.

Le rayon des possibilités que cet outil offre est tellement vaste qu’il est impossible de les résumer dans cet article, cependant on se limitera à en décrire ceux qui apparaissent être les principaux. Pour ce faire il y a deux concepts clés à retenir, les concepts de data blending et de data cleansing.

v Le data blending (littéralement “fusion de données”) est, comme le nom le suggère, l’action de prendre des données de différentes bases de données et de les fondre ensemble. RapidMiner (à la version actuelle) offre 82 opérateurs liés à cette pratique, voici les plus basiques :

§ Union : cet opérateur permet de mettre ensemble deux bases de données. Le résultat est une base de données unifiée avec les données de la première plus les données de la deuxième.

§ Join : cet opérateur permet de mettre ensemble deux bases de données sur la base d'attributs clés que les deux ont en commun. Par exemple, si on a une base de données avec une liste de produits et une base de données avec les produits vendus et que les deux ont l'attributs commun "ID produit", alors les deux bases de données peuvent être mises ensemble en prenant "ID produit" comme point de référence. Le résultat est une base de données unifiée où les données sont mises en commun en minimisant la possibilité d’avoir des données manquantes comme dans le cas de “union”.

v Le data cleansing (nettoyage de données) est une pratique, surtout utilisée en phase de prétraitement, qui permet de se libérer des données “inutiles” comme les données manquantes, des données étranges qui peuvent être causés par des erreurs de measurements, ou bien qui sont triviaux pour le type d’analyse qu’on est en train d’effectuer, RapidMiner offre 28 opérateurs pour le data cleansing, voici certains exemples :

§ Replace Missing Values : comme on peut le comprendre depuis le nom, cela permet de remplacer les données manquantes avec des données fictives qui n’ont pas d’influence relevante dans les résultats.

§ Normalize : ceci permet d’uniformiser différentes données numériques en les mettant sur la même échelle, cela permet par exemple de trouver plus facilement des valeurs étranges avec l’aide de l’opérateur Detect Outlier (Distances)

**VII.4.) Visualisations offertes**

En cliquant sur “open visualizations”, l'onglet visualizations dans la barre latérale s’ouvre. On peut choisir plusieurs types de visualisations. Les deux grandes catégories sont les graphiques (“charts”) et les cartes (“maps”). Dans la première, on retrouve 36 types de graphiques différents. Ils permettent de regarder les données de manière différente, en fonction de ce qu’on cherche à mettre en avant.



**VII.5) Points forts et limites**

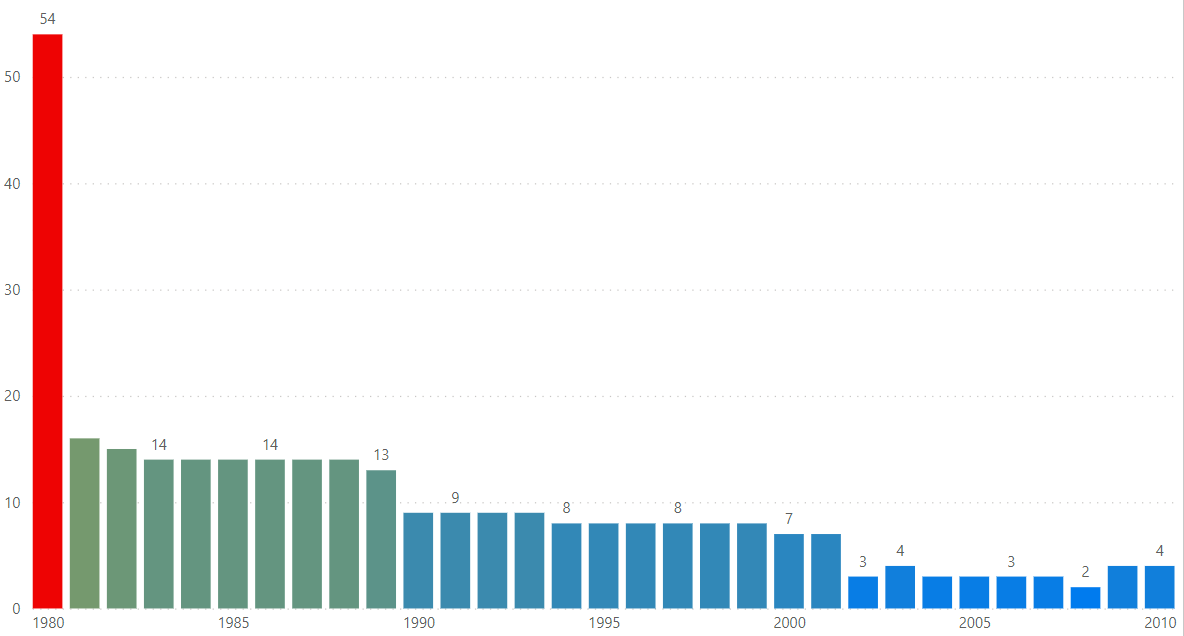
RapidMiner est certainement un outil qui offre une vaste gamme de possibilités et qui est flexible. Il possède une interface qui permet de visualiser le processus de traitement de données à travers un système de drag and drop où les opérateurs sont trascinés sur l’écran et mis en lien, cela donne ainsi au logiciel une interface intuitive et plaisante à utiliser. La quantité d’opérateurs à disposition permet aussi d’effectuer des traitements complexes sur les bases de données et d’extraire les informations qu'on nécessite de façon rapide et sans besoin de connaissances de programmation. Il y a également la possibilité de visualiser les données à travers nombreux modèles et graphiques rendant tout l’ensemble intéressant. Le fait d’avoir à disposition une version gratuite de RapidMiner pour l’éducation en permettant d’explorer toutes les fonctionnalités sans limitations, est aussi un point positif.

Cependant, malgré la présence d’un tutoriel qui nous aide à effectuer les premiers pas avec RapidMiner en nous familiarisant à certaines opérations de base, l’outil est très difficile à maîtriser, surtout pour ceux qui s’approchent pour la première fois du domaine de l’analyse et de l’extraction de données, aussi les instructions offertes par le site du logiciel ne sont pas toujours faciles à comprendre. Dans ce cas, la grande disponibilité d’opérations possibles se heurte à un temps d’apprentissage très long et une complexité qui peut effrayer les débutants et cela malgré la possibilité de visualiser les opérations.

**II. Prétraitement des données**

**1 .Identification des données manquantes**

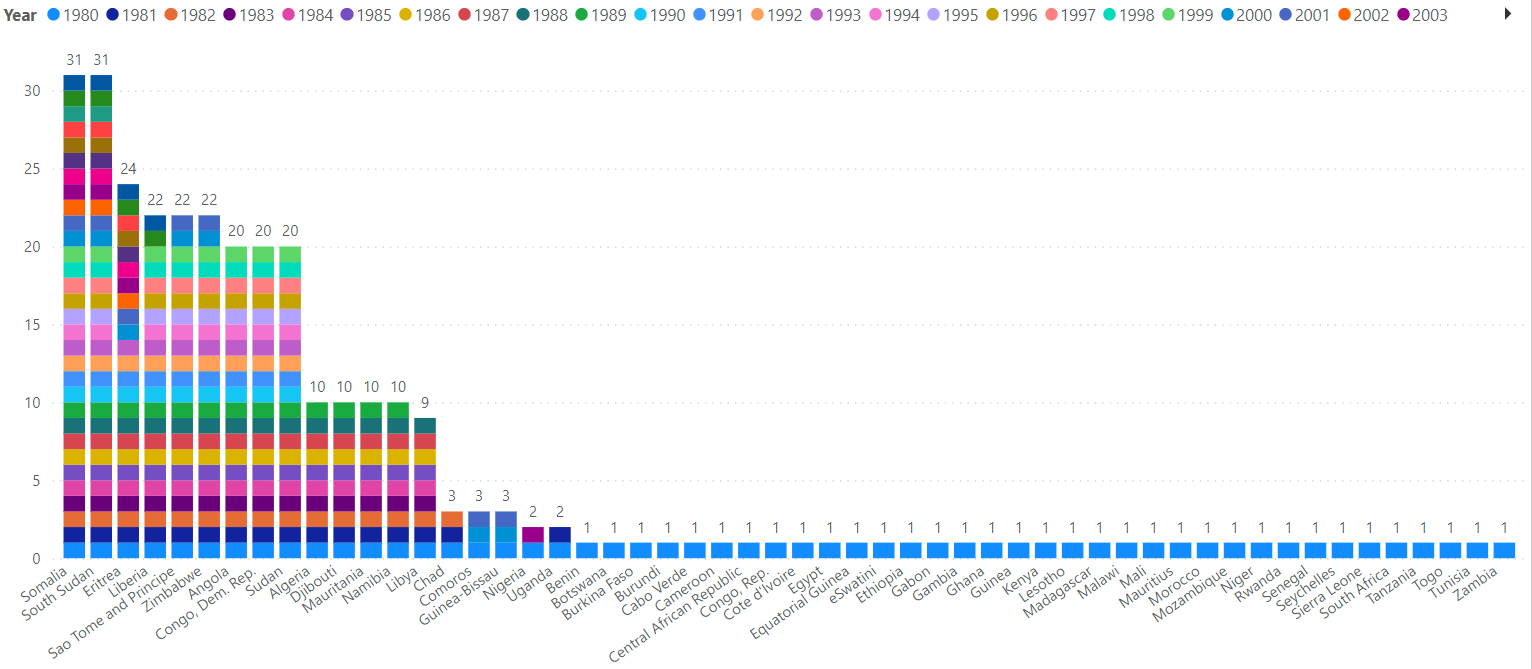
Notre jeu de données possède en total 309 lignes comportant des données manquantes, pour un total de **3193 valeurs (4.43 %** des données). Seules les colonnes **‘Year’** et **‘Country’** sont exemptes de données manquantes. Le pourcentage de lignes comportant des données manquantes varient entre  **0.4737 %** et **7.062877 %.**

****

**Repartition des données manquantes en fonction des années**

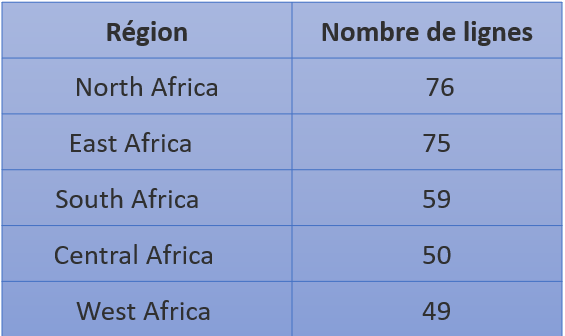
Comme on peut s’en apercevoir, l’année qui porte le plus de lignes comportant des valeurs manquantes est 1980 (54 lignes) et celle qui en comporte le moins est l’année 2008 (2 lignes). Aussi une autre observation que nous faisons est qu’après 2010 notre jeux de données ne comporte plus de données manquantes. Cette observation peut nous guider quant au traitement qui sera appliqué à ces données manquantes.

# 



**Répartition des données manquantes en fonction du pays et de l’année**

L’unedes premières observations concerne les pays comportant des données manquantes. En effet, tous les pays sans exception possèdent des lignes comportant des données manquantes. Les pays avec le plus de lignes comportant des données manquantes sont la **Somalie** et le **Sud Soudan,** certains pays comme le **Burkina Faso**, la **Côte d’Ivoire** le **Sénégal** n’en comporte qu’une seule. Une observation de ces données manquantes selon les régions de l’Afrique révèle que que ceux sont les pays d’Afrique du Nord qui en possède le plus (76 lignes) suivis par ceux de l’Afrique Oriental (75 lignes), l’ Afrique Australe (59 lignes), Afrique Centrale(50) et l’Afrique Occidentale (49 lignes).



**2. Remplacement / suppression des valeurs manquantes**

Les valeurs manquantes de notre jeu de données n’appartenant qu'à des colonnes de types numériques nous avons décidé de les remplacer par la moyenne de toutes les valeurs de leur colonnes d’appartenance.

**III. Visualisation**