**420-A55-SF ANALYSE EXPLORATOIRE DES DONNÉES groupe 12113**

TP 02

**Étudiant:**

Mohamed Toros

Matricule : 2097154

Travail présenté au :

**Professeur**

Pierre-Marc Juneau

Date : 22 Avril 2021

TABLE DES MATIÈRES

[Préparation et prétraitement des données 3](#_Toc68185292)

[Nettoyage : 3](#_Toc68185293)

[Exploration : 3](#_Toc68185294)

[Variables et facteurs influençant la productivité 3](#_Toc68185295)

[Analyse de sensibilité (Monté Carlo) (code script python) 6](#_Toc68185296)

[Impact du prix du bois ($ / pmp) sur la distribution des profits 6](#_Toc68185297)

[Note technique 7](#_Toc68185298)

# Q0 - Préparation et prétraitement des données

## Nettoyage :

## Exploration :

# Q1 - Corrélations et régressions pour les facteurs influençant le plus la production

|  |  |
| --- | --- |
| Variable dépendante | Y |
| Q1.1 | Débit de sève (L/j) |
| Q1.2 | Sucre dans la sève (%) |
| Q1.3 | Transmittance du sirop (grade) (%) |
| Q1.4 | La productivité en sève par saison (L/entaille) |

Afin d’établir un modèle régression multilinéaire, permettant d’expliquer les variables dépendantes ci-dessus :

Il faut tout d’abord cerner les variables explicatives en lien avec chaque variable Y. Effectivement, une exploration préliminaire après la phase de prétraitement des données brutes (donneesCabneSucre = pd.read\_csv('CabaneASucrev0r2.csv')) et ensuite comprendre les liens avec ces dernières, il faut filtrer nos répresseurs (par exemples ceux qu’ont en lien avec les Pixels il faut l’enlever à ce niveau et garder les autres régresseurs).

* **Q1.1 - Modèle pour : Y = Débit de sève (L/j)**

En utilisant le script python me permet d’enlever les colonnes en les variables en lie avec les entités des ‘Pixels’ pour les couleurs RGB.

donneesCabneSucre2 = donneesCabneSucre1[donneesCabneSucre1.columns.drop(list(donneesCabneSucre1.filter(regex= 'Pixel')))]

Après ça on suit le processus de produire le dit modèle selon les étapes dresser dans le tableau ci-dessous :

|  |  |
| --- | --- |
| Étape 0 :  X = donneesCabneSucre2  X = X.drop("Débit sève (L/j)",1)  X = X.drop("Date",1)  X = X.drop("Année",1)  X = X.drop("Classe Sirop",1)  Y = donneesCabneSucre2["Débit sève (L/j)"] | **Préparation de notre simulation :**  A ce niveau, on doit enlever les autres variables comme date et la variable catégoriale ‘Classe Sirop’, et préciser notre variable à expliquer sujet d’étude (Y🡪 Débit sève (L/j) |
| **Étape 1 (itération 1)**  import statsmodels.api as sm  modele = sm.OLS(Y,X.assign(const = 1))  resultats = modele.fit()  Y\_chap = resultats.predict(X.assign(const = 1))  resultats.summary() | **Exécution le script de simulation du modéle:**  A ce stade on initialise la simulation de notre modèle pour la première fois (1ére itération) avec l’ensemble initiale des régresseurs stocké dans la variable X, voir l’inventaire des résultats obtenus (objet fig. 1). Ensuite, on élimine les régresseurs dont les coefficients β dépassent 5% ( ici : P > |t| supérieure à 0.05 c’est l’objet des variables encadré dans la figure1). Et on garde les autres variables pour simulation futur a une 2éme itération (résultats de synthèse voir figure2). Pour les régresseurs qui restent candidats pour la 2eme itération est noté par (new\_features) dans notre script Python. |
| **Étape 2 (itération 2)** | On refait la simulation pour une 2éme itération avec seulement les régresseurs restantes de l’étape 1. Ainsi, les résultats de synthèses (objet du figure 2). On examine les valeurs de β pour chaque variable, il apparait qu’il faut  Éliminé à nouveau deux variables il s’agit de **‘Jour calendrier saison’** et ‘**Transmittance produit’.** |
| **Étape 3 (itération 3)** | On suit le même processus comme a l’étape 2. Mais on observe que notre modèle est finalement est stabilisé car toutes les variables ont un β inférieur à 0.05. Ainsi les variables qui expliquent note variable a expliqué ou de sortie  **"Débit sève (L/j)"** est l’ensembles des **varibles. {['Temp max.(°C)','Temp min.(°C)','Précip. Tot. Hiver (mm)','Précip. tot. (mm)','Production moyenne par entaille (L)','Sucre sève (%)','Temps bouilloire (h)','Quantité de sirop obtenue (L)']}**  **Par contre le calcul de VIF montre autres informations aussi.** |
| **Étape 4** | A ce stade pour avoir plus de visibilité sur notre modèle. Afin de réduire de plus en plus le nombre des régresseurs, et faire face aux problèmes probables en lien des fois avec la multicolniéarité entre les régresseurs. Dans cette perspective, on calcule le facteur d’inflation de variance (VIF). Effectivement, si on utilise l’ensemble des régresseurs obtenus à la 3éme itération. Pour simuler à quel ordre chaque variable **xi** est corrélé par rapport aux autres. La figure 4 donne les résultats des valeurs selon l’ordre de corrélation. |
| **Étape 5** | On simule notre modèle de synthése (avec les coffecients β) après l’élimination des variables avec VIF supérieure à 10. Mais le compromis ici on perde la qualité car le facteur R-squared est devient loin de 1  Donc on garde l’ensemble final obtenu à la simulation du modèle à l’itération 3. |
| **Régresseurs finaux est l’ensemble** | **{['Temp max.(°C)','Temp min.(°C)','Précip. Tot. Hiver (mm)','Précip. tot. (mm)','Production moyenne par entaille (L)','Sucre sève (%)','Temps bouilloire (h)','Quantité de sirop obtenue (L)']}** |

Text

Description automatically generated

Figure 1 : Résultat simulation modèle de régression multivariés (1ére itération)

A picture containing text, plaque, screenshot

Description automatically generated

On enlève à nouveau les variables :

**Jour calendrier saison**

Et **Transmittance produit** avec des coefficients β dépassent 5%. (0.112 > 0.5 0.054 > 0.05)

Figure 2 : Résultat simulation modèle de régression multivariés (2ére itération)

R-squared est pertinent

Les coefficients β sont tous significatifs

A picture containing text, plaque, screenshot

Description automatically generated

Figure 3 : Résultat simulation modèle de régression multivariés (3ére itération)

Graphical user interface, application

Description automatically generated

Malgré que les tous les β sont significatifs, mais le facteur R-squared devient plus de 1 versus celui obtenu a l’itération3

Les coefficients β sont très significatifs.

On remarque que le VIF ici dépasse 10. Donc on élimine les variables en question et on simule notre modèle à nouveau

Figure 4 : Calcul de VIF (avec les variables **Xi** de l’itération 3)

A picture containing text, plaque, screenshot

Description automatically generated

Figure 5 : Modèle simulé à l’étape 4

* **Q1.2 - Modèle pour : Y = Sucre dans la sève (%)**

Donc on suit les démarches comme le cas Q1.1. On trouve les résultats suivants :

Text

Description automatically generated

Figure 6. Résultat de simulation du modèle (1ére itération)

A picture containing text, plaque, screenshot

Description automatically generated

Tous les β sont

Significatifs avec R-squared pertinent aussi.

Figure 5. Résultat de simulation du modèle (2ére itération)

Mais ça n’empêche pas d’examiner la multicolinéarité comme avant. En calculant les valeurs de VIF. Le figure ci-dessous illustre les résultats suivants :

Graphical user interface, application

Description automatically generated

Figure 7. Valeurs de VIF à l’itération 2.

Si on supprime juste les deux premières variables par exemple, on obtient plus de visibilité, après la simulation du modèle encore une fois.

A picture containing text, plaque, screenshot

Description automatically generated

Figure 8. Résultats de synthèse à l’itération 3 (via le calcul de VIF)

Si on calcul les valeurs de VIF de nouveau pour les régresseurs restant on obtient :

A screenshot of a website

Description automatically generated with low confidence

Figure 9. Calcul des valeurs de VIF après l’itération 4.

Donc, on remarque ici, qu’un problème de multicolinéarité est réglé à ce nouveau par l’élimination les dites variables. Ce qui montre l’intérêt de calcul de VIF comme boussole pour avance la recherche d’un modèle optimal.

|  |  |
| --- | --- |
| **Régresseurs finaux est l’ensemble** | **{'Précip. Tot. Hiver (mm)','Précip. tot. (mm)',’ 'Nombre épisodes gel/dégel'}** |

* **Q1.3 - Modèle pour : Y = % de transmittance du sirop (grade)**

**A picture containing text, plaque

Description automatically generated**

Figure 10. Tableau de synthèse à l’itération 1

A picture containing text, plaque, screenshot

Description automatically generated

Figure 11. Tableau de synthèse des valeurs à l’itération 2

On remarque qu’on juste une seule variable qui reste avec une chute de la valeur de R-sqaured et loin de 1. Donc on conserve les résultats avec le modèle précédent. Ce qui valider aussi à ce niveau par le calcul de VIF, en effet on obtient les résultats suivants de VIF:

Graphical user interface, application

Description automatically generated

Figure 10. Les résultats de calcul de VIF. (Itération 01)

* **Q1.4 - Modèle pour : Y = La productivité en sève par saison (L/entaille)**

Si on suit toujours les mêmes démarches de notre processus, on trouve après simulation de notre modèle à la première itération les résultats de synthèse suivants :

A picture containing text, plaque, screenshot

Description automatically generated

Figure 11. Tableau de synthèse après simulation 1ére itération

A ce niveau on va enlever les variables avec des coefficients beta plus élevés à 5%. Et on simule encore une fois le modèle et on obtient les résultats illustrés dans le tableau de synthèse ci-dessous :

A picture containing text, plaque, screenshot

Description automatically generated

Figure 12. Tableau de synthèse après simulation 2ére itération

En. Basant les résultats du tableau ci-dessus. On remarque que tout est beau mais pour ce qui concerne des problèmes probables en lien avec la multicolinéarité entre les variables xi de la matrice X, il faut examiner à ce stade la méthode du calcul de VIF. En effet, ce calcul donne comme résultat :

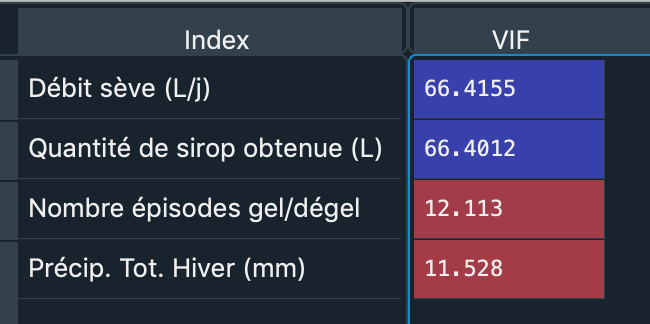


Figure 13. Calcul de VIF des variables qui correspond à l’itération 2

Ma stratégie versus cette situation repose à enlever d’abord les deux premières variables qui présentent une forte corrélation avec les autres variables.

A picture containing text, screenshot, plaque

Description automatically generated

Figure 14. Tableau de synthèse après simulation 3ére itération

Effectivement, notre modèle est optimisé avec un ensemble des régresseurs cerné a deux variables {‘Précip. Tot. Hiver (mm)’, ‘Nombre èpisode gel/dégel’}. D’autre par le calcul à nouveau les valeurs de VIF on valide une fois pour toute notre décision.

A picture containing text, outdoor, screenshot

Description automatically generated

Figure 15. Calcul de VIF des variables qui correspond à l’itération 3.

# Q2-Variations de la transmittance et du contenu en sucre de la sève d’une année à l’autre

Tout d’abord pour analyse empirique on génère un graphique de boite a moustache (boxplot) pour les années 2014, 2015 et 2016. Et pour une meilleure visibilité on procède d’abord à la suppression des valeurs abérrantes afin de ne baiser notre modèle et par la suite l’analyse associé.

* **Q2.1 – Sucre sève**

Chart, box and whisker chart

Description automatically generated

**Figure 15. Boite à moustache du pourcentage du sucre sève par année (2014,2015 et 2016).**

|  |  |  |
| --- | --- | --- |
| Année | Calcul du moyen | **μAnnée** |
| 2014 | sucre2014.mean() | **μ2014** = 2.1928235294117644 |
| 2015 | sucre2014.mean() | **μ2015** = 2.1287999999999996 |
| 2016 | sucre2014.mean() | **μ2016** = 2.6519101123595505 |
| Le calcul de P-Value pour voir la tendance globale entre les 3 années = 5.28x10-28  Comme illustré dans les résultats du de la figure 16. Donc on constate que le p-value est très petite donc il y a une tendance entre les 3 années 2014, 2015 et 2016. | | |

Table

Description automatically generated

**Figure 16. Résultats des indices du tendance globale qui concerne le pourcentage du sucre sève entre les années (2014,2015 et 2016).**

Donc ici on accépte l’hypothèse H0 : μ2014 = μ2015 = μ2016

Si on fait le même calcul, mais cette fois-ci entre deux années (deux à deux).

|  |  |
| --- | --- |
| P-Value entre deux années : | Calcul correspondant aux deux années |
| pvalue02=sts.ttest\_ind(sucre2014,sucre2015) | Graphical user interface, application  Description automatically generated  P-Value ici non significatif donc il n’y a pas une tendance. Donc on rejet le H0 : μ2014 = μ2015 |
| pvalue03=sts.ttest\_ind(sucre2014,sucre2016) | P-Value ici est significatif donc il y a une tendance entre les deux années 2014 et 2016 pour qui concerne sucre sève. Donc on accépte  L’hypothèse :  H0 : μ2014 = μ2016  Graphical user interface, application  Description automatically generated |
| pvalue04=sts.ttest\_ind(sucre2015,sucre2016) | Graphical user interface, application  Description automatically generated  P-Value ici est significatif donc il y a une tendance entre les deux années 2015 et 2016 pour qui concerne sucre sève.  Donc on accépte  L’hypothèse  H0 : μ2015 = μ2016 |

* **Q2.2 –**T**ransmittance produit (%)**

**Chart, box and whisker chart

Description automatically generated**

**Figure 17. Boite à moustache du pourcentage du transmittance produit (%)**

**selon les années (2014,2015 et 2016).**

|  |  |  |
| --- | --- | --- |
| Année | Calcul du moyen | **μAnnée** |
| 2014 | Transmittance2014.mean() | **μ2014** = 44.03659090909091 |
| 2015 | Transmittance2015.mean() | **μ2015** = 40.29711538461538 |
| 2016 | Transmittance2016.mean() | **μ2016** = 43.914719101123595 |
| Le calcul de P-Value pour voir la tendance globale entre les 3 années = 0.10211388  Comme illustré dans les résultats du de la figure 18. Donc on constate que le p-value n’est pas significatif car dépasse 0.05, donc il n’y a une tendance a observé entre les 3 années 2014, 2015 et 2016.  Donc ici on rejette le H0 : μ2014 = μ2015 = μ2016 | | |

**pvalue01=sts.f\_oneway(Transmittance2014,Transmittance2015,Transmittance2016)**

# Table Description automatically generated

**Figure 18. Résultats des indices du tendance globale qui concerne le pourcentage du transmittance du produit entre les années (2014,2015 et 2016).**

|  |  |
| --- | --- |
| P-Value entre deux années : | Calcul correspondant aux deux années |
| pvalue2=sts.ttest\_ind(Transmittance2014,Transmittance2015) | Graphical user interface, application  Description automatically generated  P-Value ici non significatif donc il n’y a pas une tendance. Donc on rejet le H0 : μ2014 = μ2015 |
| Pvalue3=sts.ttest\_ind(Transmittance2015,Transmittance2016) | Graphical user interface, application  Description automatically generated  P-Value ici est significatif donc il y a une tendance entre les deux années 2015 et 2016 pour qui concerne sucre sève. Donc on accépte  L’hypothèse :  H0 : μ2014 = μ2016 |
| pvalue4=sts.ttest\_ind(Transmittance2014,Transmittance2016) | Graphical user interface, application  Description automatically generated  P-Value ici est significatif donc il y a une tendance entre les deux années 2014 et 2016 pour qui concerne sucre sève.  Donc on accépte  L’hypothèse  H0 : μ2015 = μ2016 |

# Note technique