**Une image contenant texte

Description générée automatiquement**

Rapport du Projet Modèles de régression

Année Universitaire : 2022-2023

Intitulé du projet : Analyser une Dataset par les modèles de régression.

Le créneau du projet : Modèles de régression

**Encadré par :**

* Mme. ELGHALI Btihal
* Mme. ABOUREZQ Manar

**Réalisé par :**

* ELGHAZI Soufiane N°13
* LABYADY Ibtissam N°23
* MAASRI Amine N°26

**Remerciements :**

Nos remerciements s'adressent principalement à nos chères professeurs Mme ELGHALI et Mme ABOUREZQ de leurs efforts remarquables afin de nous bien approcher des concepts de régression et pour nous assigner ce projet dans l’intérêt de bien mobiliser nos prérequis et de découvrir des nouvelles astuces.

**Résumé :**

Nous avons analysé un jeu de données comportant l’ensemble des données météorologiques de la ville de rabat s’étalent sur une période allant de 2000 jusqu’à 2021 « Weather\_Rabat » par les modèles de régression vue en classe en tirant le maximum d’interprétations pour chaque modèle.

Ce projet nous a fait découvrir plus profondément les modèles de régression et les appliquer comme concept de machine Learning.

Introduction Générale :

Dans le cadre de l’application de la régression sur une dataset afin d’extraire le maximum des interprétations pour chaque modèle créé, on a choisi une dataset de « *Weather\_Rabat* » issu dans le site de la Nasa.

A travers le présent rapport, nous allons commencer par décrire le Domain et la problématique de notre data set et les besoins d’analyse, Après on va présenter le dictionnaire des variables et des modalités en plus de la préparation des données, Puis nous allons détailler les différentes étapes de l’analyse avec l’évaluation de chaque modèle créé en interprétant les résultats obtenus, avant d’arriver vers la fin à parler des apports de celui-ci.

**Table des matières**

[Remerciements : 2](#_Toc112938239)

[Résumé : 2](#_Toc112938240)

[Introduction générale : 2](#_Toc112938245)

[Table des matières : 3](#_Toc112938242)

[Liste des figures 4](#_Toc112938244)

[Chapitre I : introduction 5](#_Toc112938246)

A. [Domaine : **5**](#_Toc112938247)

B. [Problématique: **5**](#_Toc112938248)

[C. les Hypotheses: **5**](#_Toc112938248)

[D. Besoins d’analyse: **5**](#_Toc112938248)

[**Chapitre II : préparation des données**  6](#_Toc112938262)

A.  [Dictionnaire des variables et des modalités : **6**](#_Toc112938247)

[B. Préparation des données: **6**](#_Toc112938264)

[Chapitre III : Modeles de régression et interprétation des résultats 8](#_Toc112938270)

A.[Régression linéaire simple : **8**](#_Toc112938247)

[B. Régression linéaire multiple : **10**](#_Toc112938272)

[C. régression régularisé: **13**](#_Toc112938275)

[E. Régression non linéaire : **13**](#_Toc112938277)

[Chapitre IV : Déploiment 14](#_Toc112938246)

[Conclusion : 15](#_Toc112938278)

[Bibiliographie : 15](#_Toc112938278)

**Liste des figures :**

[Figure 1: Dictionnaire des variables et modalités 6](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746871)

[Figure 2: résumé de jeu de données 6](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746872)

[Figure 3:histogramme de température 7](#_Toc130746873)

[Figure 4:matrice de corrélation 7](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746874)

[Figure 5:variances expliqués r² 8](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746875)

[Figure 6:scatter de température avec humidité 8](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746876)

[Figure 7:comparaison des modèles de rls 9](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

[Figure 8:qualité du modèle rls1 9](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

[Figure 9:validité de modèle rls1 10](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

[Figure 10:comparaison des modèles de rlm 10](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

[Figure 11:visualition de tem en fct des autres variables 11](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

[Figure 12:acp et matrice de corrélation 11](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

[Figure 13:résultats de l’anova 12](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

[Figure 14:résidus vs prédictions & qq-plot 12](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

[Figure 15:qualité du modèle multiple 13](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

[Figure 16:qualité des modèles régularisé 13](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

[Figure 17:qualité des modèles non linéaires 13](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

[Figure 18:qualité des modèles polynomiales 14](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

[Figure 19:validité du modèle polynomiale 14](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

[Figure 20:interface 1 de l’application 15](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

[Figure 21:interface 2 de l’application 15](file:///C:\Users\massi\Downloads\Projet\Projet\Rapport_Projet_Modèles%20de%20régression.docx#_Toc130746877)

**Chapitre I : Introduction**

**A -Domaine :**

La ville de Rabat, capitale du Maroc, est une ville connue pour son climat doux et tempéré tout au long de l’année. Cependant, les variations climatiques peuvent avoir des impacts significatifs sur l’environnement et les activités humaines. Dans ce contexte, l’analyse des données météorologiques de Rabat peut apporter une meilleure compréhension de son climat, aider à la prise de décision et à la planification des activités.

**B -Problématique**:

* Existe-t-il une relation entre la Température et Humidité ?
* Qu'en est-il entre l'humidité et la Vitesse du vent ?
* Pouvons-nous prédire la température en fonction de l’humidité ?

**C -Les hypothèses :**

En se basant sur les connaissances météorologiques actuelles, on peut formuler les hypothèses suivantes :

* Il existe une relation entre la température et l'humidité, dans le sens où une humidité élevée peut conduire à une sensation de chaleur accrue (en raison de l'augmentation de la température de l'air ressentie sur la peau), tandis qu'une faible humidité peut donner une sensation de fraîcheur.
* Il peut également y avoir une relation entre l'humidité et la vitesse du vent, dans la mesure où une humidité élevée peut favoriser la formation de nuages et de précipitations, qui sont souvent associés à des vents plus forts.
* En ajoutant d’autre variables comme les précipitations et la vitesse de vent , la précision de prédiction de modèle sera augmenté.

**D-Besoins d’analyse :**

Pour répondre aux questions posées et tester les hypothèses formulées, il serait nécessaire de procéder à une analyse statistique des données météorologiques, notamment en utilisant des méthodes de régression.

Pour la première question, une analyse de corrélation entre la température et l'humidité peut être réalisée afin de déterminer s'il existe une relation linéaire entre les deux variables. Cette analyse pourrait être complétée par des graphiques de dispersion pour visualiser la relation entre les deux variables.

Pour la seconde question, une analyse de régression pourrait être utilisée pour établir un modèle statistique qui prédit la température en fonction de l'humidité et d'autres variables pertinentes. Ce modèle pourrait ensuite être validé en utilisant des données de validation pour évaluer sa précision et sa fiabilité.

Enfin, pour la troisième hypothèse, il serait nécessaire d'effectuer une analyse de régression multiple, en utilisant plusieurs variables météorologiques (température, humidité, précipitation, vitesse du vent, etc.) pour prédire la température. Cette analyse permettrait de déterminer l'importance relative de chaque variable dans la prédiction de la température et d'identifier les variables les plus significatives à inclure dans le modèle.

[**Chapitre II :**](#_bookmark8) **Préparation des données.**

**A -Dictionnaire des Variables et des Modalités :**



* Année
* Mois
* Jours
* Date
* Température à 2 Mètres (C).
* Précipitation corrigé (mm/jour).
* Humidité spécifique à 2 Mètres(g/Kg).

**Figure1 : Dictionnaire des variables et modalités**

* Une image contenant texte, lettre

  Description générée automatiquementVitesse Vent à 2 Mètres (m/s)

**B - Préparation des données :**

La préparation des données est une étape importante dans l'analyse de données, pour cela il est nécessaire de réaliser une analyse exploratoire (EDA) afin de bien comprendre sur notre jeu de donnés.

**Figure2 : résumé de jeu de données.**

Une image contenant texte

Description générée automatiquement



Tous d’abord notre jeu de donnés ne contient pas des valeurs manquantes, et contient 7761 observations et 7 colonnes, ainsi les valeurs de min et max des variables sont logique (Positives et raisonnables) ce qui signifie qu’on n’a pas des valeurs aberrantes.

Ainsi visualiser la distribution de la variable cible à l'aide d'un histogramme peut aider à comprendre la nature de la variable cible et à identifier les tendances et les relations avec d'autres variables.

***Une image contenant graphique

Description générée automatiquement***

**Figure3 : Histogramme de température**

On remarque que la distribution de la Température est symétrique, et l’absence des valeurs aberrantes qui se situent loin du reste des données.

En plus On remarque que les valeurs de la température ne sont pas distribuées d’une manière parfaitement normale.

Afin de savoir si les variables sont corrélées entre eux il est nécessaire de vérifier les corrélations en se basant sur le coefficient de Person r comme si nos variables sont quantitatives.

Une image contenant texte, lettre

Description générée automatiquementUne image contenant table

Description générée automatiquement

**Figure4 : Matrice de corrélation**

On remarque que l’humidité est corrélée positivement et fortement avec notre variable cible, en plus les valeurs de p <0.OO1, ce qui montre leurs significativités dans le cadre d’une population.

Puisque l’intervalle de confiance de coefficient r de Pearson ne contient pas 0 et il est restreint, cela signifie qu’il est significatif si on prend beaucoup d’échantillons. Et comme le r suit une loi de student la valeur de t est 173.71 avec degré de liberté de 7759. Et la valeur de r varie sur plusieurs échantillons entre [0.8872436, 0.8963485].



Une image contenant graphique

Description générée automatiquement**Une image contenant table

Description générée automatiquement**

**Figure6 : Scatter de Température avec humidité**

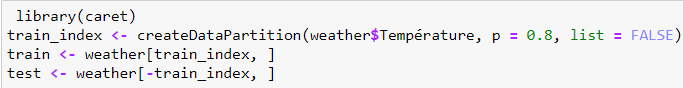
**Figure5 : Variances expliqués R²**

La variance explique autrement dit r carré, nous montre qu’on a expliqué 79.55% de la variance de Température par humidité.

Nous choisissons des caractéristiques qui ont une forte corrélation avec notre variable cible, température, pour ajuster un modèle de régression linéaire.

On remarque que les observations presque suivent une ligne droite sauf des valeurs extrêmes qui situent aux extrémités

Après on a diviser notre jeu de données en 80% pour l’entrainement des modèles et 20% pour les évaluer.



[**Chapitre III :**](#_bookmark8) **Modèles de régression et interprétation des résultats.**

Nous avons proposé plusieurs modèles de régression pour prédire la température. Nous avons commencé par les modèles de régression simple pour chaque variable météorologique, puis nous avons progressé vers les modèles de régression multiple pour prendre en compte les effets conjoints de plusieurs variables. Nous avons également utilisé des modèles de régression régularisée pour limiter les effets des variables moins significatives et des modèles de régression non linéaire pour prendre en compte des relations complexes entre les variables.

**A -Régression linéaire simple :**

Une image contenant texte, lettre

Description générée automatiquement

On a construit différents modèles entre la Températures et différente variable indépendante séparément, après on a utilisé les deux critères de d’information d’Akaike et celle Bayesien (plus ce sont petits, mieux est le modèle).

On remarque le modèle rls1 est le meilleur parmi ces modèles.

Une image contenant table

Description générée automatiquement

**Figure7 : Comparaison des modèles de RLS**

Alors après le choix du modèle, on va examiner sa qualité.

Une image contenant texte, lettre

Description générée automatiquement

* Le R² ajusté : plus il est proche de 1, plus le modèle est bon, dans notre cas R² ajusté=0.7977
* Le RSS (Residual Sum of Squares) : plus il est petit, plus le modèle est bon, RSS=1.885

Et on constate que l'intervalle de confiance pour la ponte ne contient pas zéro

([1.51, 1.55]) alors c'est bon / il y a une bonne corrélation entre les deux variables.

Une image contenant table

Description générée automatiquement

**Figure8 : Qualité de modèle RLS1**

Après on va tester la validité du modèle, Le modèle de RLS est considéré comme valide si les résidus sont

1.Indépendants :

* Test de Durbin-Watson :

Tester l’hypothèse nulle H0 = les résidus sont indépendants. Si la p-value < 0,05 l’hypothèse est rejetée

2.Distribués selon une loi normale de moyenne 0 :

* Test de Shapiro-Wilk

Tester H0 = les résidus suivent une loi normale.

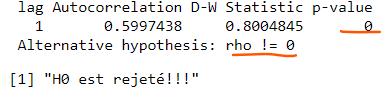
Si la p-value < 0,05 l’hypothèse est rejetée

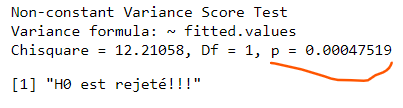
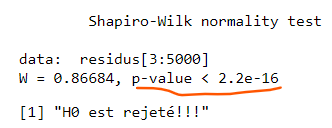
3.Distribués de façon homogène (i.e. variance constante) :

* Test de Breush-Pagan

Est-ce H0 = les résidus sont distribués de façon homogène.

Si la p-value < 0,05 l’hypothèse est rejetée





**Figure9 : Validité de modèle RLS1**

Alors le modèle rls1 n’est pas valide !

**B -Régression linéaire multiple :**

On a construit différents modèles en combinant les différentes variables indépendantes,



Une image contenant texte

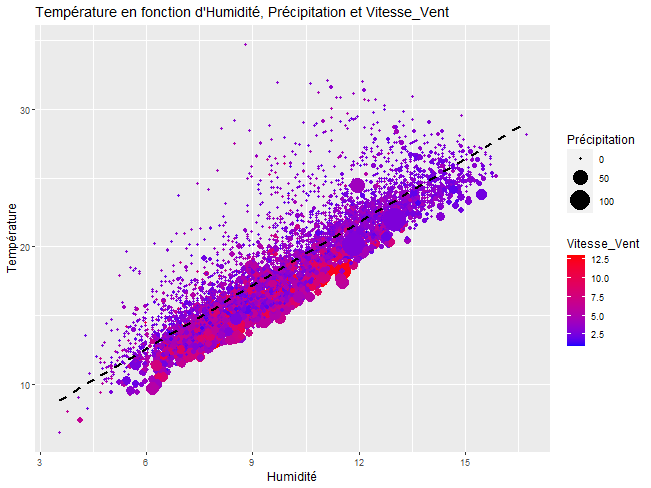
Description générée automatiquementUne image contenant texte

Description générée automatiquement

**Figure10 : Comparaison des modèles de RLM**

On remarque que le model multiple 1 qui est mieux entre ces modèles.

Après on va visualiser la Température en fonction des trois variables indépendantes.



**Figure11 : Visualisation Temp en fct des autres**

**Variables**

On observe l’existence des valeurs extrêmes et que la majorité des points sont aligné, malgré qu’il soit difficile de projeter les données en 2D mais on peut observer la condensité des valeurs au voisinage de la droite de régression.

Après dans la régression multiple il est nécessaire d’étudier la multicollinéarité qui peut influencer sur nos résultats lorsque les variables indépendantes sont fortement corrélées entre elles.

Une image contenant table

Description générée automatiquement

Tout d’abord on constate que les variables indépendantes ne sont pas fortement corrélées entre elles mais cela n’est pas suffisant.

Facteur d'inflation de la variance (VIF) : est un indicateur de la quantité de variance dans un coefficient de régression qui est due à la corrélation avec d'autres variables. Dans notre cas les valeurs de VIF sont proches de 1 (inférieure à 5), ce qui suggère que la multicollinéarité entre les variables indépendantes est faible. Par conséquent, il est peu probable que la multicollinéarité ait un impact significatif sur les résultats de notre modèle.

La ACP est une méthode pour réduire la dimensionnalité des données en combinant les variables indépendantes en nouvelles variables appelées composantes principales. Dans notre cas, PC1 explique plus de la moitié de la variance, ce qui en fait la composante principale la plus importante. Les résultats de l'analyse ACP peuvent être utilisés pour réduire la dimensionnalité des données.

Une image contenant texte, orange, capture d’écran

Description générée automatiquementUne image contenant graphique

Description générée automatiquement

**Figure12 : ACP et matrice de corrélation**

Une image contenant table

Description générée automatiquement

**Figure13 : Résultats de l’ANOVA**

En somme, ces résultats suggèrent que les variables indépendantes ont un effet significatif sur la température et que le modèle de régression linéaire multiple est un ajustement approprié pour les données.

Les résultats de l'ANOVA incluent également une évaluation de la qualité globale du modèle. Le F-value, qui mesure le rapport entre la variance expliquée par le modèle et la variance non expliquée, est élevé pour chaque variable indépendante, indiquant que le modèle est globalement significatif. Le modèle dans son ensemble explique également une proportion significative de la variance de la température, car les résidus ont une variance résiduelle faible (mean square residual = 3). Enfin, les signif. codes dans le tableau indiquent que toutes les variables indépendantes ont une influence significative sur la température.

Une image contenant graphique

Description générée automatiquementMaintenant on va tester la validité du modèle comme on a déjà fait pour le modèle rls1.

**Figure14 : résidus vs prédictions & QQ-plot**

La courbe de résidus vs valeurs prédites par modèle utilisé pour évaluer la qualité d'ajustement d'un modèle de régression, résume les trois conditions qu’on a constaté déjà.

La courbe de résidus est plate et centrée autour de zéro, cela indique que le modèle est bien ajusté aux données et que les erreurs sont aléatoires et homogènes.

La courbe de résidus ne présente pas une tendance à l'augmentation ou à la diminution de la dispersion à mesure que les valeurs

Prédites augmentent, cela indique que le modèle ne présente pas une hétéroscédasticité.

Le QQ-plot nous permet de visualiser la distribution d'un échantillon de données en comparaison avec une distribution théorique connue, telle que la distribution normale.

Les points sur le QQ-plot presque tomberont sur une ligne droite ce qui indique une adéquation n’est pas mal entre les données.

En somme, le modèle est valide !!

Une image contenant graphique

Description générée automatiquement

Une image contenant texte, lettre

Description générée automatiquementUne image contenant graphique

Description générée automatiquement

**Figure15 : Qualité du modèle multiple**

Cette courbe montre les valeurs prédites par le modèle et les vraies valeurs de température dans le jeu de données du test. On remarque une bonne prédiction comme la qualité du modèle nous montre : R²=0.8099 élevé par rapport à celle de rls1.

**C -Régression Régularisé :**

On a réalisé la régularisation par les trois méthodes, Ridge, Lasso et ElasticNet par glmnet () en modifiant la valeur de alpha ,0 pour Ridge, 1 pour Lasso et valeur entre 0 et 1 pour ElasticNet, en préparant une liste des valeurs de lambda afin de trouver la valeur optimale pour chaque modèle en calculant R² et l'erreur quadratique moyenne (EQM) de chaque modèle, on a trouvé :

R² égal à 1 est souvent un signe d'overfitting. Cela signifie que notre modèle a réussi à expliquer complètement les données d'entraînement et a donc peut-être capturé le bruit au lieu de la tendance sous-jacente dans les données.



Une image contenant graphique

Description générée automatiquement

**Figure16 : Qualité des modèles régularisé**

On peut constater l’overfitting dans ce graph.

Ce qui signifie que ces modèles ne sont pas valides !

**D -Régression Non linéaire :**

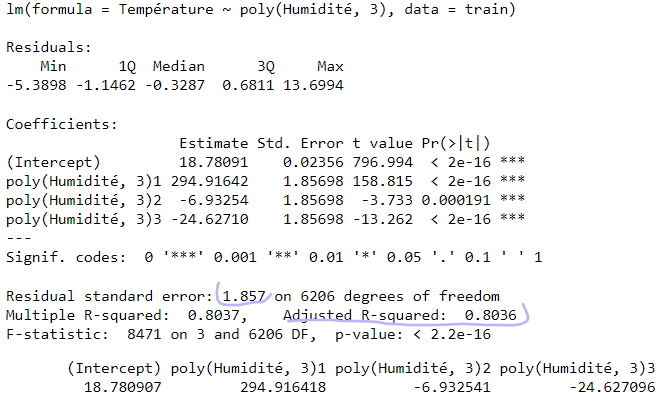
Une image contenant texte

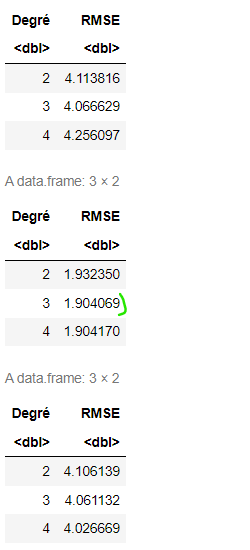
Description générée automatiquementUne image contenant graphique

Description générée automatiquement

**Figure17 : Qualité des modèles Non linéaires.**

La courbe bleue est celle du modèle de puissance qui a les meilleures mesures par rapport aux autres.

Après on a créé des modèles de régression polynomiales de degré 2, 3 et 4 pour chaque variable en calculant le RMSE de chaque modèle afin de les comparer.



Précipitation

Humidité

On remarque que le modèle polynomial de dégrée 3 d’Humidité qui donne valeur de RMSE minimale.

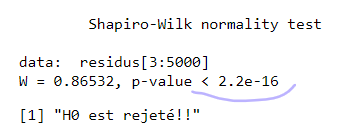
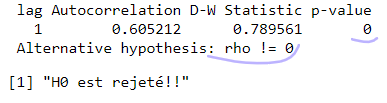
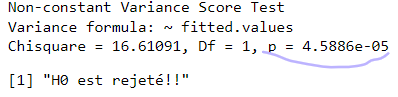
Les performances de ces modèles sont bonnes mais le modèle multiple est mieux que celui dans les deux mesures de R² et RSE.

Vent

**Figure19 : validité du modèle Polynomiale**

**Figure18 : Qualité des modèles Polynomiales**

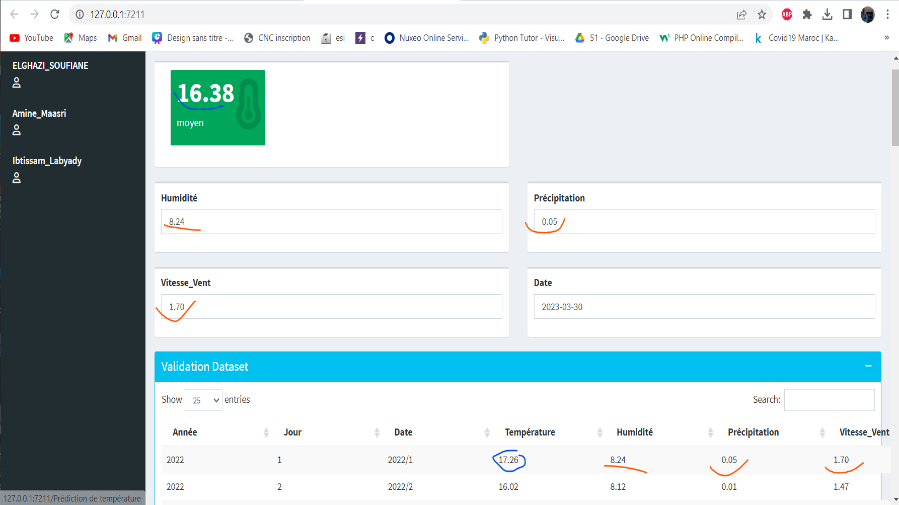
En plus on a testé la validité de ce modèle et on a trouvé qu’il n’est pas valide :



**Conclusion :** le modèle model\_multiple1 est le meilleur dans notre cas !!!

[**Chapitre IV :**](#_bookmark8) **Déploiement.**

La dernière étape d’un projet en Data Science est de le déployer, pour cela on a créé une application par la Library Shiny sous R, dans laquelle on a déployer notre modèle sous un serveur locale, tous d’abord il faut sauvegarder notre modèle sous format RDS afin de le lier avec notre application. Dans cet étape de validation de notre modèle, on a utilisé une autre dataset issue du même site de la NASA pour l’année de 2022/2023 à Rabat , afin de constater la précision de notre modèle.

******

**Figure20 : interface1 de l’application**

On entrant des valeurs d’humidité, précipitation et vitesse de vent le modèle prédit la température, en donnant l’état du climat (chaud…) et dans l’exemple au-dessus, en constate une petite différence.

******

**Figure21 : interface2 de l’application**

L’application contient aussi les graphiques de dispersion des variables, l’histogramme de température, le résumé de modèle et même le jeu de donnée de validation.

**Conclusion :**

Dans notre projet nous avons mis en œuvre une application web pour la prédiction de température en appliquant les modèles de régression sur la data set de « Weather\_Rabat » par R, l’objectif est d’appliquer les types de régression vue en classe afin extraire le maximum possible d’interprétations pour chaque modèle crée.

Pour aboutir à ce résultat, on a passé d’abord d’introduire les données à analyser et les préparer afin de les biens analyser vers la construction et l’évaluation des modèles construits.

La préparation des donnée et l’analyse exploratoire EDA sont deux étapes majeures et nécessaires, nous aide à simplifier la réalisation du projet et rend clair le Domain de notre data set ainsi notre analyse.

Quant à l’étape de l’analyse et l’interprétation elles consiste à mettre en œuvre les prérequis du cours. Elle permet d’avoir le résultat final souhaité. Elles prennent généralement le volume horaire majeur dans le traitement du sujet vu leur importances et leur complexité.

En conclusion, ce projet nous a permis également de développer notre esprit d’interprétation et de la réflexion, Il nous a permis d’enrichir et d’approfondir nos connaissances de régression, de langage R , et également du machine Learning.

**Bibliographie :**

* [Weather rabat dataset.](https://power.larc.nasa.gov/data-access-viewer/)