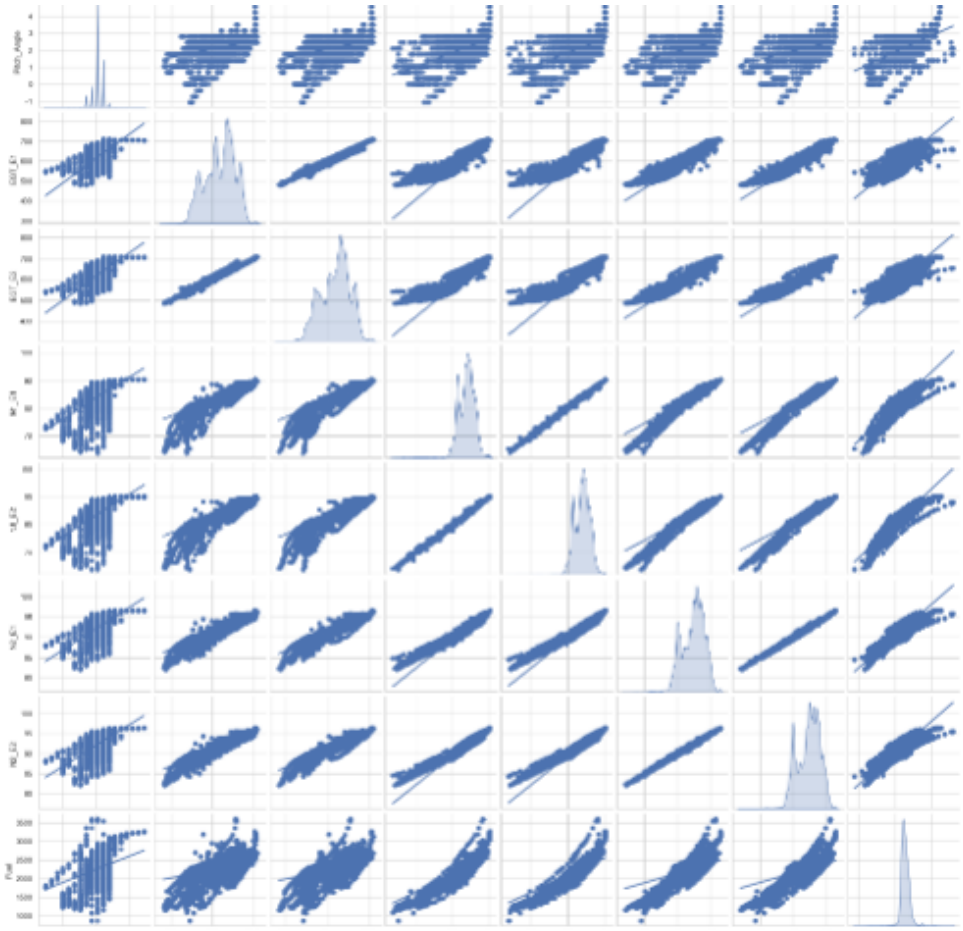
skyconseil

FÉVRIER2024

Analyse Exploratoire des Données(EDA)



INTRODUCTION

MÉTHODES UTILISÉES

ANALYSE DES RÉSULTATS

CONCLUSION

RÉALISÉ PAR : PAULSABIA

Introduction

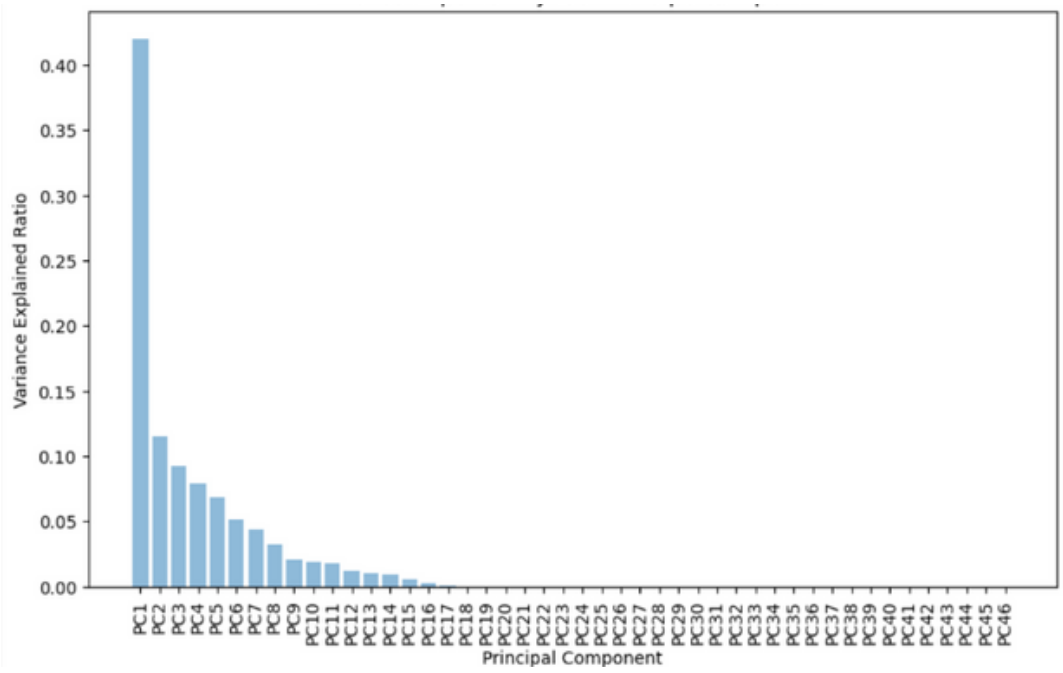
Notre EDA vise à éclairer le modèle prédictif de la consommation de carburant des avions. En examinant les données, nous identifions les variables clés et la distribution de la consommation, jetant ainsi les bases d'une prédiction précise.

Analyse en Composantes Principales

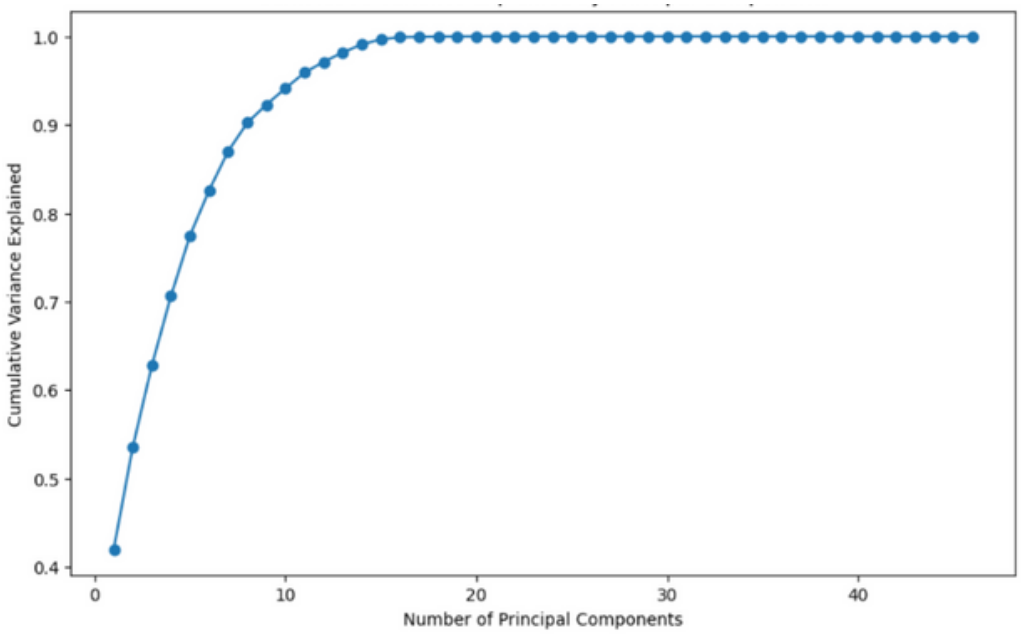
L'ACP est une technique statistique de réduction dimensionnelle qui permet de simplifier la complexité des espaces de données en les transformant en structures plus petites et plus interprétables.
L'ACP répond à la question suivante : comment réduire un grand nombre de variables tout en conservant l'essentiel de l'information ?

Remarque:

Avant l'ACP, nous avons retiré les colonnes non essentielles, telles que ["**Time_secs**", "**Vitesse_Verticale**", "**Fuel_E1**", "**Fuel_E2**"], pour leur redondance ou format inadéquat. La colonne "**Fuel**", destinée à être notre variable cible, a également été exclue de l'analyse.



1.1 Variance expliquée par chaque composante principale

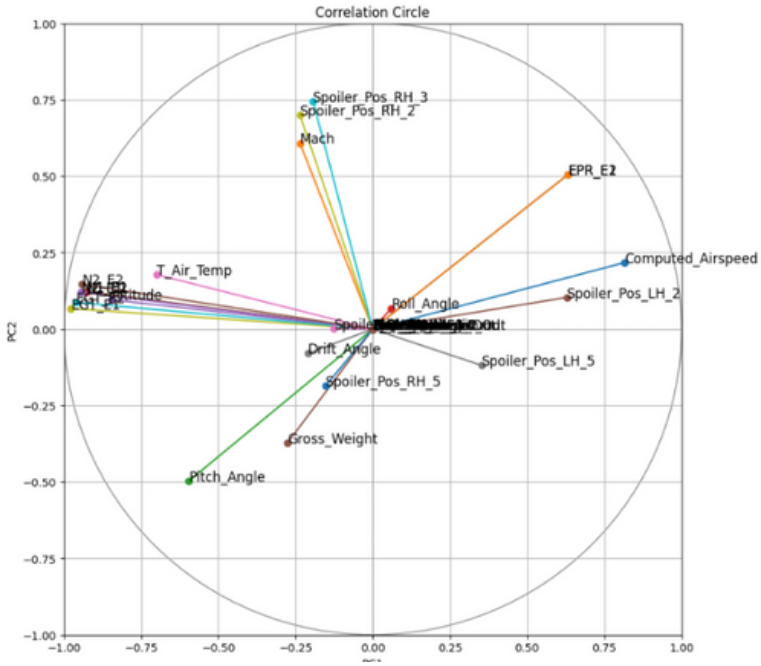


1.2 Variance cumulative expliquée par les composantes principales

Il ressort de notre analyse que l'utilisation des 17 premières variables suffirait à conserver 99% de l'information contenue dans les données.
Par ailleurs, il est notable que plus de la moitié de la variance est capturée par seulement deux composantes principales. Cela nous amène à envisager l'utilisation d'un cercle de corrélation pour examiner les liens entre les composantes principales et les différentes variables. Cette démarche nous permettra d'identifier les variables qui présentent une forte corrélation entre elles, offrant ainsi la possibilité d'éliminer les doublons en retirant les variables excessivement liées.

Cercle de corrélation

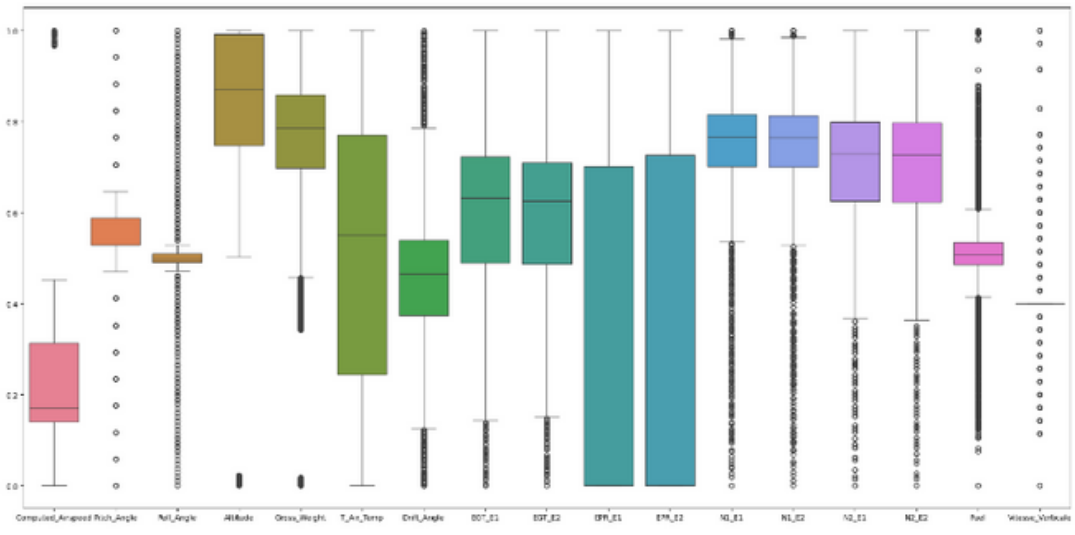
Le cercle de corrélation est un outil visuel associé à l'ACP, qui montre comment les variables d'origine se rapportent aux composantes principales et entre elles. Il clarifie les associations et les contributions des variables aux axes de l'ACP, facilitant l'interprétation des dimensions réduites.



Il ressort de notre analyse que les variables **N2_E1**, **N2_E2**, **EGT_E1**, **EGT_E2**, **l'altitude** et la **température de l'air** présentent une corrélation négative marquée avec la deuxième composante principale. Nous avons également **Spoiler_Pos_RH_3**, **Spoiler_Pos_RH_2,** et **mach** qui présentent une forte corrélation avec la première composante principale.
Il sera donc envisageable de fusionner ces variables, ou de les prioriser pour le développement du futur modèle.

Diagrammes en Boîte

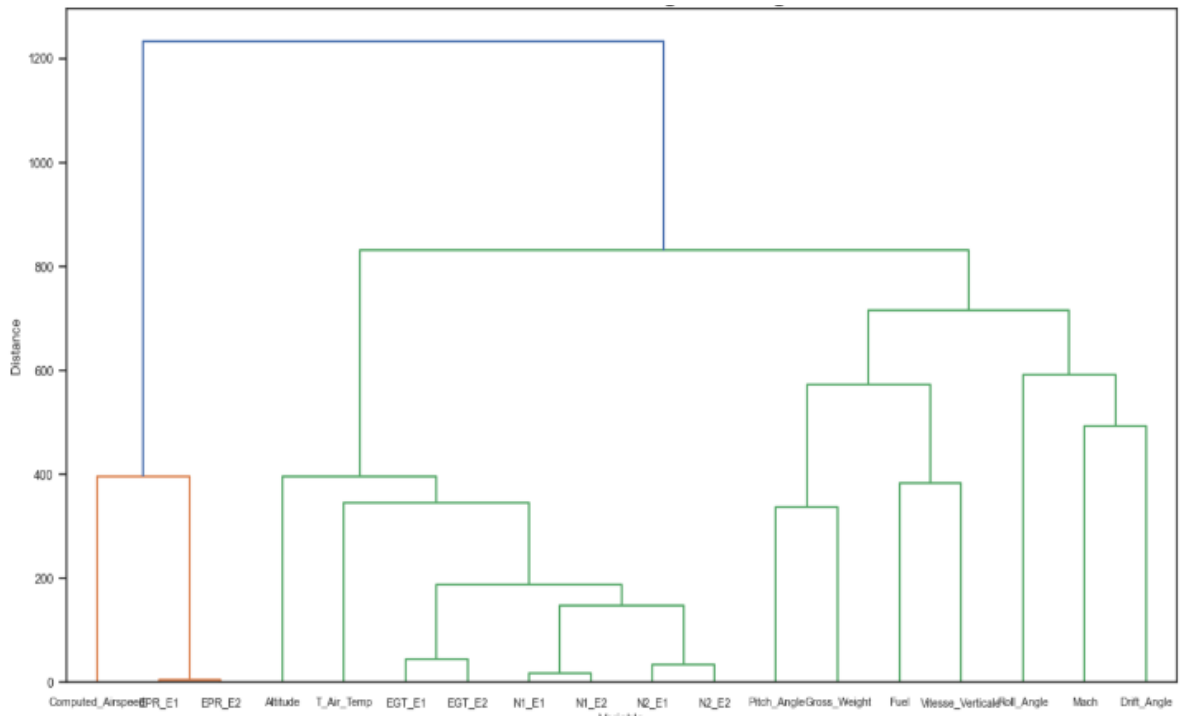
Les box-plots sont des graphiques statistiques qui résument la distribution d'un ensemble de données à travers sa médiane, ses quartiles et ses extrêmes, permettant de détecter rapidement les valeurs aberrantes et de comparer différentes distributions.



les données ont été normalisées, facilitant l'identification des variables fortement corrélées par l'observation de la hauteur et de la taille de la boîte. Ainsi, **EGT_1** et **EGT_2** apparaissent comme étroitement liés. Pour décrypter le diagramme en boîte lui-même, les bords de la boîte délimitent les premier et troisième quartiles (inférieur et supérieur, respectivement), encapsulant ainsi **50 %** des observations. La médiane est représentée par la ligne centrale, tandis que les "moustaches" s'étendant au-delà de la boîte indiquent la présence de valeurs extrêmes, signalées par des points isolés.

Dendrogramme de classification hiérarchique

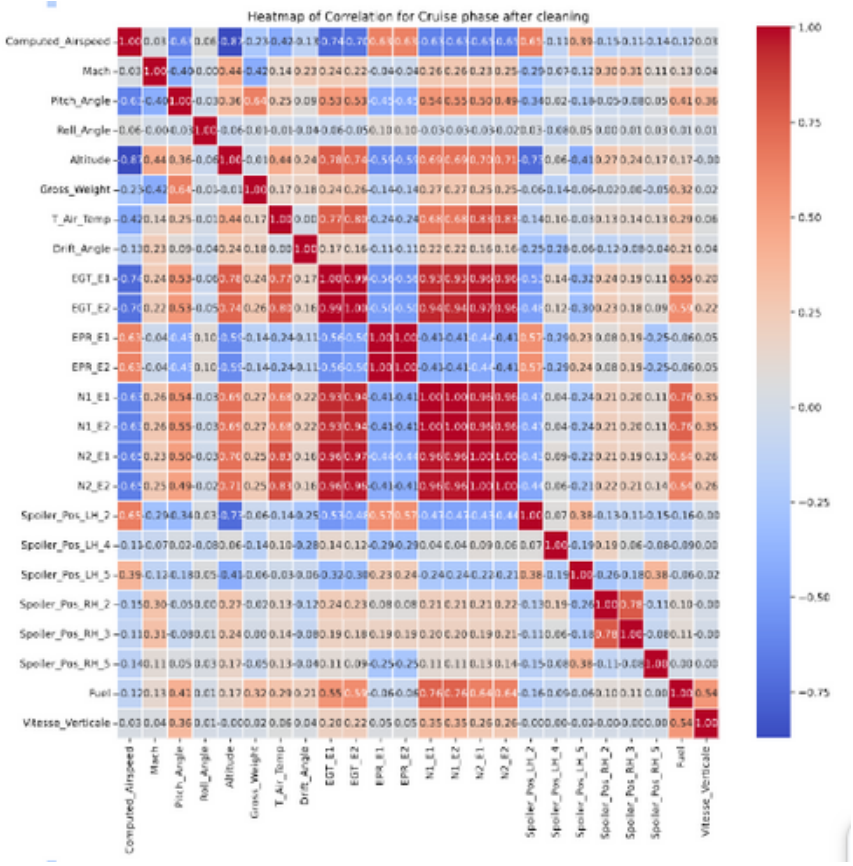
Le dendrogramme de classification hiérarchique est un diagramme en arbre qui montre les regroupements de données selon leur similarité. Il illustre comment chaque groupe est formé en fusionnant étape par étape les éléments ou clusters les plus proches, la hauteur des jonctions reflétant leur distance ou dissimilarité.



par exemple ici , il est observable que le pitch angle et la gross weight sont en corrélation avec le carburant et la vitesse verticale.

Heatmap de corrélation

La heatmap de corrélation est un tableau coloré qui montre le degré de corrélation entre les variables d'un jeu de données, où les couleurs indiquent l'intensité de la corrélation, facilitant l'identification rapide des relations entre variables.



Nous remarquons la présence d'une zone distinctement **rouge** près du centre de la heatmap, indiquant une corrélation très élevée entre certaines variables. Pour feature engineering, il pourrait être judicieux de combiner ces variables ou d'en sélectionner une sur deux. Cette observation s'applique également aux variables **EGT_E1** et **EGT_E2**. Il est également important de prendre en compte d'autres corrélations significatives, notamment entre :

- - Altitude et Computed_Air_Speed
- - Pitch angle et Computed_Air_Speed
- - Spoiler_Pos_LH_2 et Computed_Air_Speed
- - Gross_Weight et Pitch_Angle

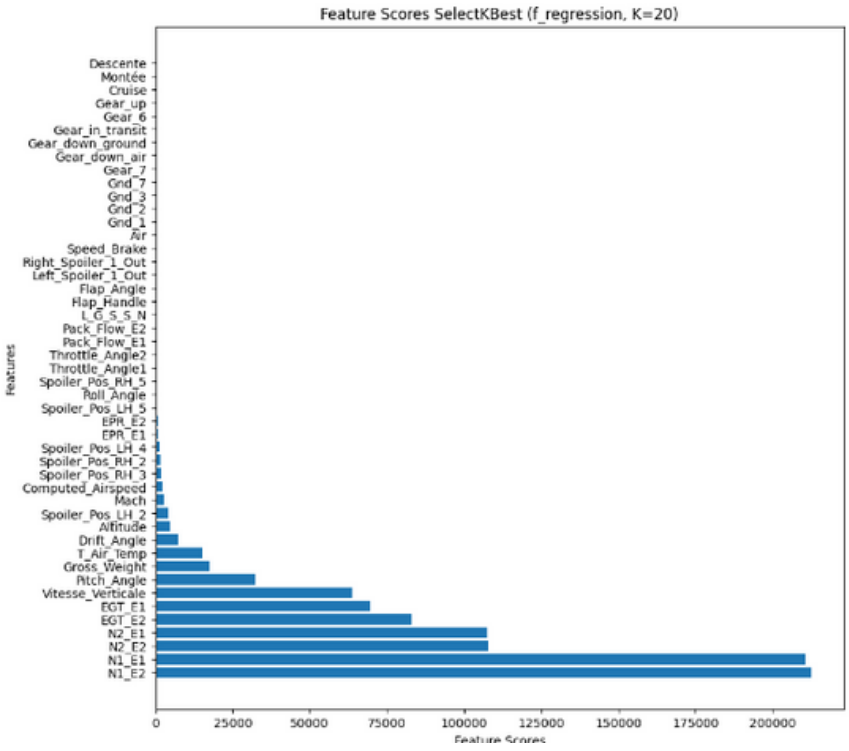
Il est à noter que la corrélation entre **Computed_Air_Speed** et les valeurs **EPR** est positive et assez similaire avec d'autres paramètres moteurs désignés par E1, E2, à l'exception d'une corrélation négative notable.

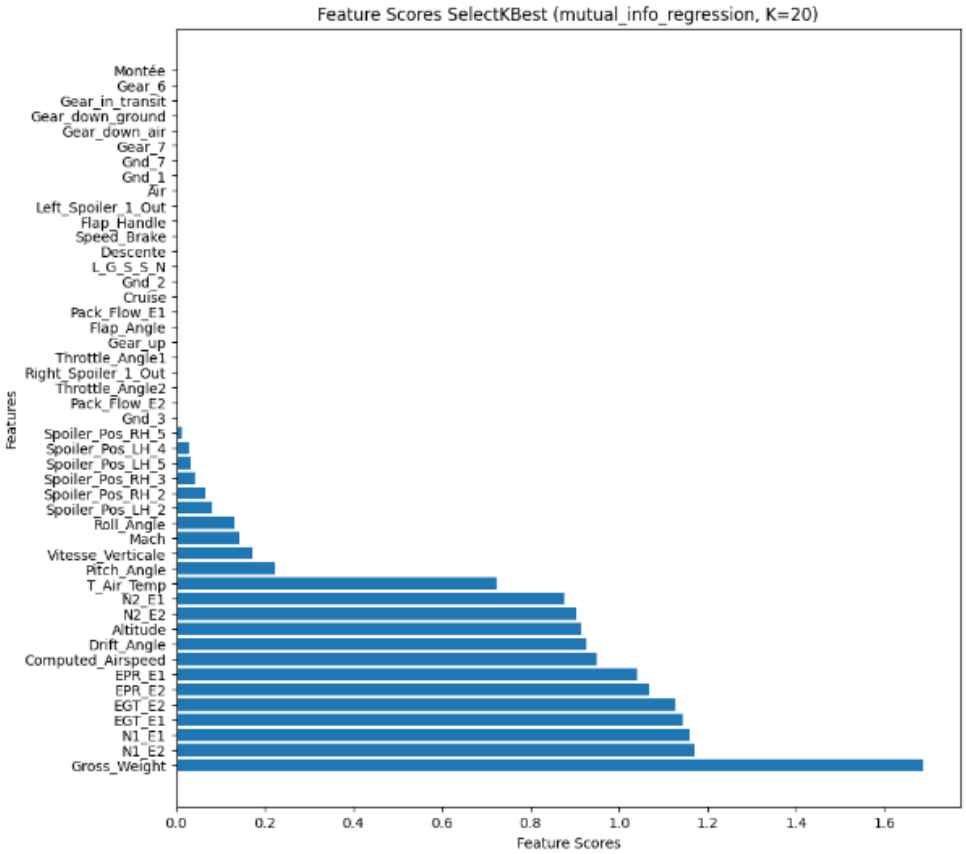
SelectKBest

SelectKBest prend deux paramètres en entrée : une fonction de score et le nombre de paramètres à sélectionner. La fonction de score est utilisée pour évaluer l'importance de chaque paramètre dans la prédiction. Nous allons explorer deux fonctions de score : f_regression, qui mesure la corrélation linéaire entre chaque paramètre et la variable cible (le carburant), et mutual_info_regression, qui mesure l'information mutuelle entre chaque paramètre et la variable cible, prenant en compte les dépendances non linéaires. Nous testerons ces fonctions avec 20 paramètres.

Les résultats initiaux montrent une sélection similaire de paramètres par les deux fonctions de score. Parmi ces paramètres communs, on trouve : 'Computed_Airspeed', 'Mach', 'Pitch_Angle', 'Altitude', 'Gross_Weight', 'T_Air_Temp', 'Drift_Angle', 'EGT_E1', 'EGT_E2', 'EPR_E1', 'EPR_E2', 'N1_E1', 'N1_E2', 'N2_E1', 'N2_E2', 'Spoiler_Pos_LH_2', 'Spoiler_Pos_RH_2', 'Spoiler_Pos_RH_3', 'Vitesse_Verticale'.

Cependant, f_regression ajoute 'Spoiler_Pos_LH_4' à la sélection, tandis que mutual_info_regression inclut 'Roll_Angle'. Ce dernier pourrait être pertinent tandis que les scores attribués aux paramètres 'Spoiler_Pos' sont relativement bas comparés aux autres. Ceci suggère que malgré leur inclusion, ces paramètres peuvent avoir une influence moindre sur la consommation de carburant, comme observé dans les graphiques qui suivent.

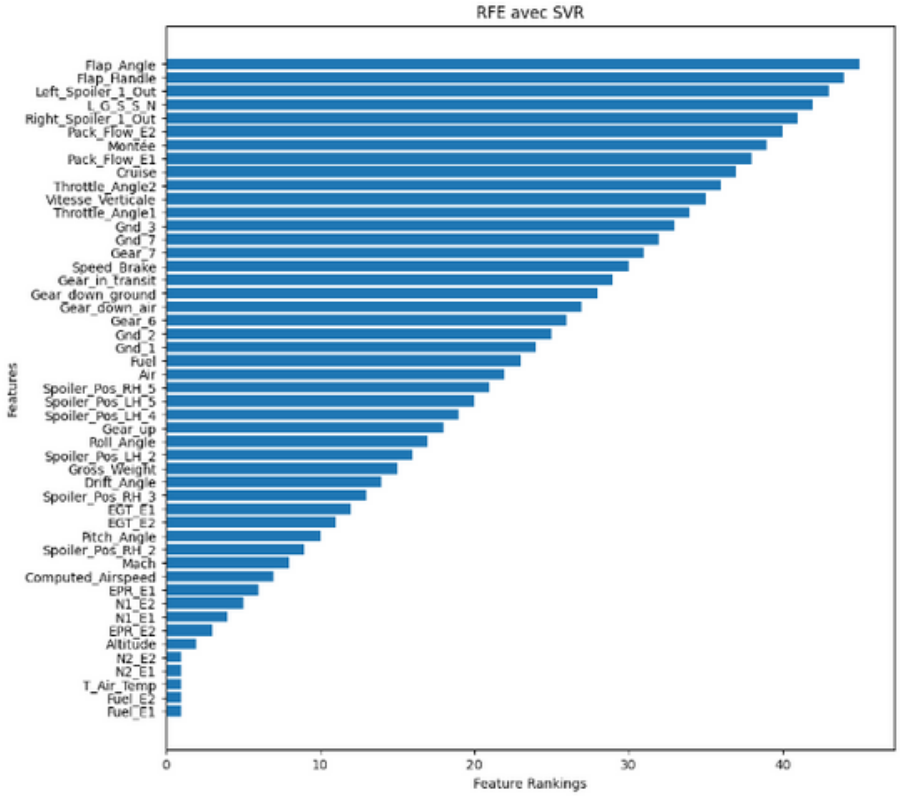




Nous avons également testé SelectPercentile, une méthode similaire à SelectKBest, qui vient confirmer les résultats précédents en sélectionnant les paramètres suivants : 'Mach', 'Pitch_Angle', 'Altitude', 'Gross_Weight', 'T_Air_Temp', 'Drift_Angle', 'EGT_E1', 'EGT_E2', 'N1_E1', 'N1_E2', 'N2_E1', 'N2_E2', 'Spoiler_Pos_LH_2', 'Vitesse_Verticale'. Cette convergence entre les méthodes renforce la fiabilité des paramètres identifiés comme les plus influents pour prédire la consommation de carburant.

Recursive Feature Elimination

La méthode RFE (Recursive Feature Elimination) est une technique de sélection de variables qui élimine successivement les caractéristiques les moins importantes pour identifier celles qui contribuent le plus à la précision du modèle.



Hypothèses sur les selected features

Parmi les variables entrantes, selon les résultats obtenus, nous pouvons émettre que N1_E1, N1_E2, N2_E1, N2_E2, EGT_E1, EGT_E2, Gross Weight, Computed Airspeed, Drift Angle, Pitch Angle, Altitude, Total Air Temperature, Mach, Spoiler_Pos_LH_2 semblent expliquer la variable Fuel.

Les coefficients de corrélation

Les coefficients de corrélation vont de **-1** à **+1**, montrant comment deux variables se rapportent linéairement.

- +1 indique une corrélation positive parfaite
- -1 une corrélation négative parfaite
- 0 aucune corrélation

Spoiler_Pos_RH_5	0.001083	Pitch_Angle	0.412037
Roll_Angle	0.005572	Vitesse_Verticale	0.535533
Spoiler_Pos_LH_5	0.059354	EGT_E1	0.551990
EPR_E2	0.061794	EGT_E2	0.586167
EPR_E1	0.062119	N2_E1	0.635720
Time_secs	0.070069	N2_E2	0.636029
Spoiler_Pos_LH_4	0.088091	N1_E1	0.755097
Spoiler_Pos_RH_2	0.098166	N1_E2	0.756596
Spoiler_Pos_RH_3	0.105359	Fuel_E2	0.996221
Computed_Airspeed	0.117001	Fuel_E1	0.996295
Mach	0.132573	Fuel	1.000000
Spoiler_Pos_LH_2	0.162213	Throttle_Angle1	NaN
Altitude	0.168981	Throttle_Angle2	NaN
Drift_Angle	0.210128	Pack_Flow_E1	NaN
T_Air_Temp	0.294205	Pack_Flow_E2	NaN
Gross_Weight	0.315133	L_G_S_S_N	NaN

Les coefficients de corrélation avec le Fuel indiquent :

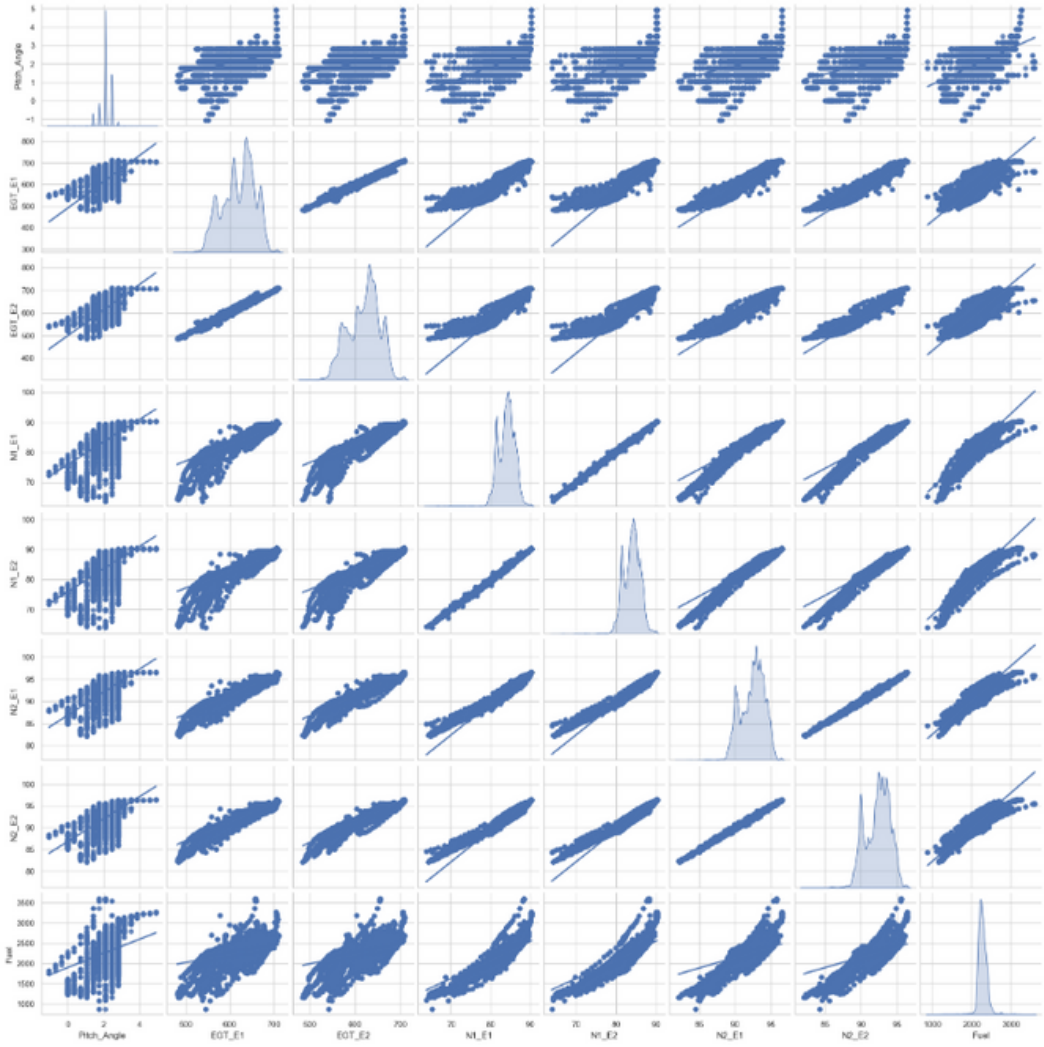
- **Pitch_Angle** (0.412): Corrélation ***Modérée***.
- **EGT_E1** (0.552) et **EGT_E2** (0.586) : Corrélation ***Modérée***.
- **N2_E1** et **N2_E2** (0.636) : ***Forte*** corrélation.
- **N1_E1** (0.755) et **N1_E2** (0.757) : ***Très*** forte corrélation.

Ces résultats soulignent l'influence des performances moteur sur la consommation

REMARQUE:
Le coefficient de corrélation est **NaN** pour des variables avec des valeurs constantes car la corrélation nécessite de la variance dans les données pour être calculée. Sans variance, comme c'est le cas avec des variables ne prenant que la valeur **0** ou **1**, le calcul de la corrélation n'est pas possible, résultant en une valeur NaN.

Pairplot

Un pairplot nous permet de voir à la fois les distributions univariées (**sur la diagonale**) et les relations bivariées (**hors diagonale**) entre les caractéristiques.



A. Les distributions univariées (sur la diagonale):

- **Pitch_Angle** : Présente une asymétrie gauche, avec des valeurs fréquemment inférieures à la moyenne de **2.13**.
- **EGT_E1/ EGT_E2** : Montre une légère asymétrie gauche avec une moyenne de **619.71**, étendant de 480 à 711 avec une tendance vers des valeurs légèrement inférieures à la moyenne.
- **N1_E1** et **N1_E2** : Montre une forte asymétrie gauche, avec des valeurs souvent inférieures à la moyenne de **83.8**
- **N2_E1** etr **N2_E2** : Présente une asymétrie gauche modérée, avec une concentration des valeurs proches de la moyenne de **92.38** mais une tendance vers des valeurs inférieures.

B. Les distributions bivariées (hors diagonale):

1. Pitch Angle et son influence sur les variables mesurées :

L'analyse montre une corrélation non linéaire entre le **Pitch Angle** et les variables mesurées, avec une augmentation suivie d'une stabilisation des valeurs, suggérant un seuil de saturation

2. EGT_E1 :

- **EGT_E1** et **Fuel:** relation non linéaire, avec des régions indiquant une efficacité variable du carburant à différents niveaux de température des gaz d'échappement.
- **EGT_E1** et **N2_E2** : Corrélation positive régulière
- **EGT_E1** et **N2_E1** : une relation positive linéaire ou presque linéaire
- **EGT_E1** et (**N1_E1/ N1_E2**) : Les données semblent suivre une tendance linéaire avec une certaine dispersion
- **EGT_E1** et **EGT_E2** : une relation fortement positive et linéaire, comme indiqué par la concentration des points le long d'une ligne ascendante.

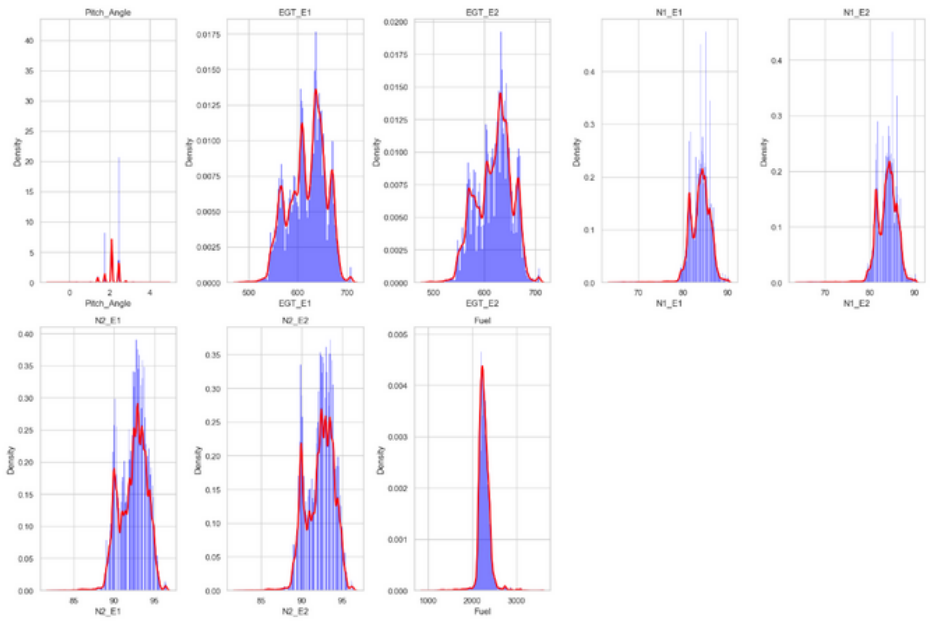
3. Fuel :

Les graphiques indiquent que la consommation de carburant augmente avec les vitesses de moteur N1 et N2. La relation est plus linéaire pour N2, tandis que pour N1, elle montre une réponse non linéaire avec un aplatissement à des vitesses plus élevées.

Tests Statistiques

1.Tests de Normalité:

A. Histogramme et courbe de densité:



B. D. Kurtosis (Mesure l'épaisseur des queues de la distribution):

Le kurtosis mesure à la fois l'aplatissement et la probabilité d'apparition des valeurs aberrantes (extrêmement élevées ou basses par rapport à la moyenne).

Une distribution normale a un kurtosis de 3.
Les distributions peuvent être classées en trois catégories en fonction de leur niveau d'aplatissement :

- **Mésokurtiques** : Aplatissement moyen, valeurs aberrantes ni très fréquentes ni très rares.
- **Platykurtiques** : Faible aplatissement, valeurs aberrantes très peu fréquentes.
- **Leptokurtiques** : Fort aplatissement, valeurs aberrantes fréquentes.

Interprétation :

Un kurtosis supérieur à 3 indique une distribution plus pointue que la normale (leptokurtique).
Un kurtosis inférieur à 3 indique une distribution plus étalée que la normale (platykurtique).

Pitch_Angle	5.169984
EGT_E1	-0.644391
EGT_E2	-0.505297
N1_E1	5.888361
N1_E2	5.908715
N2_E1	0.337595
N2_E2	0.207315
Fuel	15.594958

REMARQUE:
L'utilisation du test *ANOVA* a été écartée pour notre analyse de prédiction de la consommation de carburant, principalement en raison de la non-conformité de nos données avec l'hypothèse de normalité requise pour l'ANOVA. De plus, notre ensemble de données manque de variables catégorielles distinctes nécessaires pour grouper les données, ce qui est essentiel pour l'application de l'ANOVA. Notre objectif étant de développer un modèle prédictif, nous avons privilégié des approches plus adaptées telles que la régression et les méthodes de machine learning pour explorer les relations entre les variables continues.

L'analyse a montré que nos données ne sont pas normalement distribuées, ce qui nous amène à :

- **Choix des Tests Statistiques** : La non-normalité nous guide vers l'adoption de tests non paramétriques adaptés, qui ne présupposent pas une distribution normale des données.

- **Transformation des Données** : La nécessité de potentiellement transformer les données pour atténuer les écarts par rapport à la normalité est mise en lumière, afin de faciliter certaines analyses statistiques.

- **Sélection de Modèles** : Cette observation influe sur notre sélection de modèles de prédiction, nous orientant vers des approches robustes à la non-normalité des données, comme certains algorithmes de machine learning.

- **Identification des Valeurs Aberrantes et des Anomalies** : Une distribution fortement non normale peut souvent indiquer la présence de valeurs aberrantes ou d'anomalies dans les données.

Cette approche assure une analyse plus robuste et fiable.