



**Université Abdelmalek Essaadi**  
**Ecole Nationale des Sciences Appliquées**  
**Tétouan**



**PROJET :**

**Prédiction du secteur le plus rentable**

**Au Maroc**



**Encadré par :**

**Pr. Al Amrani Yassine**

**Réalisé par :**

**Hassan Fares**

**Nassiri Mosab**

**EL ABBASS Ossama**

**Zayd el ouaragli**

**Aly Guinga**

## Remerciement :

Nous tenons, avant de présenter notre travail, exprimer notre grande reconnaissance envers les personnes qui nous ont, de près ou de loin, apporter leurs soutiens. Qu'ils trouvent ici collectivement et individuellement l'expression de toute notre gratitude.

Nous tenons à remercier tout particulièrement et témoigner toute notre reconnaissance à **Pr. Al Amrani Yassine** pour l'expérience enrichissante et pleine d'intérêt qu'il nous fait vivre durant la période des cours et pour tous les conseils et les informations qu'il nous a prodigués.

## Liste des figures :

Figure 1: La distribution des différents indicateurs par rapport aux autres.....	11
Figure 2: la variation de la production nationale par rapport à chaque section. ....	12
Figure 3: la variation de l'excédent brut d'exploitation par rapport à chaque secteur. ....	12
Figure 4: la représentation des différents indicateurs par rapport au secteur du tourisme. ..	13
Figure 5: la représentation des différents indicateurs par rapport au secteur BTP.....	13
Figure 6: la variation du salaire brut par rapport à chaque secteur.....	14
Figure 7 : corrélation entre les indicateurs .....	15
Figure 8: Les indicateurs importants.. ....	21
Figure 9: Estimation de la mesure de la compétence par secteur. ....	22

## Les des tableaux :

Table 1: Les models utilise et leur score. ....	20
---	----

# Sommaire :

Remerciement : .....	2
Liste des figures : .....	3
Les des tableaux : .....	4
Sommaire : .....	5
INTRODUCTION : .....	1
I. Cadre du projet.....	2
A. Présentation de l'équipe : .....	2
B. Problématique : .....	2
C. Les secteurs les plus importants au Maroc : .....	3
1. Agriculture : .....	3
2. Industrie : .....	3
3. BTP (Bâtiment et Travaux Publics) : .....	3
4. Énergie : .....	3
5. Tourisme : .....	3
6. Commerce : .....	4
7. Services : .....	4
D. Les indicateurs : .....	4
1. Capitalisation boursière : .....	4
2. Consommation Finale : .....	4
3. Consommation intermédiaire : .....	4
4. Excédent brut d'exploitation (EBE) : .....	5
5. Exportations et Importations : .....	5
6. Impôts sur les produits : .....	5
7. PIB (Produit Intérieur Brut) : .....	5
8. Production nationale : .....	5
9. Salaires bruts : .....	5
10. Volume des échanges : .....	6
11. Actif occupé : .....	6
12. FBCF (Formation Brute de Capital Fixe) : .....	6
II. Cycle de vie d'une data analyste .....	6
A. Les bibliothèques : .....	6
B. Collecte des données : .....	8

III.	Nettoyage et prétraitement des données :.....	8
A.	Prétraitement des données.....	8
B.	Gestion des valeurs manquantes et duplicatas :.....	9
C.	Transformation des données : .....	9
IV.	Exploration des données :.....	10
A.	Définition : .....	10
B.	Préparation des données pour la visualisation :.....	10
C.	Description de chaque visualisation : .....	10
D.	Hypothèse d'expérience :.....	15
V.	Modélisation des données :.....	15
A.	Définition : .....	15
B.	Division des données : .....	16
C.	Sélection ou Ingénierie des Caractéristiques : .....	16
D.	Choix du ou des modèles :.....	19
1.	Régression linéaire: .....	19
2.	Régresseur de Forêt Aléatoire : .....	19
3.	SVR (Support Vector Regressor) : .....	19
E.	Entraînement du modèle : .....	19
VI.	Evaluation :.....	20
A.	Définition : .....	20
B.	Critères d'évaluation :.....	20
C.	Analyse des erreurs : .....	20
D.	Interprétation des résultats :.....	21
E.	Limitations et recommandations : .....	22
1.	Recommandations: .....	22
2.	Limites : .....	23
	Conclusion : .....	23
	Les références :.....	24

## INTRODUCTION :

Le Maroc s'est engagé dans une stratégie de transformation profonde de son économie moyennant des stratégies sectorielles visant la modernisation de son appareil productif et le renforcement de ses performances et de sa résilience.

Ainsi, l'intérêt porté par notre pays à la modernisation accélérée des activités relevant du secteur primaire obéit autant à une logique de consolidation des ressorts sectoriels de la croissance de l'économie nationale et à la mobilisation de gisements additionnels d'emplois qu'à l'impératif d'assurer une meilleure valorisation des ressources naturelles et le renforcement de leur durabilité.






Par ailleurs, les options industrielles adoptées par le Maroc au cours des deux dernières décennies ont enclenché une dynamique qui a amélioré l'attractivité du pays aux investissements étrangers et a favorisé l'émergence des métiers mondiaux du Maroc. Ces choix devraient être consolidés pour induire les changements structurels attendus notamment en termes de création conséquente de valeur ajoutée et d'emplois. Il s'agit, à cet effet, d'une grande ambition que le Maroc s'est fixée pour accélérer l'éclosion d'une industrie nationale compétitive et résiliente et répondre, par ricochet, aux besoins de développement économique et social du pays.

En outre, le Maroc, en tant que nation en développement, recherche constamment des moyens d'améliorer son économie et d'identifier les secteurs les plus prometteurs pour stimuler la croissance. Dans ce contexte, ce projet s'inscrit dans une démarche visant à exploiter les principes du data mining et l'utilisation du langage de programmation Python pour prédire les secteurs économiques les plus rentables au Maroc.

# I. Cadre du projet

## A. Présentation de l'équipe :

L'équipe est composée de cinq élèves-ingénieurs de la filière Big Data et Intelligence Artificielle de l'Ecole Nationale des Sciences Appliquées de Tétouan. Il s'agit notamment de :

-  Hassan Fares
-  Nassiri Mosab
-  EL ABBASS Ossama
-  Zayd el ouaragli
-  Aly Guinga

En tant qu'élèves-ingénieurs en 1<sup>ère</sup> année cycle ingénieur, relever un tel défi nous honore et est un moyen d'améliorer nos compétences afin d'être aptes à faire face à nos futures missions.

## B. Problématique :

La problématique au cœur de ce projet réside dans la synergie entre les enseignements en data mining et les compétences en programmation Python pour l'analyse approfondie des données économiques recueillies auprès de divers secteurs au Maroc. Comment mettre en œuvre de manière cohérente ces connaissances pour développer un modèle de prédiction robuste et précis afin d'identifier les secteurs économiques les plus rentables dans le contexte marocain ? La réalisation de cette tâche complexe nécessite à la fois une maîtrise technique des outils de data mining et une compréhension approfondie des nuances économiques du Maroc. Comment répondre aux défis posés par la diversité des données sectorielles tout en maximisant la pertinence et la fiabilité du modèle de prédiction ? En traitant ces questions, le projet a pour objectif de développer une solution innovante qui peut aider à prendre des décisions économiques en identifiant les opportunités de croissance les plus prometteuses pour le Maroc.



## C. Les secteurs les plus importants au Maroc :

### 1. Agriculture :

L'agriculture englobe les activités liées à la culture des terres, à l'élevage du bétail, à la production de cultures vivrières, de fruits, de légumes, de céréales et d'autres cultures. C'est le secteur de base qui fournit des matières premières alimentaires et non alimentaires.

### 2. Industrie :

Le secteur industriel regroupe les activités de transformation des matières premières en biens manufacturés. Cela comprend la production de biens matériels tels que machines, équipements, produits chimiques, textiles, et divers produits manufacturés.

### 3. BTP (Bâtiment et Travaux Publics) :

Le secteur du BTP englobe toutes les activités liées à la construction, à la rénovation et à l'entretien des infrastructures physiques. Cela comprend la construction de bâtiments résidentiels et commerciaux, de routes, de ponts, de barrages, d'aéroports et d'autres projets d'ingénierie.

### 4. Énergie :

Le secteur de l'énergie englobe la production, la distribution et la fourniture d'énergie. Cela comprend l'énergie électrique (produite à partir de sources telles que le charbon, le gaz, l'hydroélectricité, l'énergie éolienne, etc.), ainsi que l'exploration et la production d'énergie fossile et renouvelable.

### 5. Tourisme :

Le secteur du tourisme englobe toutes les activités liées aux voyages et aux loisirs. Cela comprend l'hébergement (hôtels, auberges), la restauration, les transports (aériens, maritimes, terrestres), les activités de loisirs et de divertissement, ainsi que les services connexes tels que les agences de voyage.

## 6. Commerce :

Le secteur du commerce concerne toutes les activités liées à l'achat et à la vente de biens et de services. Il se divise en commerce de détail (vente directe aux consommateurs), commerce de gros (vente en grandes quantités aux détaillants) et distribution. Les formes de commerce incluent le commerce électronique, le commerce traditionnel, etc.

## 7. Services :

Le secteur des services regroupe une variété d'activités qui fournissent des services plutôt que des biens matériels. Cela inclut les services financiers, les services de santé, l'éducation, les technologies de l'information, les services juridiques, les services de conseil, la restauration, et bien d'autres. Les services jouent un rôle essentiel dans l'économie moderne.

# D. Les indicateurs :

## 1. Capitalisation boursière :

La capitalisation boursière représente la valeur totale des actions d'une entreprise cotée en bourse. Elle est essentielle pour évaluer la taille et la valeur d'une entreprise sur le marché financier, ce qui peut influencer les investisseurs et les décisions de gestion.

## 2. Consommation Finale :

La consommation finale mesure la dépense totale des ménages et des institutions sans réinvestissement. Cet indicateur est crucial pour évaluer la demande globale dans une économie, ce qui peut indiquer la santé financière et la stabilité.

## 3. Consommation intermédiaire :

La consommation intermédiaire représente les biens et services utilisés dans le processus de production. Son suivi est important pour évaluer l'efficacité et la productivité d'un secteur,

influençant ainsi les décisions de gestion.

#### 4. Excédent brut d'exploitation (EBE) :

L'EBE mesure la rentabilité d'une entreprise en calculant les revenus générés avant déduction des charges de personnel et des impôts. C'est un indicateur clé pour évaluer la performance financière d'une entreprise.

#### 5. Exportations et Importations :

Ces indicateurs mesurent le volume de biens et de services échangés avec d'autres pays. Ils sont cruciaux pour évaluer la compétitivité d'une économie sur le marché international, impactant ainsi la rentabilité des secteurs orientés vers l'exportation ou dépendants des importations.

#### 6. Impôts sur les produits :

Les impôts sur les produits représentent les taxes liées à la production et à la vente de biens et services. Ils influencent la rentabilité des entreprises en impactant les coûts de production.

#### 7. PIB (Produit Intérieur Brut) :

Le PIB mesure la valeur totale des biens et services produits dans une économie. Il est fondamental pour évaluer la croissance économique et la performance globale d'un pays.

#### 8. Production nationale :

La production nationale représente la quantité totale de biens et de services produits à l'intérieur des frontières d'un pays. Elle est essentielle pour évaluer la capacité productive d'une économie.

#### 9. Salaires bruts :

Les salaires bruts reflètent les coûts liés à la main-d'œuvre. Leur suivi est important pour évaluer la charge salariale d'une entreprise et son impact sur la rentabilité.

## 10. Volume des échanges :

Le volume des échanges mesure la quantité de biens et de services échangés. Il est crucial pour évaluer l'activité économique et la compétitivité d'un secteur sur le marché.

## 11. Actif occupé :

Représente l'ensemble des personnes en âge de travailler qui sont employées, que ce soit dans le secteur formel ou informel de l'économie. Cet indicateur englobe les employés salariés, les travailleurs indépendants, les entrepreneurs, et toute personne contribuant à la production de biens et de services.

## 12. FBCF (Formation Brute de Capital Fixe) :

Évaluer le niveau d'investissement dans une économie. Elle reflète la capacité d'un pays à accroître sa production future, à améliorer son infrastructure et à rester compétitif. Une FBCF élevée est souvent associée à une croissance économique soutenue à long terme.

# II. Cycle de vie d'une data analyste

## A. Les bibliothèques :

Dans ce projet, nous avons utilisé les bibliothèques suivantes :

### ❖ Pandas :

Est une bibliothèque permettant la manipulation et l'analyse des données [1]. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles.

### ❖ Numpy :

Est une bibliothèque destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux [2].

### ❖ Matplotlib :

Matplotlib [3] est une bibliothèque complète pour créer des visualisations statiques, animées et interactives en Python. Matplotlib rend les choses faciles et les choses difficiles possibles :

- Créez des graphiques de qualité de publication.
- Faire des figures interactives qui peuvent zoomer, panoramique, mise à jour.
- Personnalisez le style visuel et la mise en page.
- Exporter vers de nombreux formats de fichiers.

#### ❖ **Scikit-learn(sklearn) [4] :**

Est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle propose dans son framework de nombreuses bibliothèques d'algorithmes à implémenter, clé en main. Ces bibliothèques sont à disposition notamment des data scientists. Cette bibliothèque fournit les bibliothèques suivantes :

##### **sklearn.preprocessing.StandardScaler :**

Est une classe de la bibliothèque scikit-learn qui est utilisée pour standardiser des données, c'est-à-dire les mettre à l'échelle de manière à ce qu'elles aient une moyenne nulle et une variance unitaire.

##### **sklearn.preprocessing.LabelEncoder:**

Est une classe de la bibliothèque scikit-learn utilisée pour convertir des étiquettes de classes textuelles en nombres.

##### **Sklearn.model\_selection :**

Utiliser pour importer train\_test\_split qui permet de diviser un ensemble de données en ensembles d'entraînement et de test.

##### **La classe LinearRegression :**

Est une classe du module linear\_model de scikit-learn. Elle représente un modèle de régression linéaire, qui est un type de modèle linéaire supposant une relation linéaire entre les caractéristiques d'entrée et la variable cible.

##### **sklearn.ensemble.RandomForestRegressor :**

De la bibliothèque scikit-learn est utilisée pour implémenter un modèle de régression basé sur une forêt aléatoire.

##### **sklearn.svm.SVR:**

Est une classe de la bibliothèque scikit-learn qui implémente le modèle de régression par machines à vecteurs de support.

 **sklearn.metrics.r2\_score :**

Est une fonction de scikit-learn est utilisée pour calculer le coefficient de détermination, également appelé le coefficient  $R^2$ . Le coefficient  $R^2$  mesure la proportion de la variance de la variable dépendante qui est prédictible à partir des variables indépendantes dans un modèle de régression. Il donne une indication de la qualité de l'ajustement du modèle aux données.

 **sklearn.model\_selection.cross\_val\_score:**

Est une fonction de la bibliothèque scikit-learn qui est utilisée pour effectuer une validation croisée d'un modèle d'apprentissage automatique.

❖ **Seaborn :**

Est une bibliothèque permettant de créer des graphiques statistiques Elle est basée sur Matplotlib, et s'intègre avec les structures Pandas [5].

## **B. Collecte des données :**

Les données utilisées dans ce projet sont issues du site du Ministère [6], Nous avons importé des fichiers Excel contenant des statistiques sur les secteurs du Maroc par rapport à chaque indice entre 2010 et 2020. Nous avons rassemblé ces données dans un seul ensemble de données afin de les analyser de manière approfondie.

# **III. Nettoyage et prétraitement des données :**

## **A. Prétraitement des données**

L'analyse de données, en tant que processus itératif, dépend largement de la qualité des données utilisées. Les données brutes, souvent collectées à partir de différentes sources, peuvent contenir des imperfections, des valeurs manquantes, des duplicatas ou d'autres anomalies qui compromettent la fiabilité des résultats d'analyse. C'est là que la phase de nettoyage et prétraitement des données entre en jeu.

## B. Gestion des valeurs manquantes et duplicatas :

Tout d'abord, les dimensions du DataFrame sont (77, 15). Ensuite, les valeurs manquantes dans chaque colonne sont identifiées à l'aide de la fonction `isnull()`, qui renvoie des valeurs booléennes (True, False) et pour obtenir le total des valeurs manquantes, la fonction `isnull().sum().sum()` est utilisée, donnant comme résultat 207. L'utilisation de `isnull().sum()` permet également de présenter le nombre de valeurs manquantes par colonne. Les colonnes avec un nombre significatif de valeurs manquantes sont supprimées à l'aide de la fonction `dropna()`. Premièrement, les colonnes ayant plus de 37 valeurs manquantes sont éliminées, résultant en la suppression des colonnes 'FBCF' et 'Actifs occupés'. Deuxièmement, les lignes ayant plus de 4 valeurs manquantes sont supprimées, entraînant la suppression de toutes les lignes correspondant aux années 2010, 2019 et 2020. Les valeurs manquantes spécifiquement dans les colonnes 'Importations' et 'Exportations' sont remplacées par des zéros à l'aide de la fonction `fillna(0)`. Enfin, une imputation en avant (`ffill()`) est appliquée pour remplir les valeurs manquantes restantes. Ces étapes visent à nettoyer le DataFrame, réduisant ainsi le nombre de valeurs manquantes et préparant les données pour l'analyse ultérieure.

## C. Transformation des données :

### Premièrement :

Nous avons constaté la présence de valeurs numériques qui utilisent la virgule "," au lieu du point ".". Cela peut poser problème en Python, car la virgule est généralement utilisée comme séparateur décimal dans certaines régions, tandis que le point est utilisé dans d'autres. Pour garantir la cohérence, nous devons remplacer les virgules par des points.

### Deuxièmement :

Nous voulons convertir le type de toutes les valeurs des indicateurs en flottant (float). Pour ce faire, nous utilisons la ligne de code suivante :

- `df[column_name] = df[column_name].apply(lambda x: float(x.replace(',', '.')) if isinstance(x, str) else x).`

Cette ligne de code applique une fonction à chaque valeur de la colonne spécifiée (column\_name). La fonction vérifie d'abord si la valeur est une

chaîne de caractères (str). Si c'est le cas, elle remplace la virgule par un point et convertit la valeur en flottant. Si la valeur n'est pas une chaîne de caractères, elle la laisse inchangée.

## **IV. Exploration des données :**

### **A. Définition :**

La visualisation des données joue un rôle central dans le processus d'analyse en améliorant la compréhension des données, en facilitant l'exploration, en communiquant des résultats, en soutenant la prise de décision et en renforçant la collaboration entre les parties prenantes. Elle transforme les données en informations exploitables, en faisant d'elle un outil indispensable pour les analystes de données et les décideurs.

### **B. Préparation des données pour la visualisation :**

D'après l'étape de nettoyage et Prétraitement des Données, nous avons effectué les actions suivantes :

- Suppression des deux indicateurs "actifs occupée" et "FBCF".
- Suppression des années 2010, 2019 et 2020.
- Conversion du type de tous les indicateurs en flottant (float).

### **C. Description de chaque visualisation :**



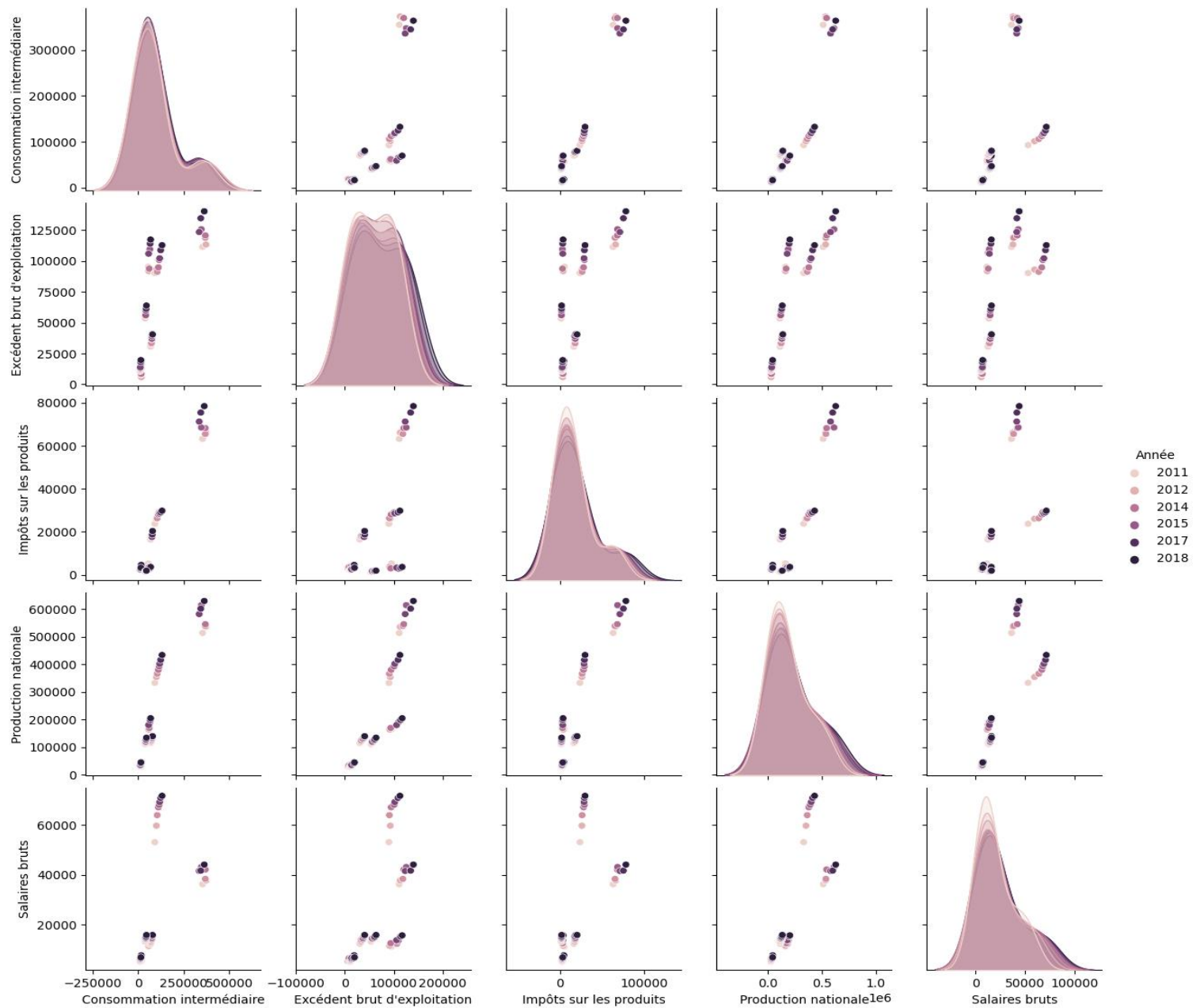


Figure 1: La distribution des différents indicateurs par rapport aux autres.

Dans ce graphe on peut découvrir la distribution des indicateurs par rapport aux autres, ce qui nous permet de choisir le bon modèle pour notre expérience.

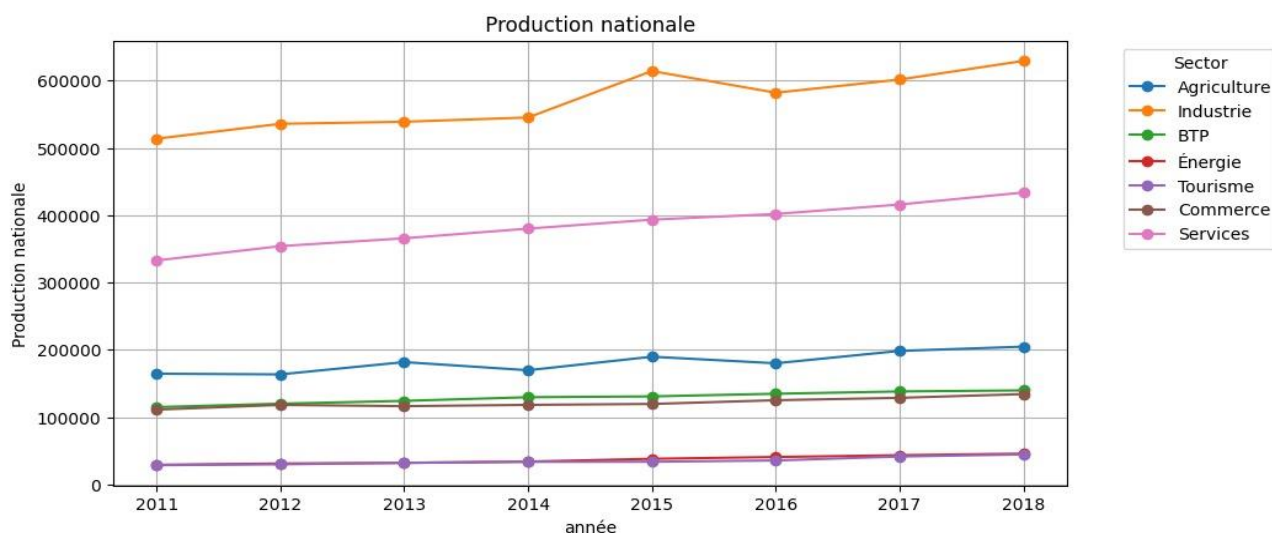


Figure 2: la variation de la production nationale par rapport à chaque section.

Ce graphique reprend la distribution de la production nationale par rapport à chaque modèle.

La production national dans le secteur industrie et services est plus grand aux autre, et les secteur agriculture, BTP et commerce sont une peu moins , et en dernière les deux secteurs services et énergie sont les plus moins (figure 2)

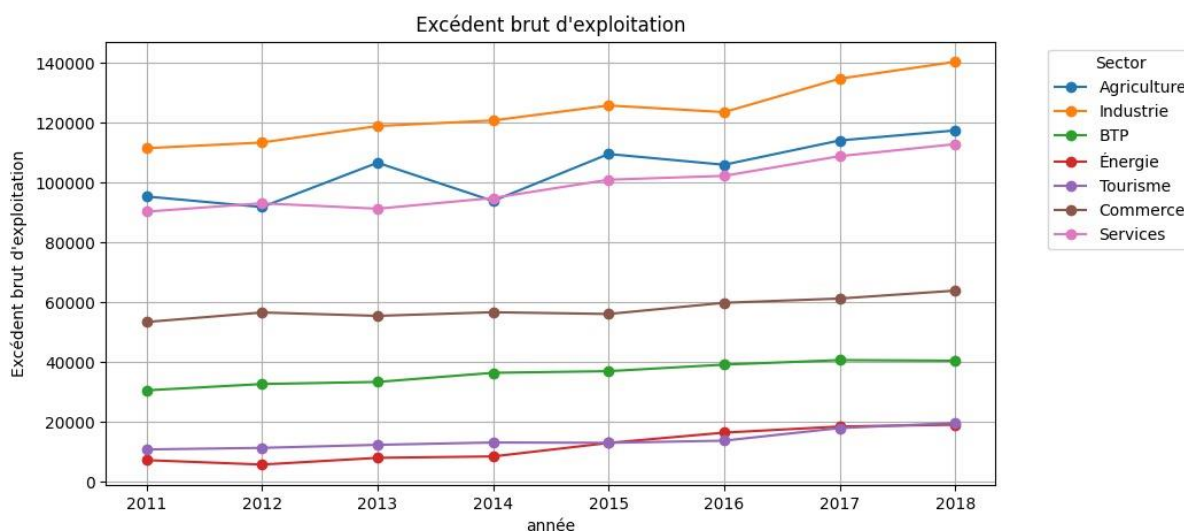


Figure 3: la variation de l'excédent brut d'exploitation par rapport à chaque secteur.

On voit qu'excédent brut d'exploitation dans le secteur industrie est plus grande que les autres, et dans les secteurs tourisme et énergie plus faible que les autres.

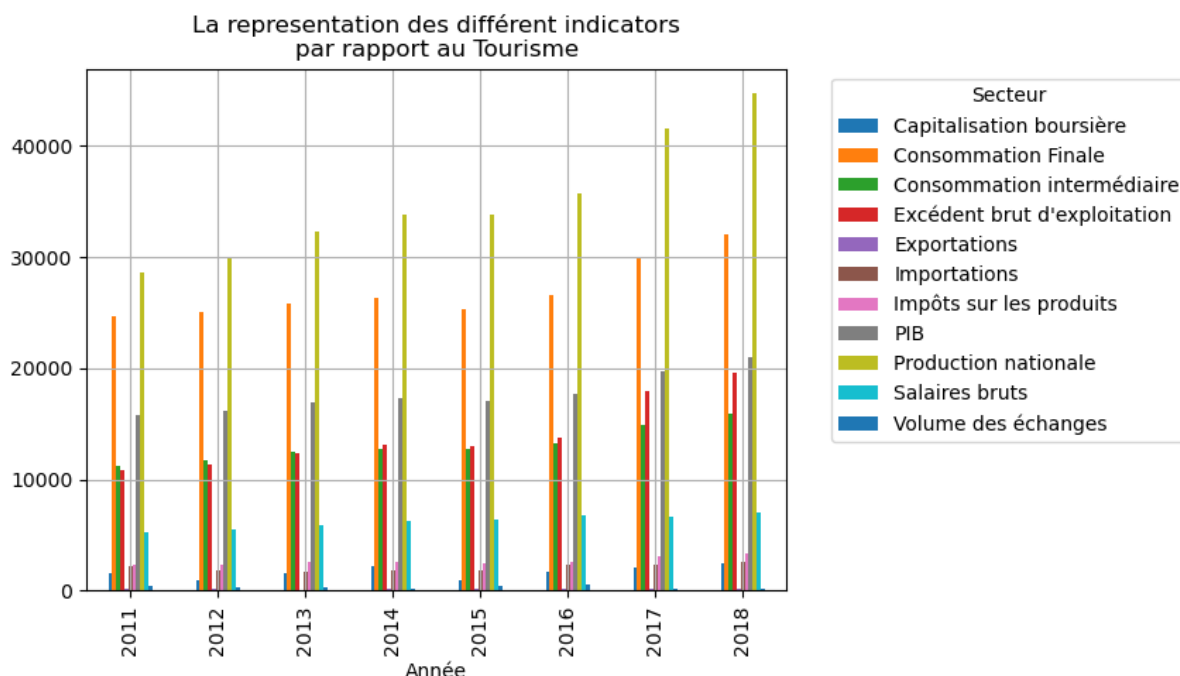


Figure 4: la représentation des différents indicateurs par rapport au secteur du tourisme.

Ce graphique nous donne la répartition des différents indicateurs économiques dans le secteur du tourisme.

Comme il est indiqué sur la figure 4 la production nationale est en progression croissance chaque année, y compris la consommation intermédiaire et l'excédent brut d'exploitation.

Au contraire, l'indice d'impôt sur les produits et de salaire bruts est stable sur toutes les années.

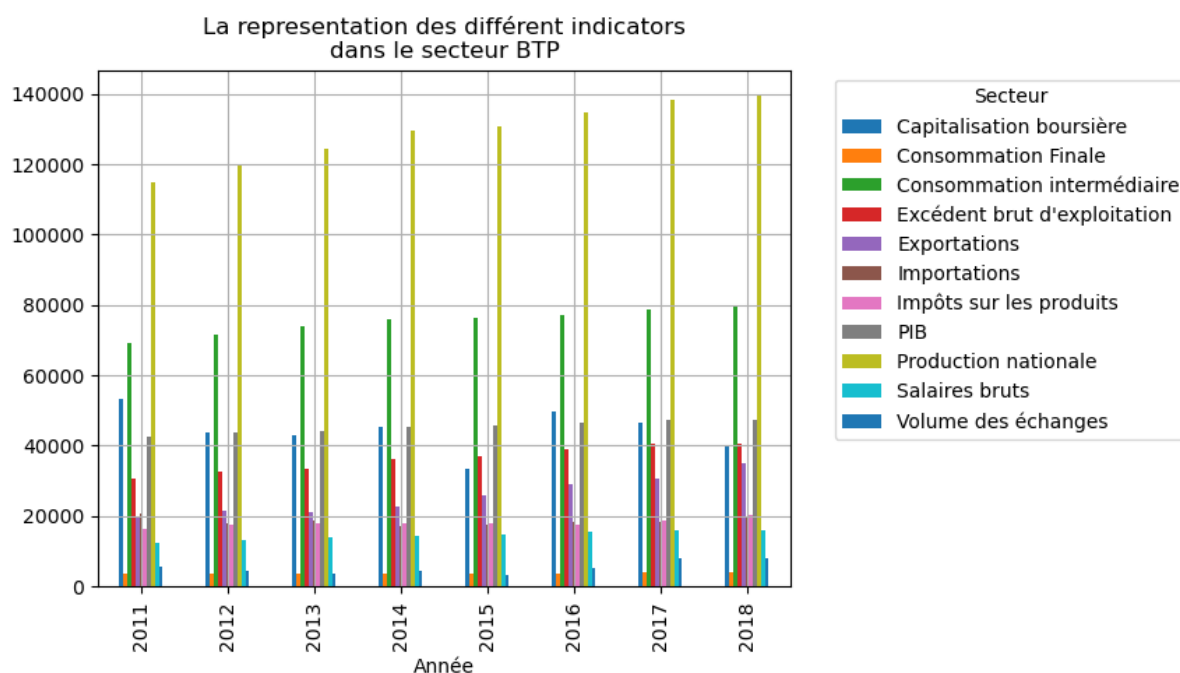


Figure 5: la représentation des différents indicateurs par rapport au secteur BTP.

Ce graphique nous donne la répartition des différents indicateurs économiques dans le secteur du BTP (Bâtiment et Travaux Publics).

De même, comme il est indiqué sur la figure 4 la production nationale est en progression croissance chaque année, y compris la consommation intermédiaire et l'excédent brut d'exploitation.

Au contraire, l'indice d'impôt sur les produits et de salaire bruts est stable sur toutes les années.

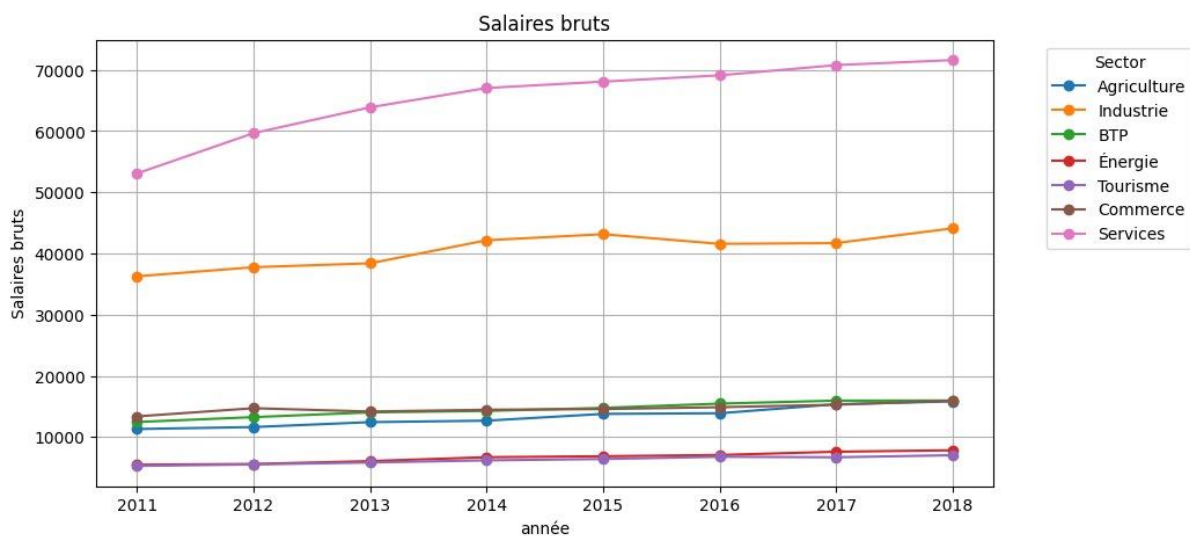


Figure 6: la variation du salaire brut par rapport à chaque secteur.

On voit que le salaire brut est plus élevé dans le secteur services par rapport les autres secteurs, les secteurs énergie et tourisme ils ont des salaires bas, par rapport les autres secteurs.

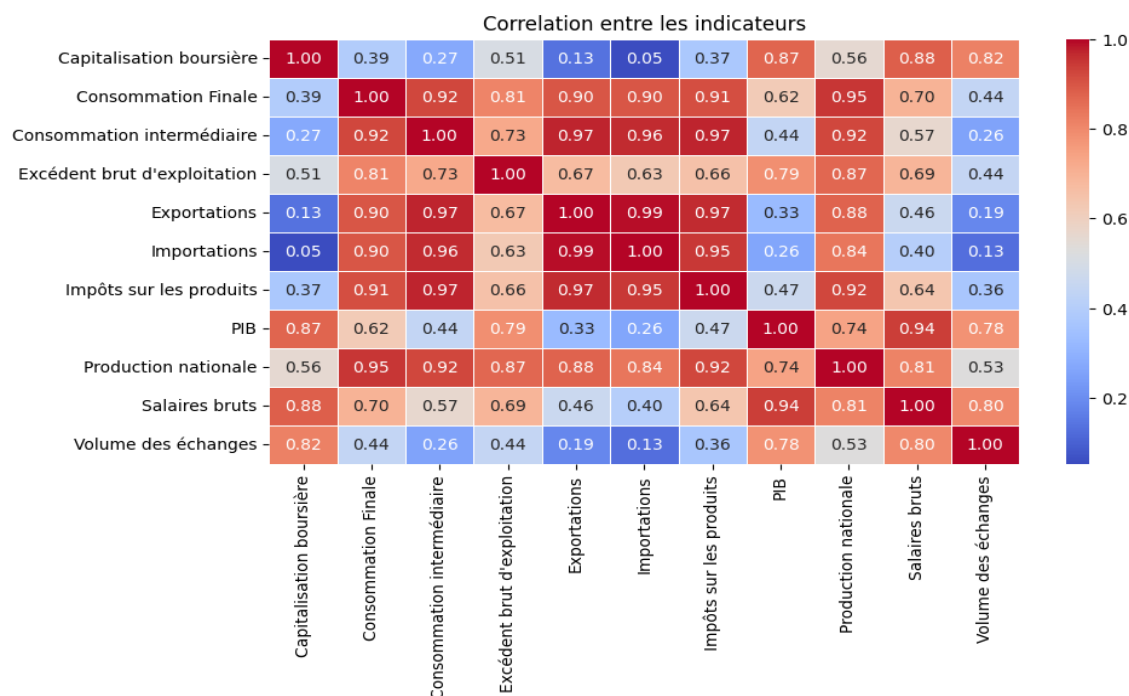


Figure 7 : corrélation entre les indicateurs.

On note que la plupart des coefficients de corrélation entre les indicateurs, varient entre 0,6 et 1 cela indique une forte relation linéaire positive entre les indicateurs.

## D. Hypothèse d'expérience :

D'après ces résultats, on peut constater que le secteur industriel présente des indicateurs plus élevés. Cela suggère que l'industrie pourrait être le secteur le plus rentable au Maroc.

## V. Modélisation des données :

### A. Définition :

La phase de modélisation est une étape cruciale du processus d'analyse de données, où des modèles mathématiques et computationnels sont utilisés pour effectuer des prédictions, classer des données ou découvrir des motifs au sein d'un ensemble de données. Cette phase revêt une importance particulière pour extraire des informations précieuses et des connaissances des données, facilitant la prise de décisions basée sur les données et la résolution de problèmes complexes.

## B. Division des données :

Ce processus est fondamental pour s'assurer que le modèle est entraîné sur un ensemble de données et testé sur un autre, ce qui permet d'évaluer sa capacité à généraliser à de nouvelles données non vues lors de l'entraînement.

- + La fonction **train\_test\_split** est utilisée pour diviser **X** et **y** en ensembles d'entraînement (**X\_train**, **y\_train**) et de test (**X\_test**, **y\_test**).
- + Nous avons choisi un paramètre **test\_size de 0.2** qui indique que 20% des données seront réservées pour l'ensemble de test, tandis que les 80% restants constitueront l'ensemble d'entraînement.
- + Le **random\_state=42** est un paramètre qui assure la reproductibilité de la division. Cela signifie que chaque fois que le code est exécuté, la division des données sera la même.

## C. Sélection ou Ingénierie des Caractéristiques :

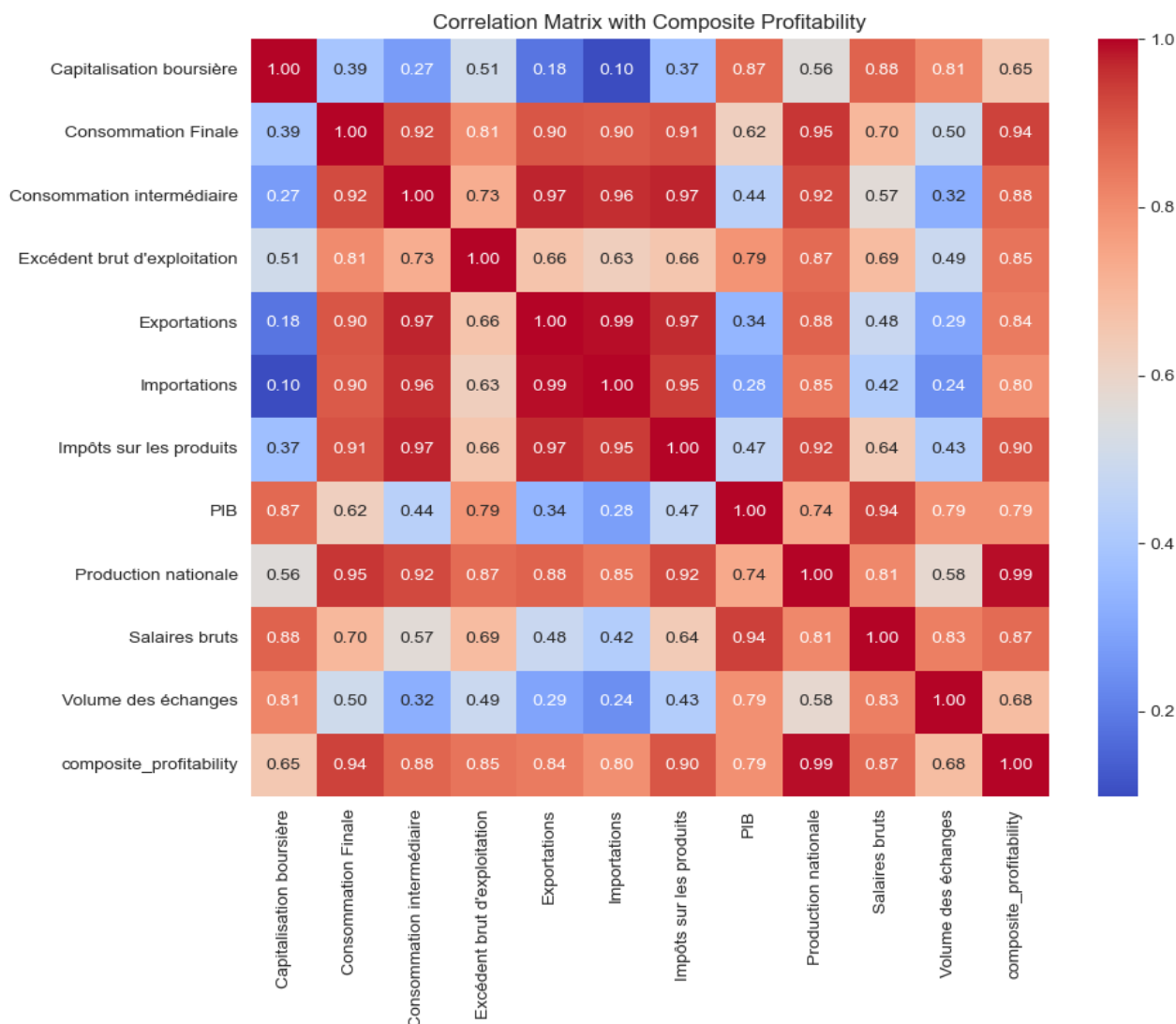
### Normalisation et L'ajout du rentabilité composite :

- + Sélection des indicateurs économiques pertinents en excluant les colonnes non numériques ou non indicatives comme 'Secteur' et 'Année'.
- + Normalisation de ces indicateurs à l'aide de **StandardScaler** pour garantir que toutes les caractéristiques contribuent équitablement au modèle, sans être biaisées par leur échelle.
- + Création d'un nouveau dataframe composé des caractéristiques normalisées.
- + Calcul de la rentabilité composite pour chaque observation en prenant la moyenne des indicateurs normalisés.

Le but de cette étape est de transformer les données de manière à ce que le modèle puisse les interpréter de façon plus efficace, en réduisant les biais potentiels dus aux différences d'échelles entre les indicateurs. Cela

permet d'obtenir une mesure standardisée de la rentabilité qui peut être utilisée comme variable cible pour la prédiction.

### Corrélations de la rentabilité composite avec d'autres variables :



Pour analyser la relation entre la rentabilité composite et les autres variables économiques, nous suivons un processus de calcul de la matrice de corrélation. Voici les étapes clés de ce processus :

- Calcul de la matrice de corrélation pour toutes les variables, y compris la rentabilité composite. Cette matrice nous aide à comprendre comment chaque variable est liée à la rentabilité composite.
- Sélection des corrélations spécifiques entre la rentabilité composite et les autres variables pour identifier les relations les plus fortes.

- + Visualisation de ces corrélations pour faciliter l'interprétation des relations. Cela peut être fait à l'aide de graphiques qui montrent la force et la direction de la corrélation entre la rentabilité composite et les autres variables.
- + Définition d'un seuil de corrélation, par exemple 0.7, pour identifier les caractéristiques qui sont fortement liées à la rentabilité composite.
- + Sélection des caractéristiques qui ont une corrélation supérieure au seuil établi. Ces caractéristiques sont considérées comme ayant une influence significative sur la rentabilité composite et sont donc pertinentes pour le modèle prédictif.
- + Exclusion des caractéristiques qui ont une corrélation inférieure au seuil, car elles sont considérées comme ayant une influence moindre sur la rentabilité composite.

Le but de cette étape est de réduire la dimensionnalité des données en éliminant les caractéristiques qui ont peu ou pas de relation avec la variable cible. Cela permet de simplifier le modèle, de réduire le risque de surajustement (overfitting) et d'améliorer la performance du modèle sur de nouvelles données.

➤ **Vérifier la multicollinéarité :**

Pour vérifier la multicollinéarité parmi les caractéristiques restantes et réduire la complexité du modèle, nous suivons un processus structuré :

- + Calculer la matrice de corrélation pour les caractéristiques restantes afin de détecter la présence de multicollinéarité, qui est une situation où deux ou plusieurs variables sont fortement corrélées linéairement.
- + Définir un seuil de corrélation élevé pour identifier la multicollinéarité, typiquement une valeur au-dessus de 0.7 peut indiquer une forte multicollinéarité.
- + Identifier les paires de caractéristiques où le coefficient de corrélation dépasse le seuil élevé de corrélation.
- + Retirer une caractéristique de chaque paire fortement corrélée pour réduire la multicollinéarité. Cela implique de choisir entre les variables corrélées celles à conserver et celles à éliminer en se basant sur leur importance, leur pertinence ou leur redondance.



- ✚ Supprimer les caractéristiques identifiées du dataset pour finaliser la préparation des données.

Le but de cette étape est d'éliminer la redondance dans les données, ce qui peut améliorer la précision des prédictions du modèle et réduire le risque de surajustement. En éliminant la multicollinéarité, nous nous assurons que notre modèle est plus généralisable et que chaque caractéristique apporte une information unique à la prédiction.

## D. Choix du ou des modèles :

### 1. Régression linéaire:

Est une technique statistique utilisée pour modéliser la relation linéaire entre une variable dépendante, dans notre cas c'est le seuil de rentabilité, et une ou plusieurs variables indépendantes (indicateurs).

### 2. Régresseur de Forêt Aléatoire :

Est un modèle d'apprentissage automatique qui est utilisé pour résoudre des problèmes de régression, où l'objectif est de prédire une variable continue plutôt que de classer des exemples dans des catégories discrètes. Il appartient à la famille des modèles d'ensemble et est construit à partir de plusieurs arbres de décision.

### 3. SVR (Support Vector Regressor) :

Est une technique d'apprentissage automatique utilisée pour résoudre des problèmes de régression. Contrairement à la régression linéaire traditionnelle, qui cherche à modéliser la relation linéaire entre les variables.

## E. Entraînement du modèle :

Pour entraîner un modèle de machine learning et évaluer sa performance, nous suivons un processus structuré :

### ➤ Entraînement du modèle (.fit) :

- ✚ Utiliser la méthode **.fit** pour entraîner le modèle sur l'ensemble d'apprentissage.

- ✚ Le modèle apprend à faire des prédictions en ajustant ses paramètres internes pour minimiser l'erreur de prédiction.
- **Prédiction sur l'ensemble de test :**
  - ✚ Utiliser la méthode **.predict** pour générer des prédictions sur l'ensemble de test.
  - ✚ Cela permet d'évaluer comment le modèle performe sur des données qu'il n'a pas vues pendant l'entraînement.

## VI. Evaluation :

### A. Définition :

La phase d'évaluation revêt une importance cruciale dans le processus d'analyse des données. C'est à ce stade que nous évaluons les performances des modèles que nous avons développés, afin de déterminer dans quelle mesure ils répondent aux objectifs fixés. L'évaluation permet de prendre des décisions éclairées sur la pertinence et l'efficacité des solutions proposées.

### B. Critères d'évaluation :

- **Calcul des métriques (r2\_score) :**
  - ✚ Utiliser **r2\_score** pour calculer le coefficient de détermination, qui mesure la qualité de la prédiction du modèle.
  - ✚ Un score  $R^2$  proche de 1 indique que le modèle explique une grande partie de la variance de la variable cible.

Le but de ce processus est de construire un modèle prédictif fiable et de comprendre son fonctionnement. En évaluant la performance du modèle et l'importance des caractéristiques, nous pouvons améliorer le modèle et prendre des décisions éclairées sur les caractéristiques à utiliser pour les prédictions.

### C. Analyse des erreurs :

Table 1: Les model utilise et leur score.

Model	score
-------	-------

LinearRegression_model	99.3 %
RandomForest_model	99.8 %
SupportVector_model	98.6 %

## D. Interprétation des résultats :

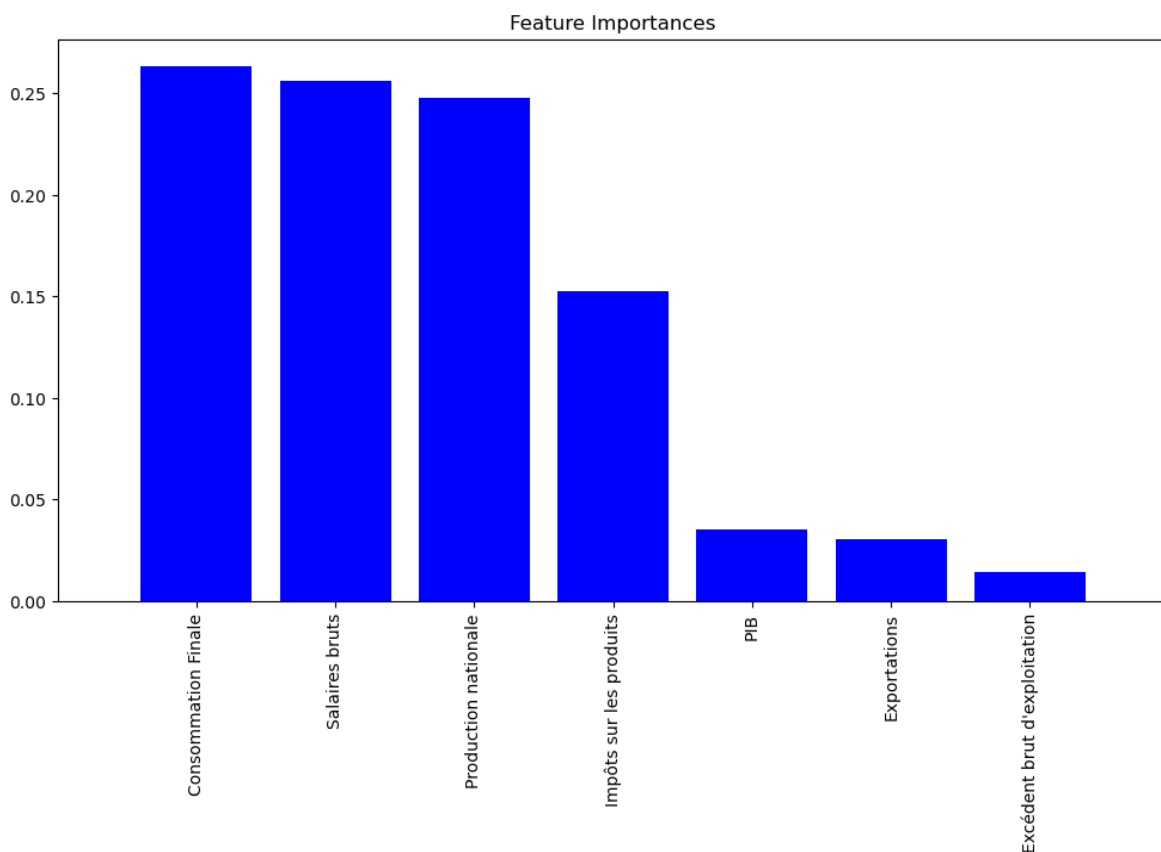


Figure 8: Les indicators importants..

### ➤ Investigation de l'importance des caractéristiques (.feature\_importances\_) :

- ✚ Utiliser l'attribut **.feature\_importances\_** pour déterminer l'importance de chaque caractéristique dans les prédictions du modèle.
- ✚ Cela aide à comprendre quelles caractéristiques contribuent le plus à la prédiction et peuvent être utiles pour l'interprétation du modèle et la sélection des caractéristiques.

## E. Limitations et recommandations :

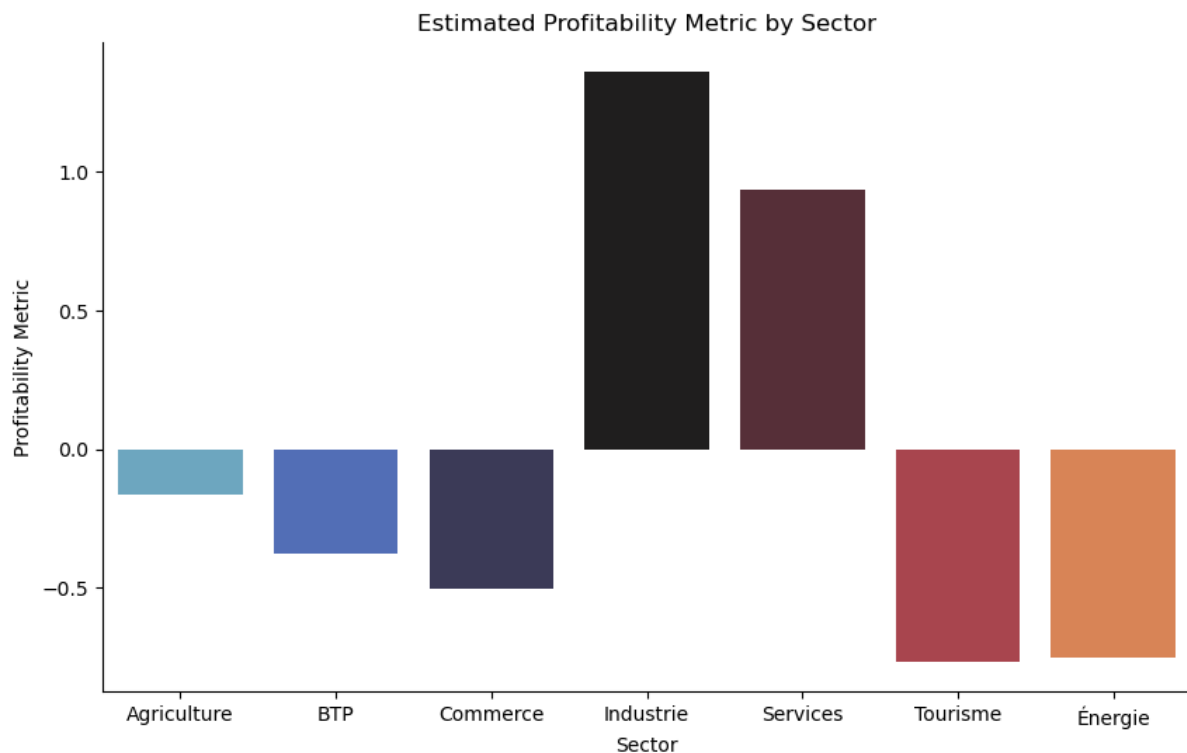


Figure 9: Estimation de la mesure de la compétence par secteur.

### 1. Recommendations:

- ✚ Secteur industriel : Compte tenu de la rentabilité du secteur industriel, il est recommandé d'investir davantage dans les progrès technologiques, l'innovation et les infrastructures pour améliorer la productivité et la compétitivité.
- ✚ Secteur du tourisme : Malgré sa faible rentabilité actuelle, le secteur du tourisme présente un potentiel de croissance important. Les recommandations peuvent inclure l'amélioration des infrastructures, la promotion du tourisme culturel et écologique et l'amélioration de l'expérience touristique globale.
- ✚ Secteur de l'énergie : Pour améliorer la rentabilité du secteur de l'énergie, il est recommandé de se concentrer sur les sources d'énergie renouvelables, l'efficacité énergétique et les initiatives de développement durable.

## 2. Limites :

- ✚ Secteur industriel : Bien que le secteur industriel soit actuellement le plus rentable, il peut être confronté à des défis liés à la durabilité environnementale, à la gestion des ressources et aux fluctuations du marché mondial.
- ✚ Secteur de l'énergie : les limites du secteur de l'énergie peuvent inclure la dépendance à l'égard de ressources non renouvelables, les contraintes réglementaires et la nécessité d'investissements initiaux substantiels.

## Conclusion :

En conclusion, le secteur industriel présente le potentiel de rentabilité le plus élevé au Maroc, tandis que les secteurs du tourisme et de l'énergie nécessitent des interventions stratégiques pour libérer leur plein potentiel économique. En répondant aux recommandations et en surmontant les limites, le Maroc peut parvenir à un paysage économique plus équilibré et plus durable.

## Les références :

[1] : <https://pandas.pydata.org/>.

[2] : <https://numpy.org/>.

[3] : <https://matplotlib.org/>.

[4] : <https://scikit-learn.org/stable/>

[5] : <https://seaborn.pydata.org/>

[6] : <https://manar.finances.gov.ma/manar/initAccueilInscription> .