Université Abdelmalek Essaadi

Faculté des sciences et techniques Tanger

Logiciels & Systèmes Intelligents

***Mini projet***



***Réalisé par :***

*Asmaa OUTALEB*

*Redwane MAHJOUBI*

***Encadré par :***

*Pr. Lotfi EL AACHAK*

*Remerciements :*

*L'année universitaire touchant à sa fin, nous tenions à vous dire que vous avez été un excellent professeur ! Nous vous remercions d'avoir partagé vos connaissances avec nous, d'avoir toujours été juste dans votre éducation monsieur Lotfi EL AACHAK et de nous avoir toujours soutenus et aidés. Merci encore de la part de tous vos étudiants.*

*Tables des matières :*

[I. Introduction : 6](#_Toc103705524)

[II. **Qu’est ce que kaggle:** 8](#_Toc103705525)

[III. Description de la compétition : 10](#_Toc103705529)

[VI. Conclusion : 18](#_Toc103705548)

# Introduction :

L'apprentissage automatique est un sous-domaine de l'intelligence artificielle (IA). L'objectif de l'apprentissage automatique est généralement de comprendre la structure des données et d'intégrer ces données dans des modèles qui peuvent être compris et utilisés par les gens.

Bien que l'apprentissage automatique soit un domaine de l'informatique, il diffère des approches informatiques traditionnelles. Dans l'informatique traditionnelle, les algorithmes sont des ensembles d'instructions explicitement programmées utilisées par les ordinateurs pour calculer ou résoudre des problèmes. Les algorithmes d'apprentissage automatique permettent à la place aux ordinateurs de s'entraîner sur les entrées de données et d'utiliser l'analyse statistique afin de générer des valeurs comprises dans une plage spécifique. Pour cette raison, l'apprentissage automatique facilite la création de modèles par les ordinateurs à partir d'échantillons de données afin d'automatiser les processus de prise de décision basés sur les entrées de données.

Aujourd'hui, tout utilisateur de technologie a bénéficié de l'apprentissage automatique. La technologie de reconnaissance faciale permet aux plateformes de médias sociaux d'aider les utilisateurs à marquer et à partager des photos d'amis. La technologie de reconnaissance optique de caractères (OCR) convertit les images de texte en caractères mobiles. Les moteurs de recommandation, alimentés par l'apprentissage automatique, suggèrent les films ou les émissions de télévision à regarder ensuite en fonction des préférences de l'utilisateur. Les voitures autonomes qui s'appuient sur l'apprentissage automatique pour naviguer pourraient bientôt être disponibles pour les consommateurs.

L'apprentissage automatique est un domaine en constante évolution. Pour cette raison, vous devez garder à l'esprit certaines considérations lorsque vous travaillez avec des méthodologies d'apprentissage automatique ou analysez l'impact des processus d'apprentissage automatique.

Dans ce didacticiel, nous examinerons les méthodes courantes d'apprentissage automatique supervisé et non supervisé, ainsi que les approches algorithmiques courantes de l'apprentissage automatique, notamment l'algorithme k-plus proche voisin, l'apprentissage par arbre de décision et l'apprentissage en profondeur. Nous explorerons les langages de programmation les plus utilisés dans l'apprentissage automatique, en vous fournissant certains des attributs positifs et négatifs de chacun. De plus, nous discuterons des biais qui sont perpétués par les algorithmes d'apprentissage automatique et examinerons ce qui peut être gardé à l'esprit pour éviter ces biais lors de la création d'algorithmes.

# Qu'est-ce que kaggle ?:

Kaggle est une plateforme communautaire en ligne pour les scientifiques des données et les passionnés d'apprentissage automatique. Kaggle permet aux utilisateurs de collaborer avec d'autres utilisateurs, de rechercher et de publier des ensembles de données, d'utiliser des blocs-notes intégrés au GPU et de rivaliser avec d'autres scientifiques des données pour résoudre les défis de la science des données. L'objectif de cette plateforme en ligne (fondée en 2010 par Anthony Goldbloom et Jeremy Howard et acquise par Google en 2017) est d'aider les professionnels et les apprenants à atteindre leurs objectifs dans leur parcours en science des données grâce aux puissants outils et ressources qu'elle fournit. À ce jour (2021), il y a plus de 8 millions d'utilisateurs enregistrés sur Kaggle.

L'une des sous-plateformes qui a fait de Kaggle une ressource si populaire est leurs compétitions. De la même manière que HackerRank joue ce rôle pour les développeurs de logiciels et les ingénieurs informaticiens, les « Kaggle Competitions » ont une importance significative pour les data scientists ; vous pouvez en savoir plus à leur sujet dans notre Kaggle Competition Guide et apprendre à analyser un ensemble de données étape par étape dans notre Kaggle Competition Tutorial . Dans les compétitions de science des données comme celles de Kaggle ou de DataCamp , les entreprises et les organisations partagent une grande quantité de tâches stimulantes en science des données avec de généreuses récompenses dans lesquelles les scientifiques des données, des débutants aux expérimentés, s'affrontent à leur achèvement. Kaggle fournit également le Kaggle Notebook, qui, tout comme DataCamp Workspace, vous permet de modifier et d'exécuter votre code pour les tâches de science des données sur votre navigateur, afin que votre ordinateur local n'ait pas à faire tout le gros du travail et que vous n'ayez pas besoin de configurer vous-même un nouvel environnement de développement.

Kaggle fournit des ressources puissantes sur le cloud et vous permet d'utiliser un maximum de 30 heures de GPU et 20 heures de TPU par semaine. Vous pouvez télécharger vos ensembles de données sur Kaggle et télécharger également les ensembles de données d'autres personnes. De plus, vous pouvez vérifier les ensembles de données et les blocs-notes d'autres personnes et lancer des sujets de discussion à leur sujet. Toute votre activité est notée sur la plateforme et votre score augmente à mesure que vous aidez les autres et partagez des informations utiles. Une fois que vous commencez à gagner des points, vous serez placé sur un classement en direct de 8 millions d'utilisateurs de Kaggle.

Kaggle convient à différents groupes de personnes, des étudiants intéressés par la science des données et l'intelligence artificielle aux data scientists les plus expérimentés au monde. Si vous êtes débutant, vous pouvez profiter des cours dispensés par Kaggle. En rejoignant cette plateforme, vous pourrez évoluer au sein d'une communauté de personnes de différents niveaux d'expertise, et vous aurez la chance de communiquer avec de nombreux data scientists très expérimentés. Au fur et à mesure que vous gagnez des points et des médailles Kaggle, qui sont la preuve de vos progrès, il est tout à fait possible que vous finissiez même par attirer des chasseurs de têtes et des recruteurs, et débloquer de nouvelles opportunités d'emploi.

Enfin, lorsque vous postulez à des emplois en science des données, mentionner votre expérience Kaggle a définitivement un impact positif. Il va sans dire que tous ces avantages s'appliquent également aux data scientists très expérimentés. Quelle que soit votre expérience, cette plateforme offre des possibilités d'apprentissage et d'amélioration continues, et, bien sûr, les récompenses en espèces qui peuvent accompagner les compétitions sont tout aussi intéressantes.

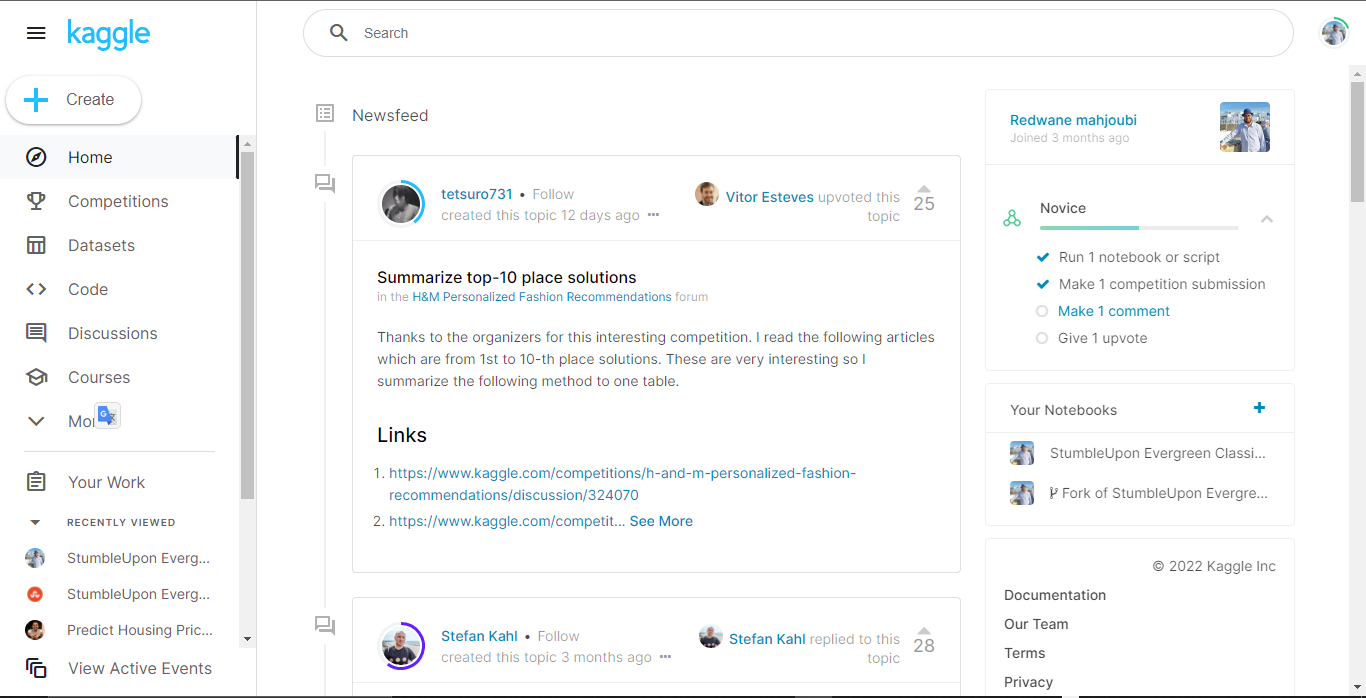


Figure Kaggle.

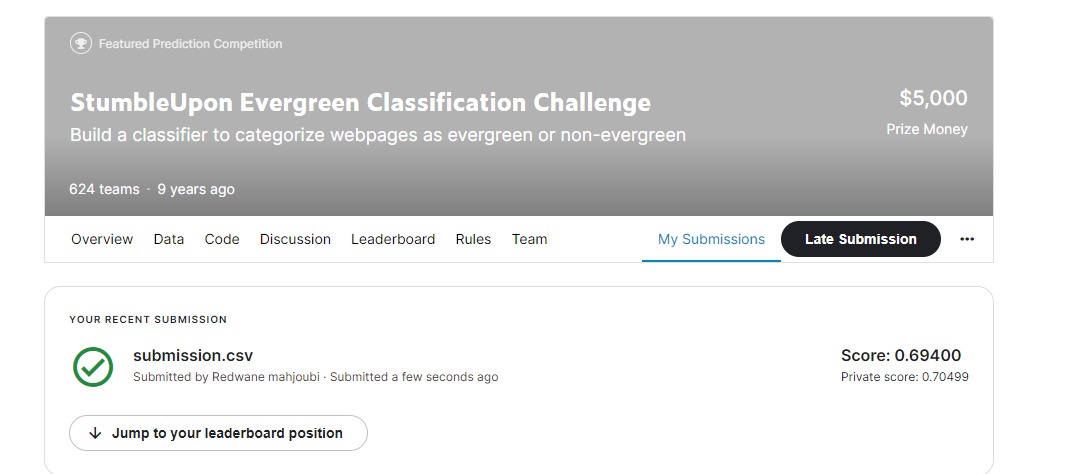
# Description de la compétition :

StumbleUpon est un moteur de découverte de contenu Web organisé par les utilisateurs qui recommande des pages et des médias pertinents et de haute qualité à ses utilisateurs, en fonction de leurs intérêts. Alors que certaines pages que nous recommandons, telles que les articles d'actualité ou les recettes de saison, ne sont pertinentes que pendant une courte période, d'autres conservent une qualité intemporelle et peuvent être recommandées aux utilisateurs longtemps après leur découverte. En d'autres termes, les pages peuvent être classées comme "éphémères" ou "persistantes". Les notes que nous obtenons de notre communauté nous donnent des signaux forts indiquant qu'une page n'est peut-être plus pertinente - mais que se passerait-il si nous pouvions faire cette distinction à l'avance ? Une prédiction de haute qualité de "éphémère" ou "persistant" améliorerait grandement un système de recommandation comme le nôtre.

Beaucoup de gens connaissent le contenu à feuilles persistantes lorsqu'ils le voient, mais un algorithme peut-il faire la même détermination sans l'intuition humaine ? Votre mission est de créer un classificateur qui évaluera un grand nombre d'URL et les étiquettera comme persistantes ou éphémères. Pouvez-vous surclasser StumbleUpon ? Comme incitation supplémentaire au prix, une solide performance dans ce concours peut conduire à un stage de lancement de carrière dans l'un des meilleurs endroits où travailler à San Francisco.



Figure2 StumbleUpon.

 Figure 2 participation dans la compétition

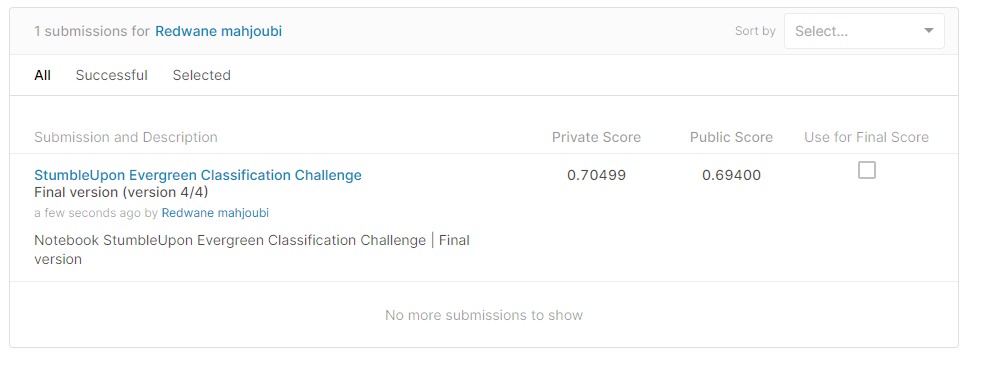
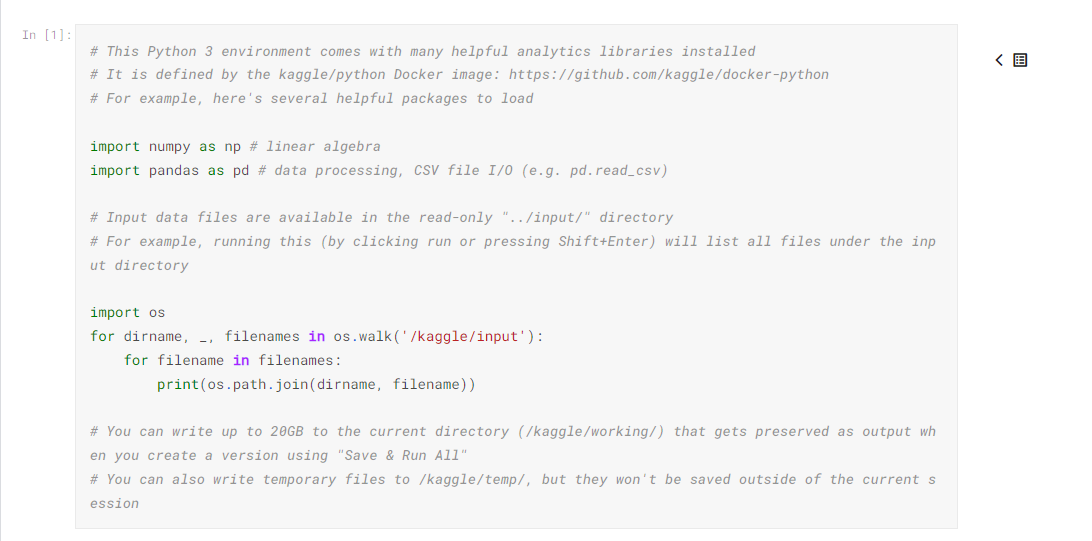


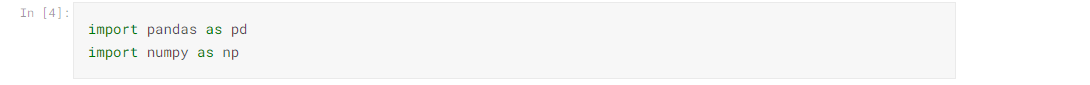
Figure3 score prive et public de notre participation

# Explication de code source :

Importation des données :

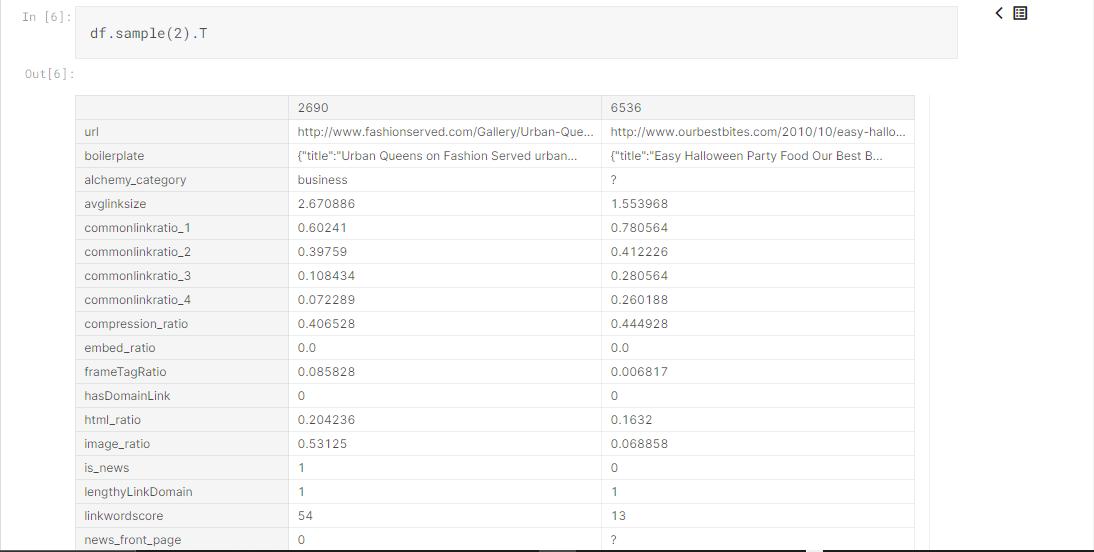




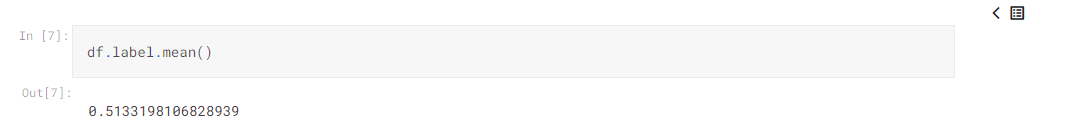




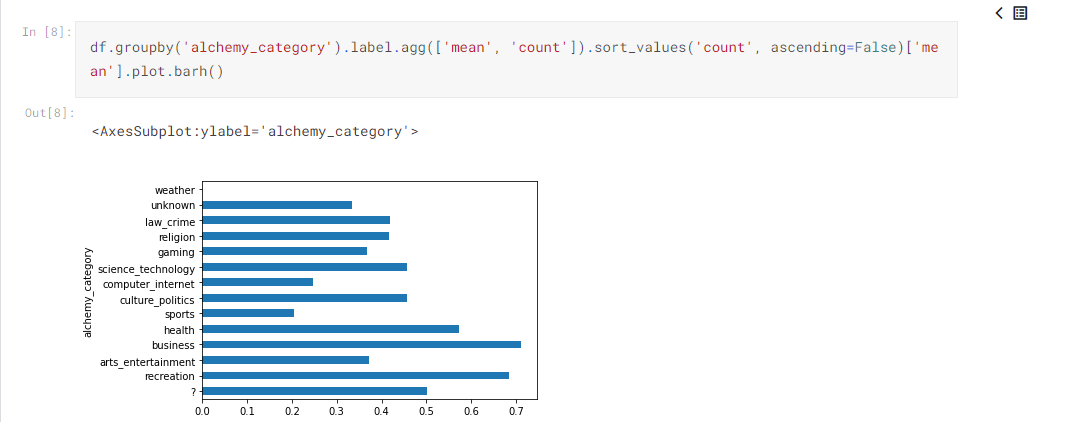
Affichage deux ligne de manière transposer



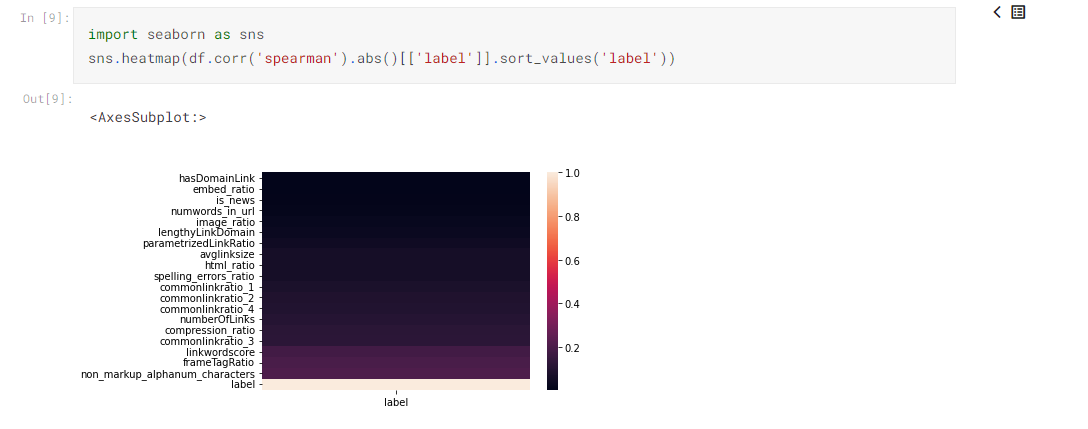
Affichage de la moyenne



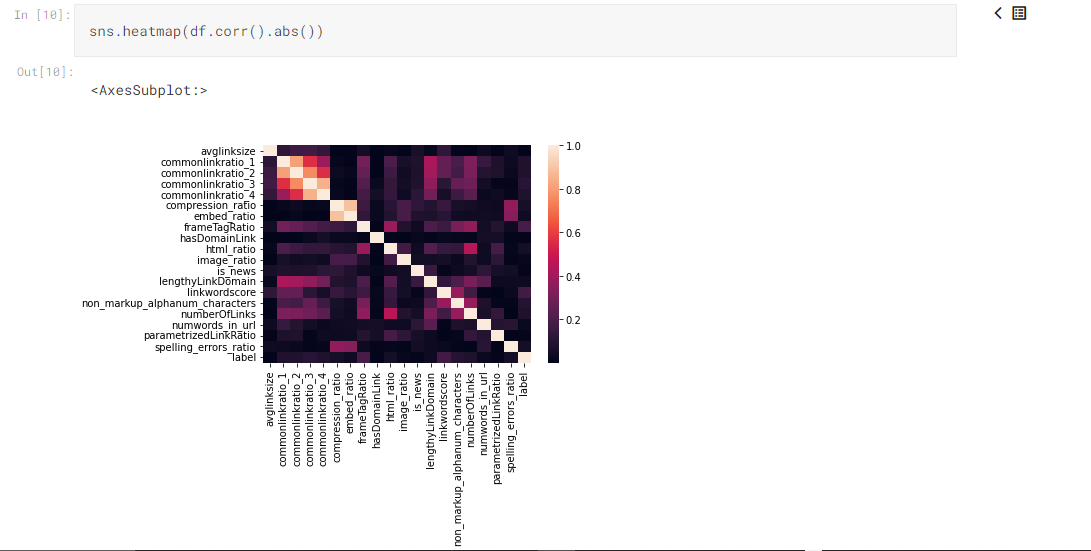
Designer un plot dont se trouve la moyenne de label pour chaque categorie



La visualisation de corrélation entre les features et label se forme de graph avec couleurs



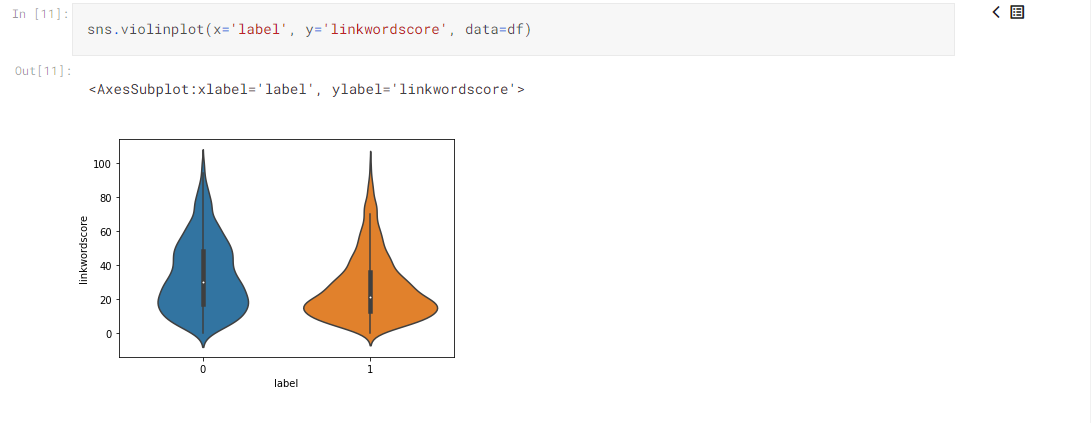
La visualisation de corrélation entre les features se forme de graph avec couleurs



On a devisé data en deux :

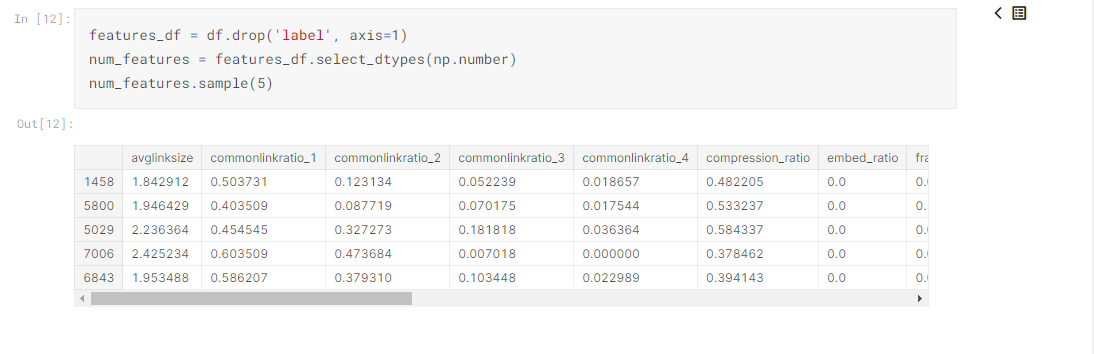
0 -> les données perdre savalibilité apres le temps

1-> les données reste toujours valable

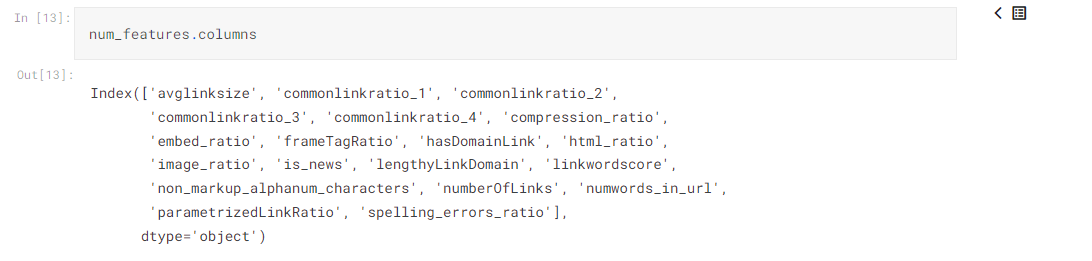


La suppression de données qui sont corrélés 100%

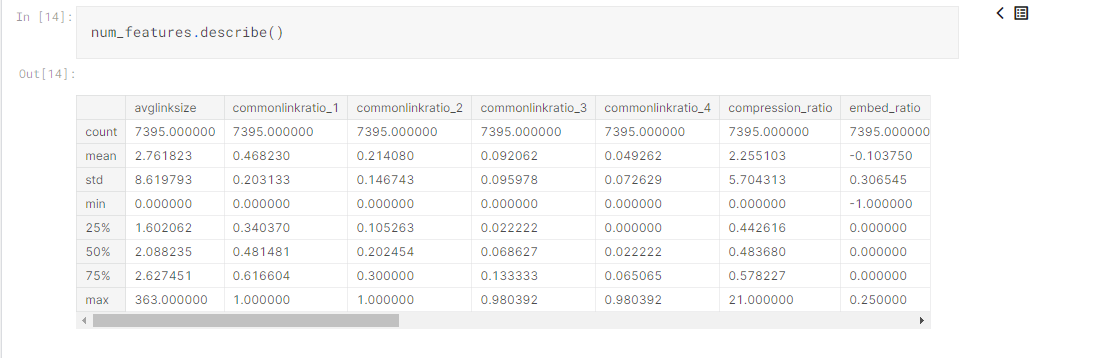
Et afficher les cinq premières lignes



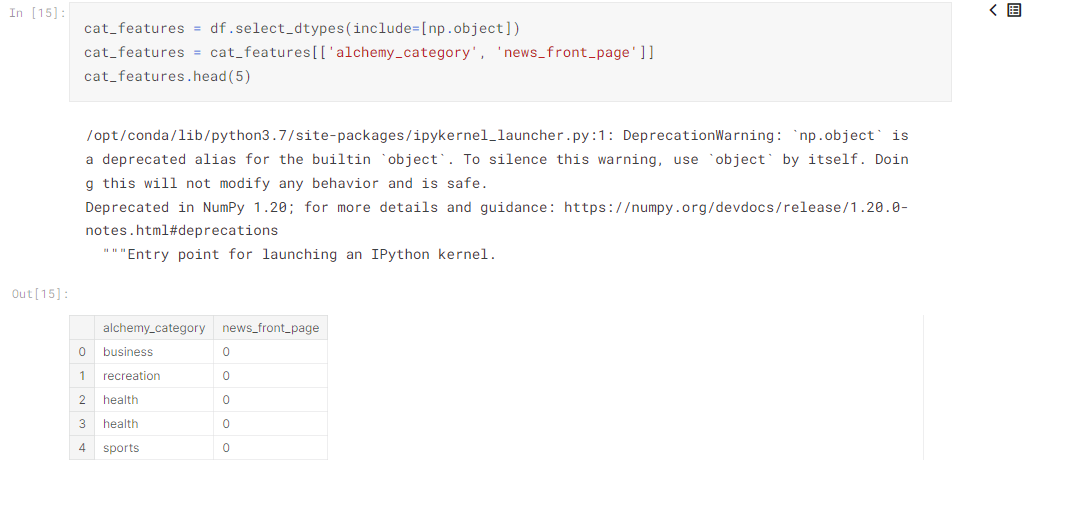
Afficher les noms des features

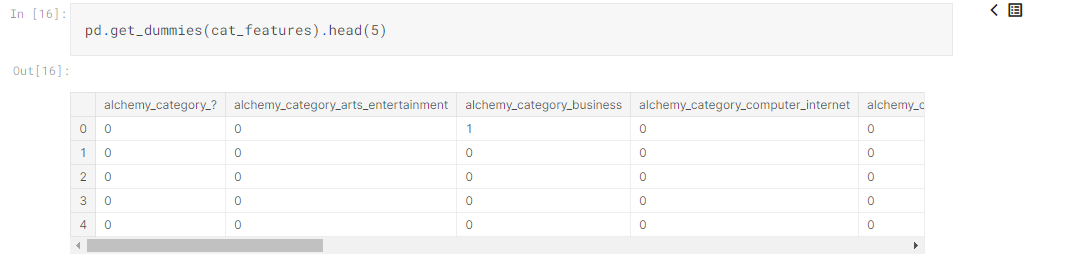


Afficher la description des features



Sélectionner deux features et afficher les cinq premiers lignes

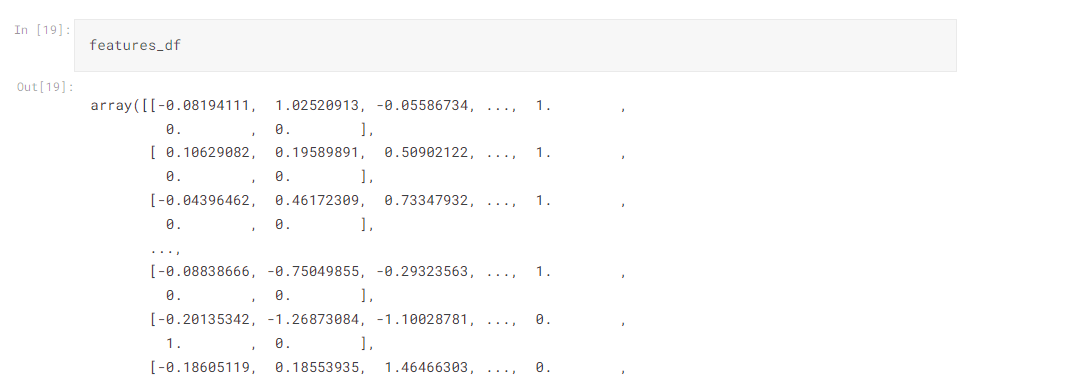


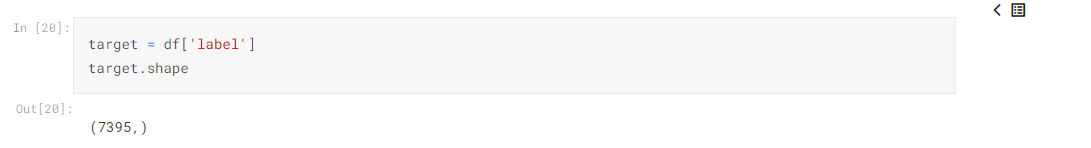


Préparation de données pour l’entrainer









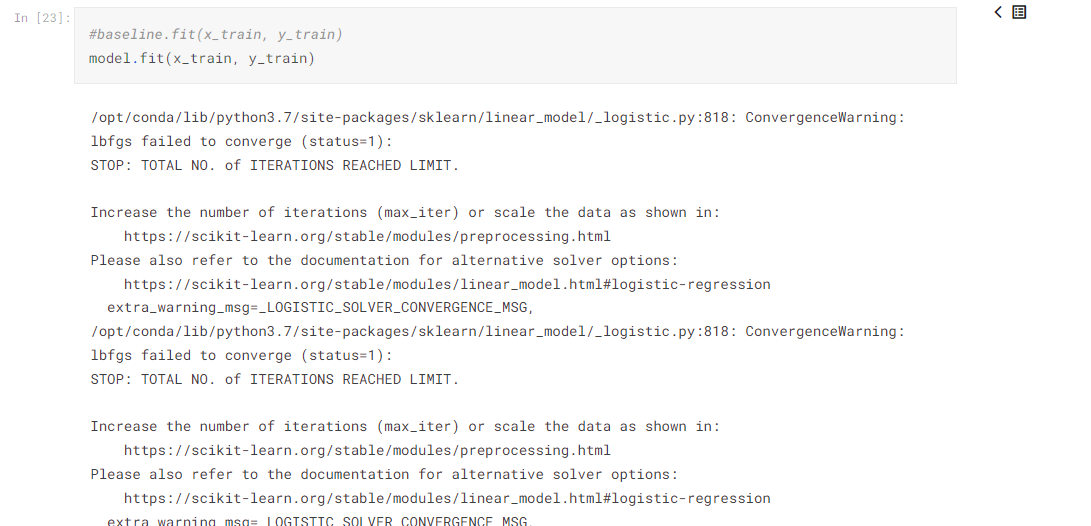
Deviser le Dataset en deux données de test et d’entrainement



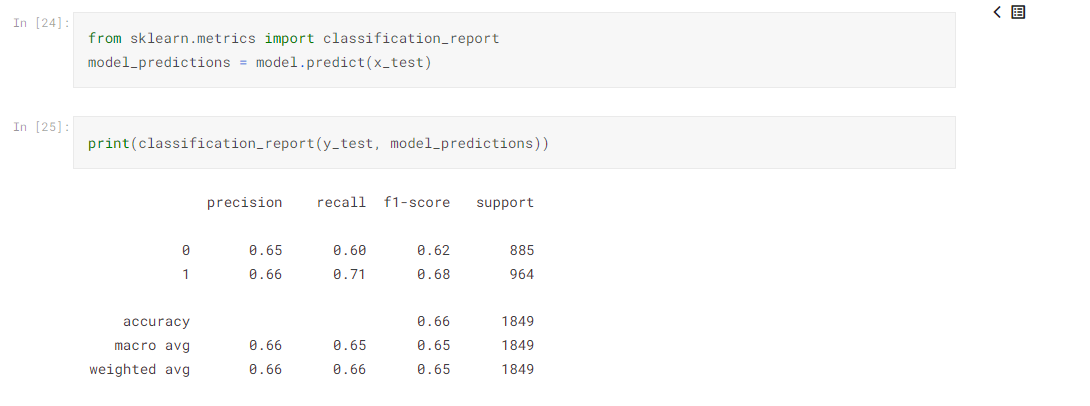
Créer le model



Appliquer le model sur les données d’entrainement



Afficher le rapport sur notre model







# Conclusion :

L'apprentissage automatique peut être supervisé ou non supervisé. Si vous avez moins de données et des données clairement étiquetées pour la formation, optez pour l'apprentissage supervisé. L'apprentissage non supervisé donnerait généralement de meilleures performances et de meilleurs résultats pour les grands ensembles de données.