Rapport du TP de machine learning : Groupe 3

Données extraites de
https://www.kaggle.com/datasets/jtrofe/beer-recipes?select=recipeData.csv

Introduction

Le but de ce travail est de mettre en œuvre les connaissances acquises lors de nos cours d'Introduction d'IA et de machine Learning avec Sylvain Gault.

Le sujet du projet est la création d'une IA pour une société fictive (BREW) afin d'accélérer le développement de nouvelles recettes de bières. Le but principal de cette IA est de prédire l'amertume et la teneur en alcool du produit à partir des autres éléments de cette qui seront fournis par un utilisateur final.

Le projet a pour but d'être utilisé par des utilisateurs finaux qui n'ont aucune connaissance en informatique, il nous faudra donc réaliser en plus une interface graphique ergonomique afin que ces utilisateurs puissent utiliser le projet.

Données

Pour cela, nous nous sommes basés sur un large dataset de recettes de bière provenant de Kaggle. Après exploration, on s'est vite rendu compte que ces données ne sont pas parfaites, en plus d'avoir des données inutiles pour nos travaux, certaines de ces données importantes contiennent des aberrations.
Un travail de prétraitement à dû être mis en place, à commencer par catégoriser les types de données et définir leur signification.

Non de la donnée	Description	Utilisation de la donnée	Raisonnement
BeerID	Identifiant de la bière	Non utilisé	L'identifiant de la bière n'a pas d'importance pour nos besoins

Name	Nom de la bière	Non utilisé	Le nom de la bière n’est pas important pour les résultats recherchés
URL	URL lié la bière	Non utilisé	L’URL n’a pas d’importance pour nos besoins
Style	Nom du style auquel la bière est associé	Sortie	Permet d’identifier la bièrece comme appartenant a un style
StyleID	ID du style de la bière	Sortie	Correspond a un style de bière
Size(L)	Quantité de bière réalisée avec la bière listée	Sortie	Ne fait pas partie des sorties demandées, mais est le résultat de la recette
OG	Original Gravity : Gravité originale de la bière, Densité du moût avant fermentation	Entrée	Densité du moût: est déterminé par la quantité de sucre dans le moût, Donc un élément de recette.
FG	Final Gravity : Gravité finale de la bière, Densité du moût après fermentation		Densité du moût après l'étape de fermentation. Peut être prédite selon certains critères, mais pas décidée à l'avance
ABV	Alcohol By Volume: volume d’alcool par litre	Sortie	Une des sorties demandées par le client
IBU	International Bitterness Unit: Unite internationale d’amertume	Sortie	Une des sorties demandées par le client
Color	Couleur de la bière	Sortie	N’est pas déterminable à l’avance, dépend de la recette
BoilSize	Quantite de liquide au debut de l’ébullition	Non utilisé	etape intermédiaire de la recette
BoilTime	Durée de l’ébullition	Entrée	Temps durant lequel

			le moût doit bouillir. Change drastiquement le type de bière.
BoilGravity	Gravité de l'ébullition	Entrée	
Efficiency	Quantité de sucres extraits après la cuisson	Entrée	
MashThickness	Epaisseur de la purée	Non utilisé	Trop peu de données
SugarScale	Echelle de teneur en sucre		Données communes pour la grande majorité des recettes
BrewMethod	Methode de brassage de bière	Colonne transformée (One-hot)	4 valeurs possibles: BIAB, All-grain, Extract, Partial Mash
PitchRate	Taux d'ensemencement		Trop peu de données renseignées
PrimaryTemp	Temperature de fermentation primaire		Trop peu de données renseignées
PrimingMethod	Methode de priming	Non utilisé	Trop peu de données renseignées
PrimingAmount	Quantité de priming		Trop peu de données renseignées
UserID	Identifiant de l'utilisateur	Non utilisé	Donnée n'ayant pas d'intérêt dans notre cas

Ensuite, certaines données semblaient posséder des aberrations.
En utilisant le langage python et la librairie pandas, il nous était possible de visualiser les détails des données.

	StyleID	Size(L)	OG	FG	ABV	IBU	Color	BoilSize	BoilTime	BoilGravity	Efficiency	BrewMethod
count	70871.000000	70871.000000	70871.000000	70871.000000	70871.000000	70871.000000	70871.000000	70871.000000	70871.000000	70871.000000	70871.000000	70871
unique	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4
top	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	All Grain
freq	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	47778
mean	59.801738	44.849624	1.416692	1.077640	6.130476	44.782874	13.338915	50.795041	65.126046	1.353955	66.214736	NaN
std	56.793392	183.982380	2.229167	0.438967	1.877925	42.708998	11.896019	197.085525	15.037039	1.930989	14.157417	NaN
min	1.000000	1.000000	1.000000	-0.003000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	NaN
25%	10.000000	18.930000	1.051000	1.011000	5.070000	23.900000	5.150000	21.000000	60.000000	1.040000	65.000000	NaN
50%	34.000000	20.820000	1.058000	1.013000	5.790000	36.160000	8.350000	28.000000	60.000000	1.047000	70.000000	NaN
75%	109.000000	24.000000	1.068000	1.017000	6.820000	56.740000	16.670000	30.000000	60.000000	1.060000	75.000000	NaN
max	176.000000	9200.000000	34.034500	23.424600	54.720000	3409.300000	50.000000	9700.000000	240.000000	52.600000	100.000000	NaN

Avec l'étude des données précédentes et des recherches internet sur le métier de brassage de bière afin de s'assurer de la véracité des données, nous avons conclu que certaines des données ne pouvaient être gardées pour la suite de notre travail, notamment sur les colonnes:

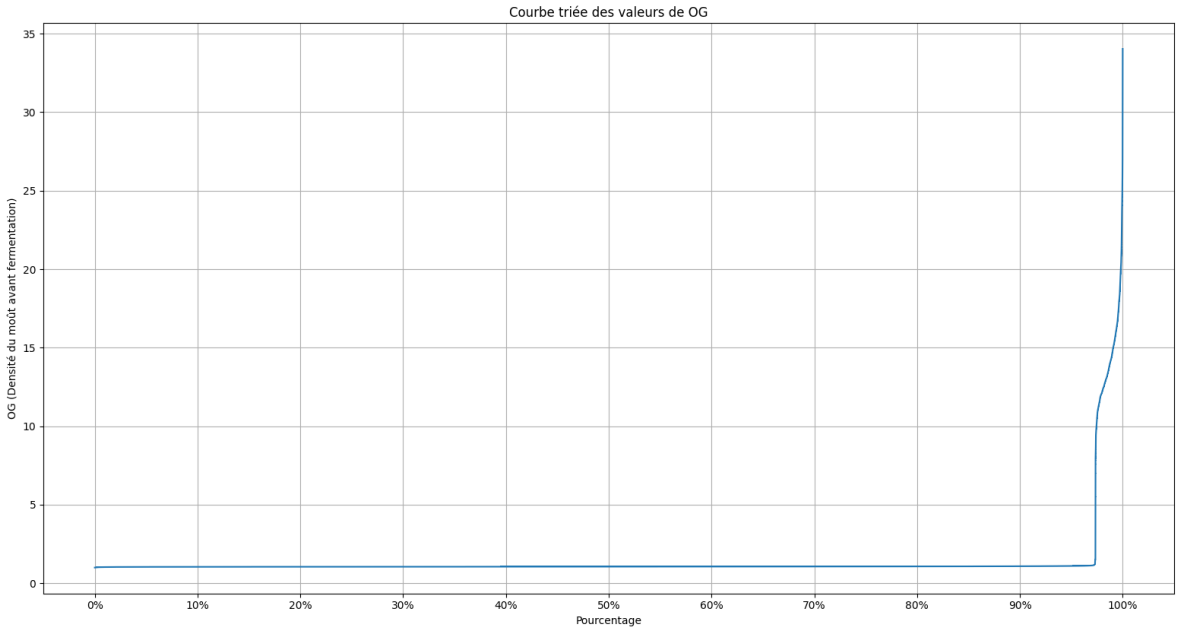
- Size(L)
- OG
- FG
- IBU
- BoilSize

Pour l'expliquer, nous avons pris comme exemple les données de la colonne OG (Densité du moût avant fermentation). Avec la librairie Pandas, il nous était facilement possible de nous renseigner sur les détails de ces données.

	OG
count	70871.000000
mean	1.416692
std	2.229167
min	1.000000
25%	1.051000
50%	1.058000
75%	1.068000
max	34.034500

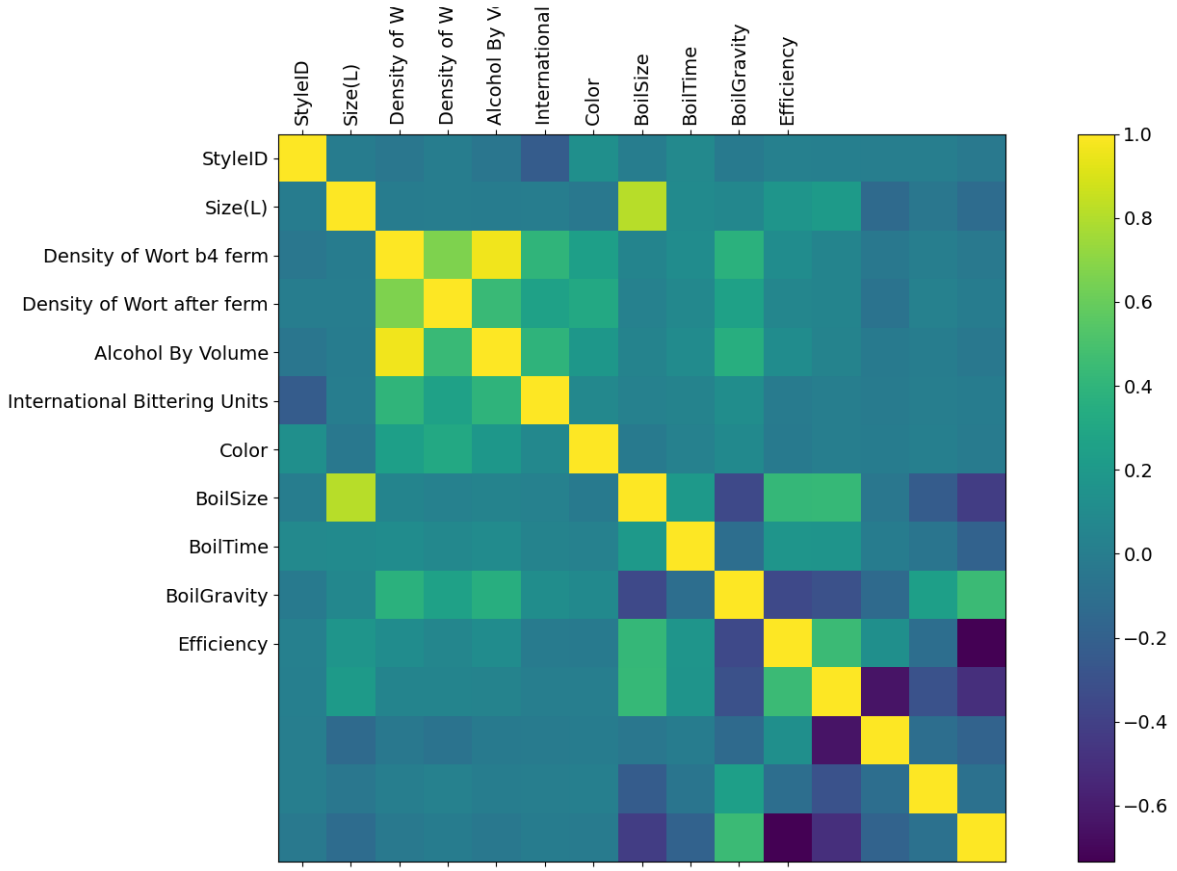
Nous avions constaté que pour un total de 70871 valeurs, nous avons une moyenne (mean) à 1,416692 avec comme valeur minimale à 1,0 et valeur maximale à 34,034500.
Cependant lorsqu'on observe les quartiles, nous n'atteignons toujours pas la valeur maximale même à 75%.

Nous devions donc avoir une meilleure visibilité de ces valeurs afin de déterminer quelles valeurs étaient à filtrer.

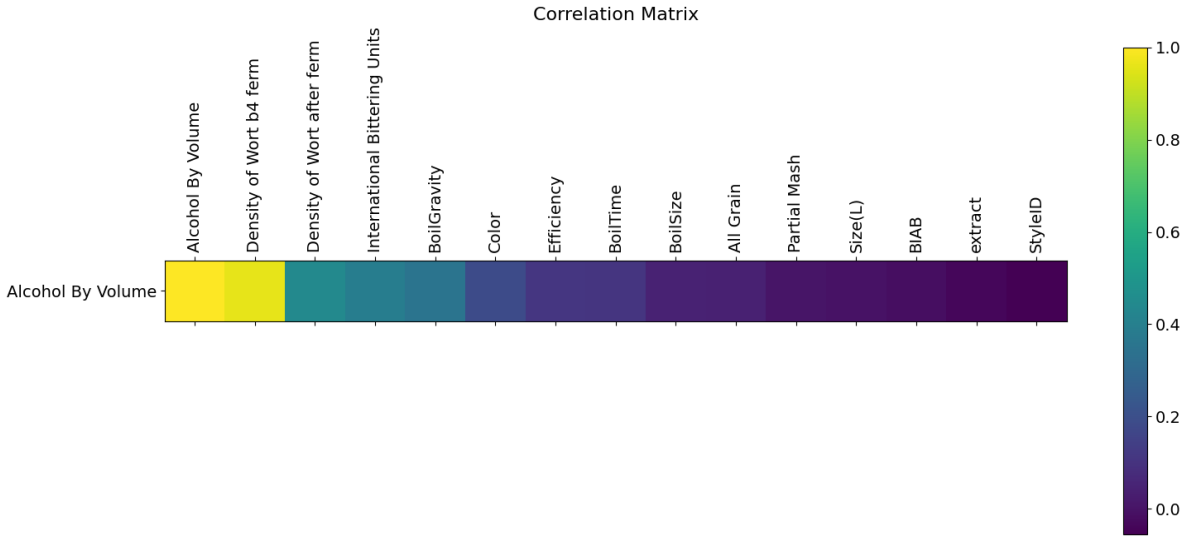


Pour en être sûr, nous avions tracé une courbe à partir de ces données triée (asc).
Nous avons donc constaté que les données deviennent très élevées par rapport aux autres à partir d'environ 95% de la quantité. Ces valeurs nous semblent fausses. En plus que d'avoir que très peu de valeurs aussi hautes et donc peu importantes, avec quelques recherches sur le métier la plupart des mesures de densité avant fermentation se trouvent très souvent entre 1040 et 1100.
Nous avons donc pris la décision de filtrer ces données.

Ce travail a été fait sur chacune des colonnes importantes à nos yeux.
Après avoir filtré toutes les aberrations et les absences de données dans les enregistrements, nous nous sommes retrouvés de 73861 enregistrements totaux à 57237 enregistrements, ce qui représente 80% de données totales. Cette réductions d'enregistrements ne devrait pas impacter notre travail étant donné qu'il nous reste suffisamment d'enregistrements.



Suite à nos recherches, nous avons pu identifier une forte corrélation entre le degré d'alcool et la densité du moût avant fermentation.



L’amertume cependant, ne semble pas avoir de corrélation aussi forte avec d’autres paramètres.