Master Économétrie et Statistiques, parcours Économétrie Appliquée

Machine Learning sous Python
SVM et réseaux de neurones

MEZIANI Sara

MORTREUIL Louis

Année Universitaire : 2021/2022

Résumé

L’objectif visé à travers notre étude est d'analyser et d’explorer l’état d’apprentissage numérique aux Etats Unis durant la COVID 19, spécifiquement durant l'année 2020. Pour cela, nous avons réalisé des statistiques descriptives des trois bases de données mise à disposition par les organisateurs de cette compétition kaggle.Pour déterminer les facteurs explicatifs de la variable 'engagement-district' nous avons réalisé trois modélisations: arbre de régression, réseau de neurone MLP et une régression OLS. Les résultats obtenus sont proches et ne diffèrent pas d’une méthode à une autre et ils nous ont permis de tirer pratiquement les mêmes constats notamment en ce qui concerne la qualité d’ajustement sur le jeu de données ‘train’.

Abstract

The objective of our study is to analyze and explore the state of digital learning in the United States during COVID 19, specifically during the year 2020. To do so, we performed descriptive statistics on three databases provided by the organizers of this kaggle competition. To determine the explanatory factors of the variable 'engagement-district' we performed three models: regression tree, MLP neural network and an OLS regression. The results obtained are similar and do not differ from one method to another and they allowed us to draw practically the same conclusions, especially concerning the goodness of fit on the 'train' data set.

SOMMAIRE

1 PRESENTATION DE L'ETUDE

Depuis la fin de 2019, le monde ne fonctionne plus comme avant en raison de la pandémie du Covid 19 qui a bouleversé les systèmes économiques, sociaux, éducatifs, etc. L'un des principaux impacts est celui des étudiants du monde entier. Ils ne peuvent pas aller à l'école pour apprendre en face à face, ce qui a un impact important sur leur vie sociale. En effet, selon la Banque Mondiale, près de **1,5 milliard** d'élèves dans plus de 170 pays ne vont plus à l'école, leurs établissements ayant été fermés par les gouvernements pour lutter contre la propagation du virus. Dans ce contexte, les ministères de l'éducation du monde entier tentent désormais d'assurer la continuité des apprentissages par le biais de l'enseignement à distance. Dans la plupart des cas, cela implique l'utilisation de plateformes numériques et d'outils technologiques dédiés à l'éducation dans le but de rendre les espaces d'apprentissage aussi ouverts et stimulants que possible.

Aux Etats Unis , la pandémie de COVID-19 a perturbé l'apprentissage de plus de 56 millions d'étudiants. Jusqu'à aujourd'hui, les inquiétudes concernant la fracture numérique exacerbée et la perte d'apprentissage à long terme parmi les apprenants les plus vulnérables d'Amérique continuent de croître.

L'objectif de notre étude est de participer à la compétition Kaggle " LearnPlateform Covid-19 Impact On Digital Learning " pour analyser et déterminer l'impact de la Covid 19 sur l'apprentissage des étudiants tout en utilisant les données numériques aux USA et présenter notre démarche et nos résultats obtenus.

Pour répondre à cette problématique, nous allons utiliser dans ce travail plusieurs méthodes de machine learning. Les données que nous allons exploiter sont présentées en trois fichiers de données fournis par les organisateurs de la compétition et qui sont :

- **Les données engagement**: sont basées sur l'extension Student Chrome de LearnPlatform. L'extension collecte les événements de chargement de page de plus de 10 000 produits de technologie éducative dans notre bibliothèque de produits, notamment des sites Web, des applications, des applications Web, des logiciels, des extensions, des livres électroniques, des matériels et des services utilisés dans les

établissements d'enseignement. Les données d'engagement ont été agrégées au niveau du district scolaire, et chaque fichier représente les données d'un district scolaire.

- **Les données products**: comprend des informations sur les caractéristiques des 372 meilleurs produits avec la plupart des utilisateurs en 2020.
- **Les données districts**: comprend des informations sur les caractéristiques des districts scolaires, y compris des données du NCES et de la FCC.

2 L'État de l'apprentissage numérique en 2020 (Analyse exploratoire)

Dans cette partie nous allons présenter quelques statistiques descriptives des trois bases de données que nous possédons et explorer les informations qu'elles contiennent pour déterminer l'état de l'apprentissage numérique aux USA en 2020 suite à la pandémie du Covid 19.

2.1 La base de données District

Après avoir importé le fichier csv représentant la base de données "district" à l'aide de la librairie "***Pandas***", nous affichons les premières lignes de cette dernière en appelant la commande "head" pour s'assurer qu'elle est bien importée.

Figure 1: Aperçu de la base de données district

	district_id	state	locale	pct_black/hispanic	pct_free/reduced	county_connections_ratio	pp_total_raw
0	8815	Illinois	Suburb	[0, 0.2[	[0, 0.2[	[0.18, 1[	[14000, 16000[
1	2685	NaN	NaN	NaN	NaN	NaN	NaN
2	4921	Utah	Suburb	[0, 0.2[	[0.2, 0.4[	[0.18, 1[	[6000, 8000[
3	3188	NaN	NaN	NaN	NaN	NaN	NaN
4	2238	NaN	NaN	NaN	NaN	NaN	NaN

Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

A partir de la figure ci-dessus, on constate que la base de données est bien importée et chaque chaque donnée est composée des variables suivantes :

district_id : L'identifiant unique du district scolaire

state : L'état dans lequel se trouve le district

locale : Classification locale du NCES qui classe le territoire américain en quatre types de zones : Ville, banlieue, ville et zone rurale.

pct_black/hispanic : Pourcentage d'élèves des districts identifiés comme étant noirs ou hispaniques sur la base des données 2018-19 du NCES.

pct_free/reduced : Pourcentage d'élèves des districts éligibles pour un déjeuner gratuit ou à prix réduit sur la base des données 2018-19 du NCES.

countyconnectionsratio : ratio (connexions haut débit fixes résidentielles de plus de 200 kbps dans au moins une direction/ménages) basé sur les données au niveau du comté provenant de FCC From 477 (version de décembre 2018). Voir les données de la FCC pour plus d'informations.

pptotalraw : Dépenses totales par élève (somme des dépenses locales et fédérales) provenant du projet Edunomics Lab's National Education Resource Database on Schools (NERD$). Les données relatives aux dépenses sont établies école par école, et nous utilisons la valeur médiane pour représenter les dépenses d'un district scolaire donné.

Maintenant, pour visualiser les principales statistiques descriptives de l'ensemble des variables , on exécute la commande "***discribe***" et on obtient le tableau ci-dessous:

Tableau 1: Statistiques descriptives

	district_id	state	locale	pct_black/hispanic	pct_free/reduced	county_connections_ratio	pp_total_raw
count	233.000000	176	176	176	148	162	118
unique	NaN	23	4	5	5	2	11
top	NaN	Connecticut	Suburb	[0, 0.2[	[0.2, 0.4[	[0.18, 1[	[8000, 10000[
freq	NaN	30	104	116	48	161	30
mean	5219.776824	NaN	NaN	NaN	NaN	NaN	NaN
std	2595.751581	NaN	NaN	NaN	NaN	NaN	NaN
min	1000.000000	NaN	NaN	NaN	NaN	NaN	NaN
25%	2991.000000	NaN	NaN	NaN	NaN	NaN	NaN
50%	4937.000000	NaN	NaN	NaN	NaN	NaN	NaN
75%	7660.000000	NaN	NaN	NaN	NaN	NaN	NaN
max	9927.000000	NaN	NaN	NaN	NaN	NaN	NaN

Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

A partir des données du district, nous avons 233 lignes avec 7 colonnes. Ainsi, la base de données "district" contient plusieurs valeurs manquantes marquées par 'NAN' et ceci est dû à la suppression des données pour maximiser l'anonymat de ces dernières.

On constate que la variable "district-id" varie entre 1000 et 9927. Ces nombres représentent des identifiants et non des valeurs numériques.

Nous pouvons aussi , distinguer le nombre de valeurs manquantes dans chaque colonne :

Figure 2 : Le nombre de valeurs manquantes dans chaque colonne

```
Number of missing Values in every column:
district_id                  0
state                       57
locale                      57
pct_black/hispanic          57
pct_free/reduced            85
county_connections_ratio    71
pp_total_raw               115
dtype: int64
```

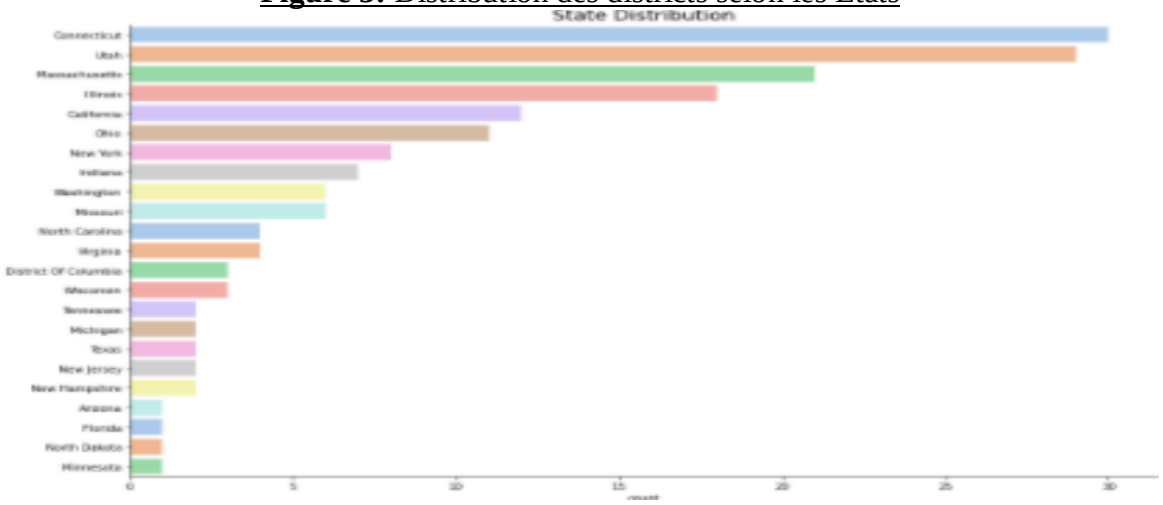
Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

On constate que toutes les variables de la base district contiennent des valeurs manquantes à l'exception de la variable "district". La variable qui a le plus de valeurs manquantes est "les dépenses totales par élève". Ces valeurs manquantes , nous devons les nettoyer dans la phase de la modélisation , pour obtenir des résultats plus robustes.

Maintenant, nous allons faire des croisements des variables existantes dans la base "district" pour visualiser le maximum d'informations possibles. Le code utilisé sous python est rpicnipaement issu de la libraire "***Matplotlib***"

Dans ce qui va suivre nous allons représenter les répartitions des districts selon les Etats où ils se trouvent. Les résultats sont présentées dans la figure ci-dessous:

Figure 3: Distribution des districts selon les Etats

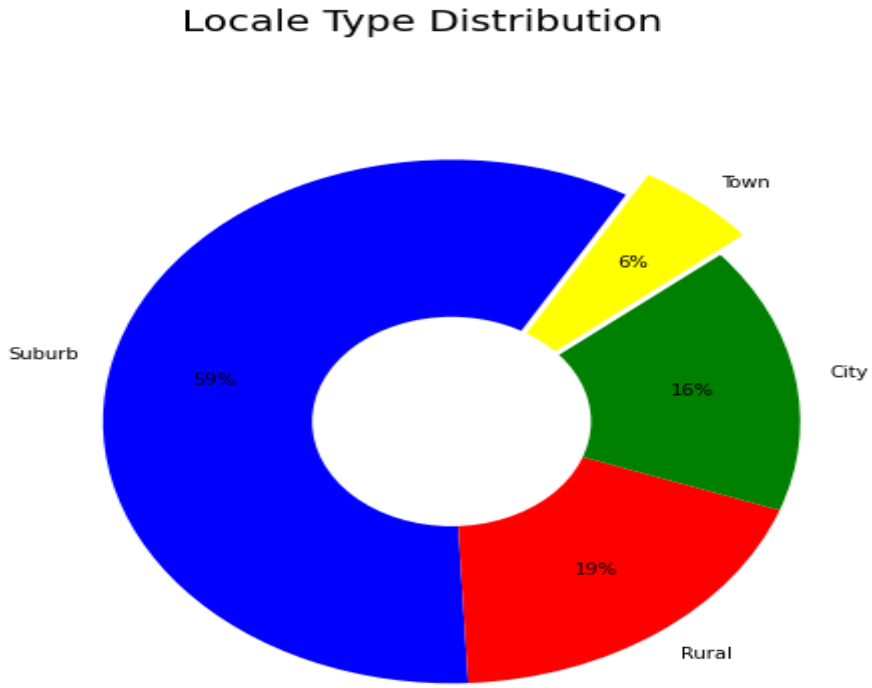


Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

D'après la figure ci-dessus, on constate que les Etats qui ont le plus de districts, autrement dit la fréquence des districts est plus grande dans l'Etat de Connecticut, suivi par les Etats: Utah, Massachusetts et Illinois. Les États qui contiennent le moins de districts sont : Minnesota, Florida, North Dakota et Arizona.

Nous pouvons aussi présenter la répartition des districts selon les zones géographiques:

Figure 4: La répartition des districts selon la zone géographique

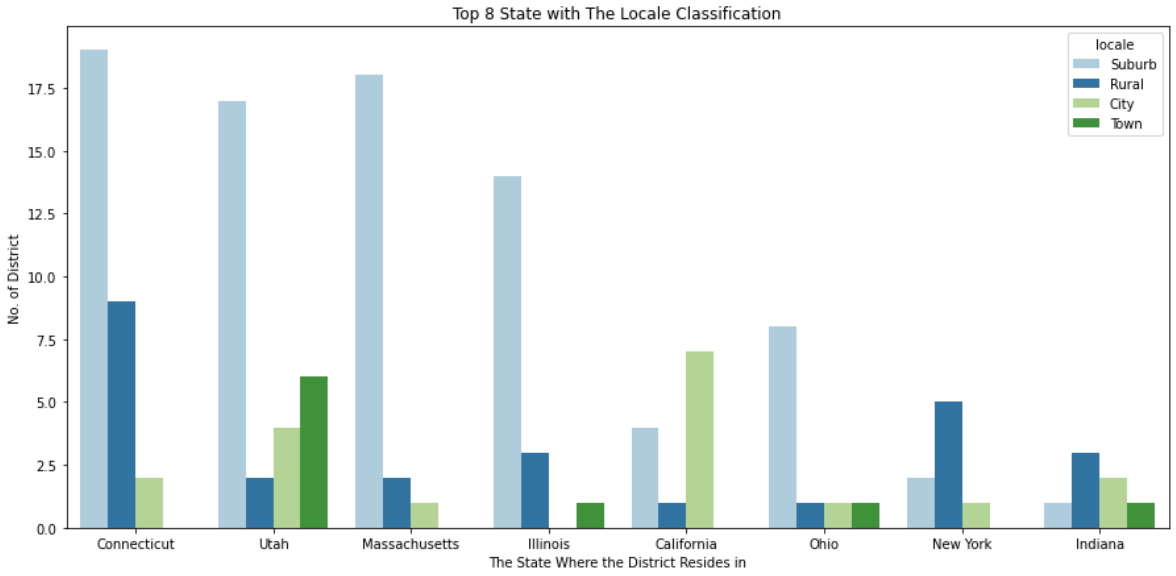


Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

La figure n°4 , montre que la région suburbaine est la plus représentée et continent le plus grand nombre de districts scolaires, avec 59%, suivi par la région rurale avec 19%. Les villes sont les moins représentées avec 6%.

A partir des deux représentations précédentes, nous pouvons également représenter la répartition des districts scolaires en fonction de la zone géographique et de l'Etat auquel ils appartiennent. Nous avons choisi de nous arrêter au 8 premiers Etats.

Figure 5: Les 8 premiers Etats en fonction de la zone géographique

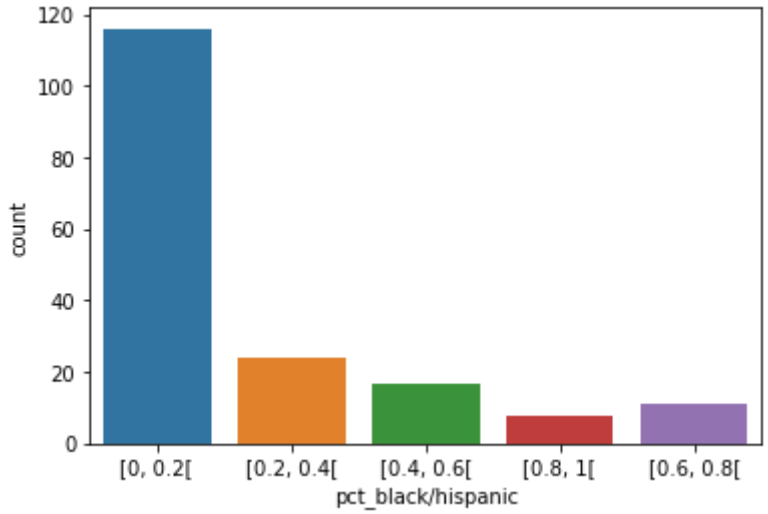


Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

On constate que pour tous les Etats , c'est la zone suburbaine qui contient le plus grand nombre de districts scolaires; sauf les Etats de New York et Indiana qui ont la zone rurale comme dominante. Ce qui confirme le constat de la figure n°4. On peut remarquer que, mise à part dans l'Etat de l'Utah, la catégorie n'est presque pas représentée. Cela confirme les observations faites précédemment où la localisation town ne représente que 6% des districts.

Nous nous sommes ensuite intéressés au pourcentage de la population afro-américaine et hispanoaméricaine dans les districts. En effet, ces communautés proviennent régulièrement de milieux défavorisés, il est donc probable que leur accès à l'apprentissage numérique soit plus compliqué.

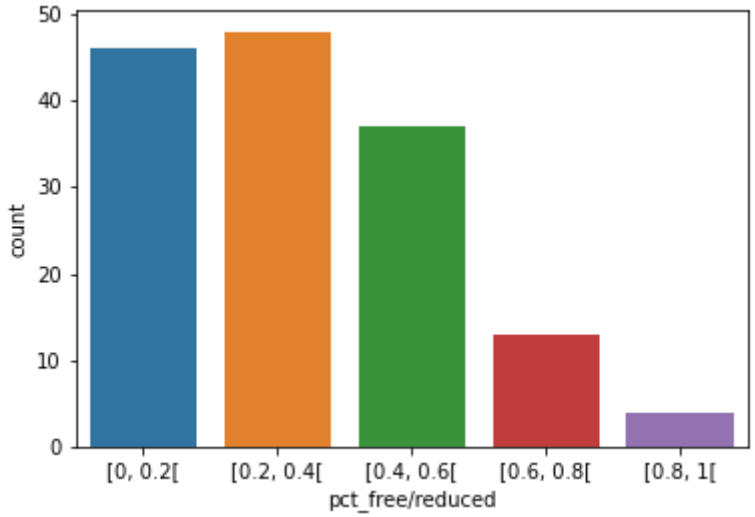
Figure 6: Représentation en pourcentage des étudiants afro ou hispanique dans les différents districts des Etats-Unis



Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

Dans la majorité des districts présents aux Etats-Unis, on peut voir que le pourcentage des étudiants afroaméricains ou hispanoaméricains représente moins de **20%** des étudiants du district. Mais dans certains districts, on remarque que ces étudiants d'origine africaine ou d'Amérique latine peuvent représenter une part importante des étudiants.

Figure 7: Représentation en pourcentage des étudiants ayant droit à des réductions ou des repas gratuits à l'école dans les différents districts des Etats-Unis



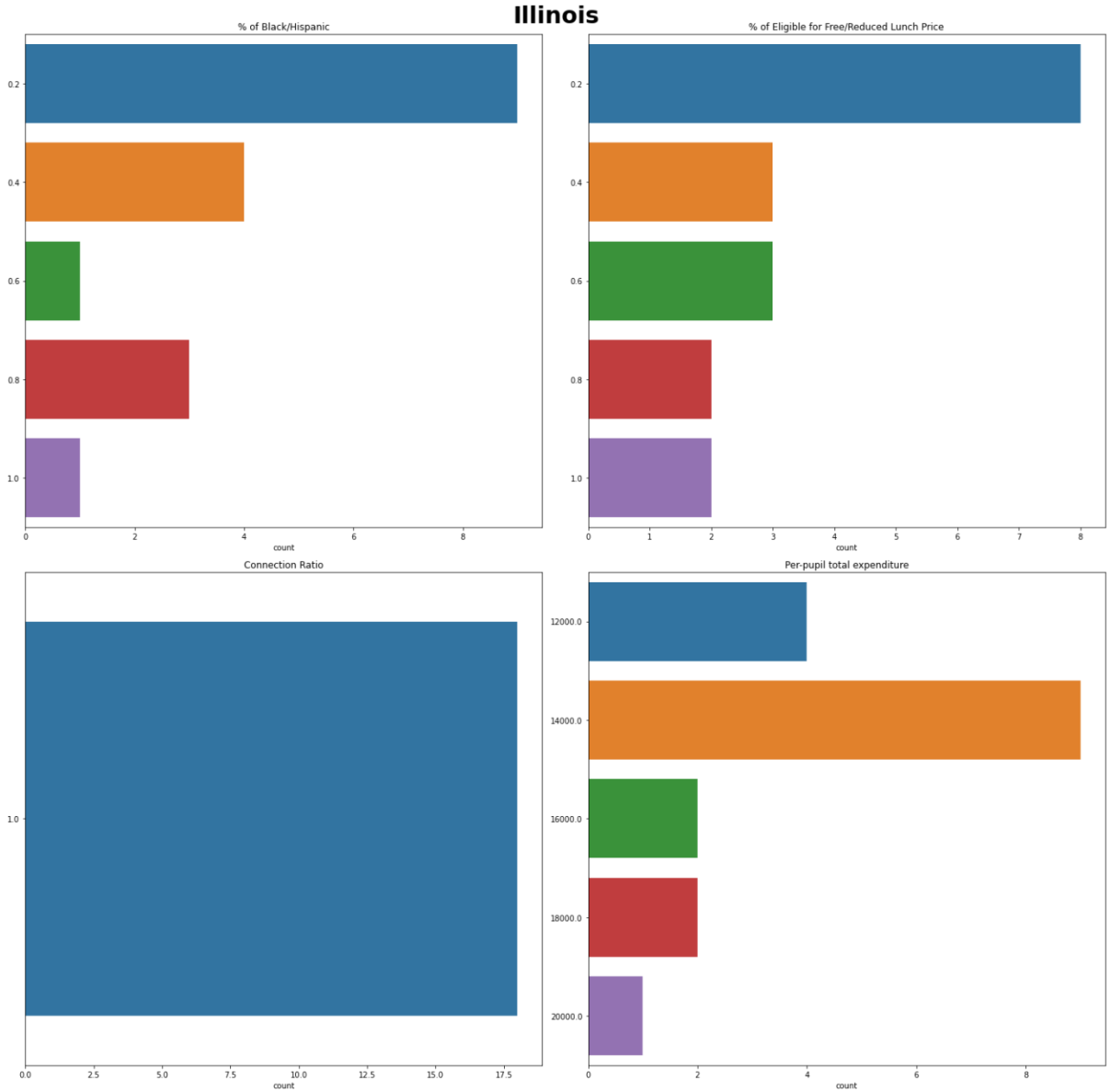
Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

On peut remarquer que peu de districts se trouvent dans les deux dernières catégories, ainsi le pourcentage des étudiants ayant accès à des repas gratuits ou des réductions est plus ou moins important dans les différents districts.

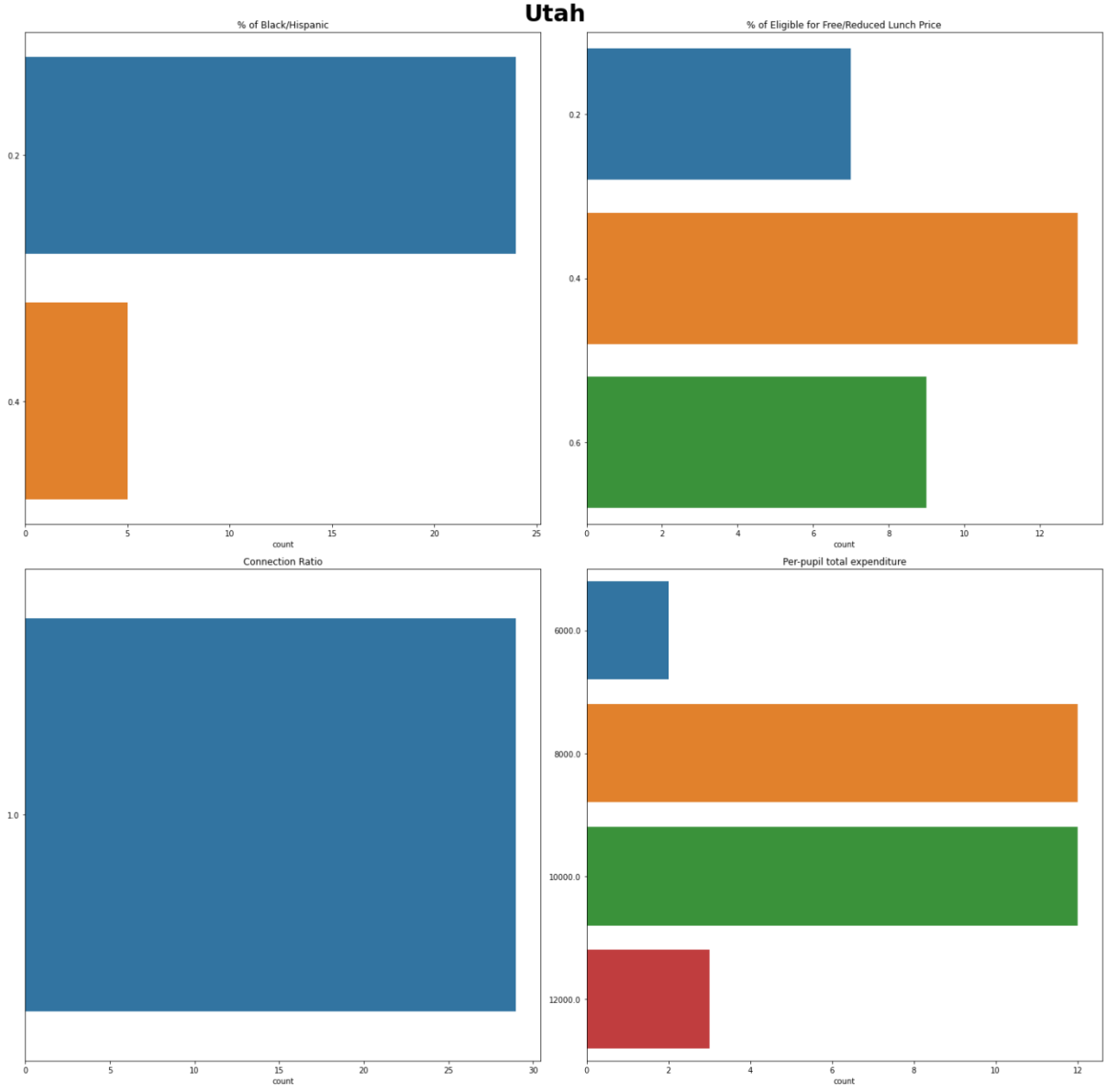
Il est possible d'observer certaines données de districts plus précisément par état que sur l'ensemble du pays. Nous avons choisi de le faire sur les 4 États choisis aléatoirement.

Les résultats sont résumés ci-après:

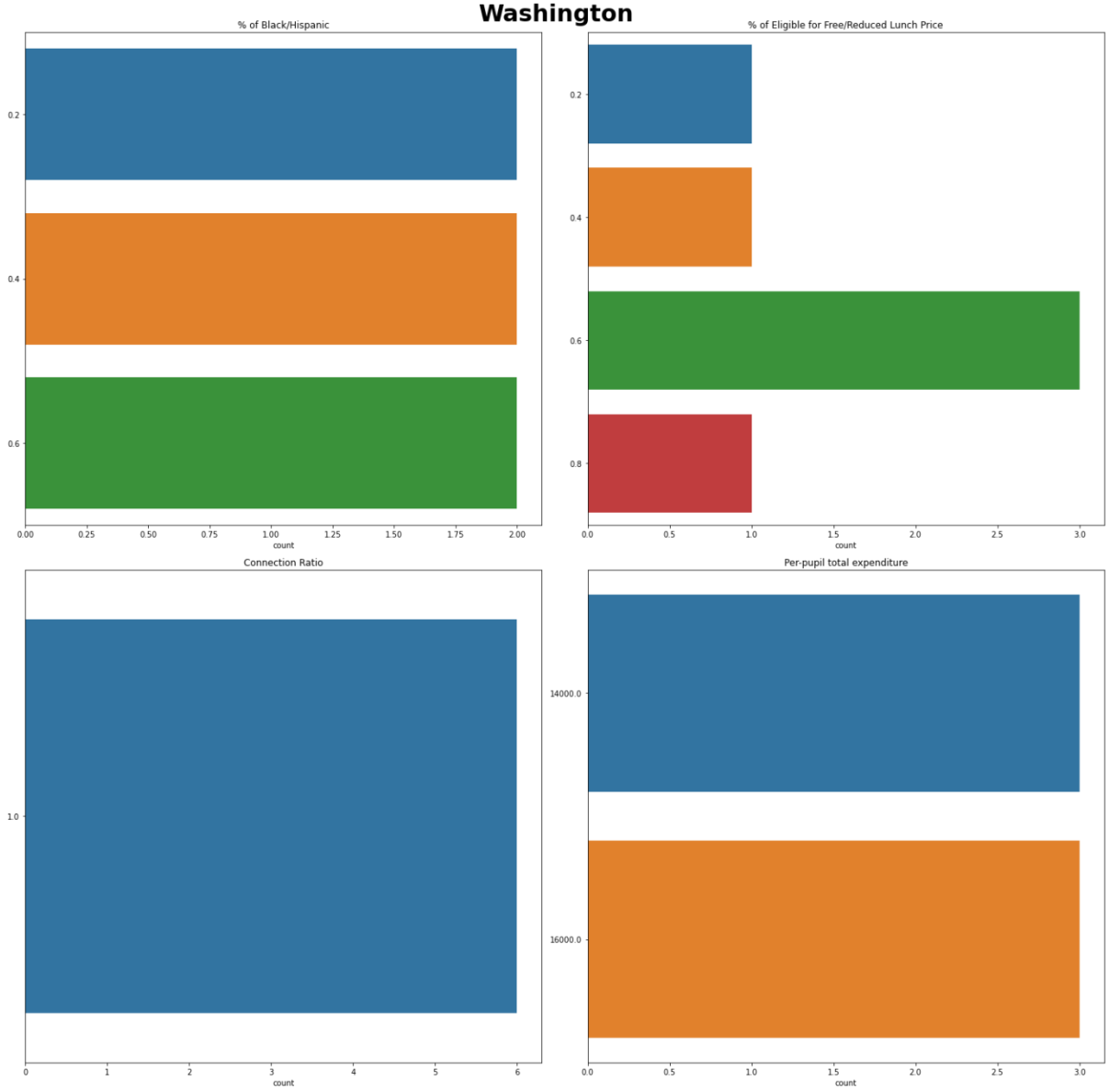
Figure 8: Quelques statistiques descriptives sur 4 Etats des Etats-Unis

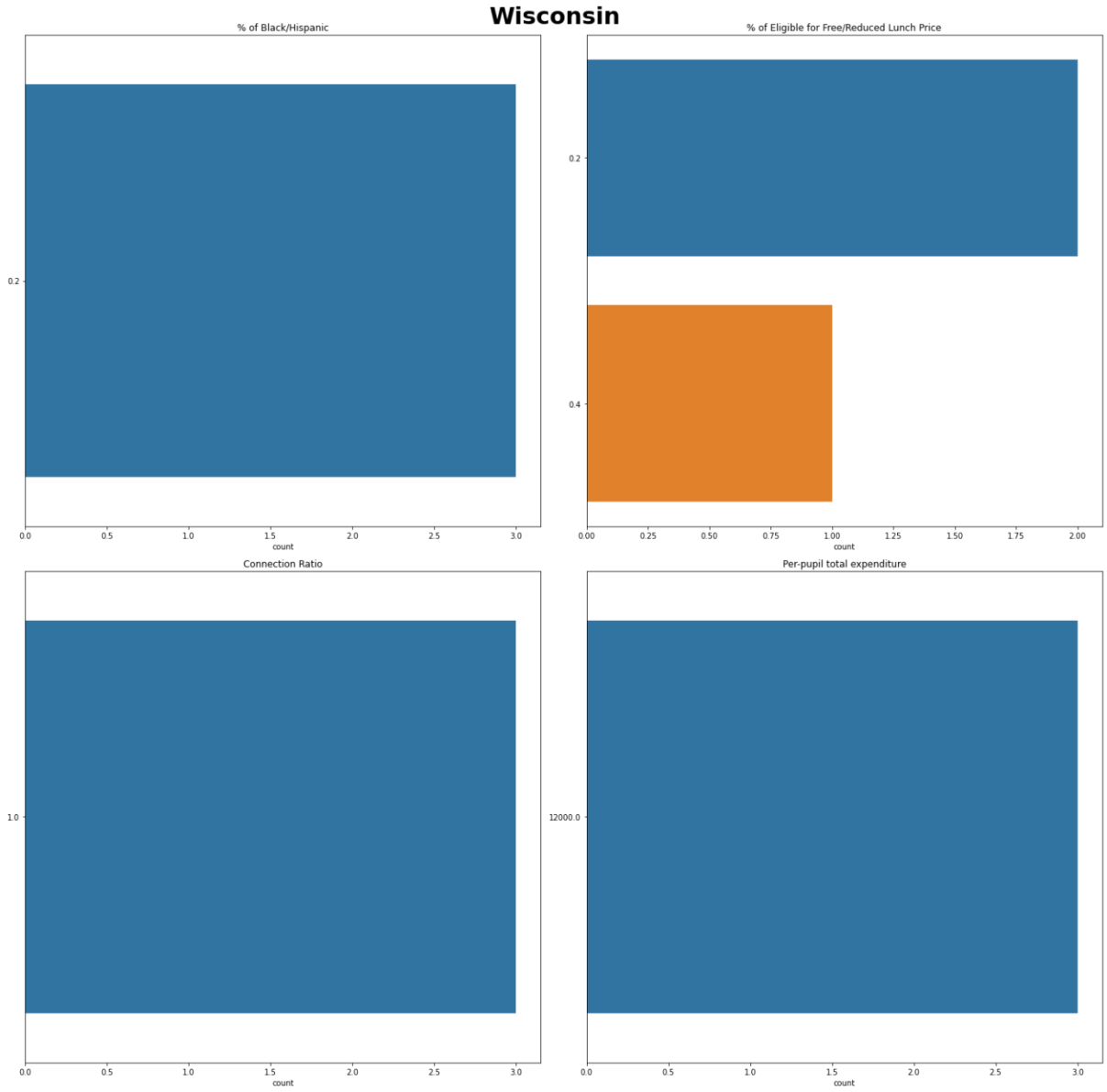


Utah


% of Black/Hispanic
0.2
0.4
0 5 10 15 20 25
count
% of Eligible for Free/Reduced Lunch Price
0.2
0.4
0.6
0 2 4 6 8 10 12
count
Connection Ratio
1.0
0 5 10 15 20 25 30
count
Per-pupil total expenditure
6000.0
8000.0
10000.0
12000.0
0 2 4 6 8 10 12
count

Washington


% of Black/Hispanic
0.2
0.4
0.6
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
count
% of Eligible for Free/Reduced Lunch Price
0.2
0.4
0.6
0.8
0.0
0.5
1.0
1.5
2.0
2.5
3.0
count
Connection Ratio
1.0
0
1
2
3
4
5
6
count
Per-pupil total expenditure
14000.0
16000.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
count



Pour l'Etat de l'Illinois, on remarque que la majorité des districts se comportent moins de 40% de population afro-américaine ou hispanoaméricaine et que la majorité des districts ont un pourcentage d'étudiants ayant droit à des repas gratuits en dessous de 40%. Pour l'Utah, la population d'afro-américain ou d'hispanoaméricain ne dépasse pas les 40% dans aucun district et le pourcentage d'étudiants ayant accès à un repas gratuit ne dépasse pas les 60% pour tous ses districts. En ce qui concerne Washington, le nombre de districts dans les catégories moins de 20%, entre 20% et 40% et entre 40% et 60% de population afro-américaine et hispanoaméricaine sont égaux. La majorité des districts se situent dans la catégorie [40%, 60%[des étudiants qui ont accès à un repas gratuit. Pour finir, l'Etat du Wisconsin possède un pourcentage de moins de 20% de population afro-américaine ou

hispanoaméricaine dans la totalité de ses districts qui sont au nombre de trois. Un se situe dans la catégorie [20%, 40%[des étudiants ont le droit à un repas gratuit, les deux autres se situent dans la catégorie moins de 20%. Grâce à ces graphiques, on remarque que la population des étudiants est différente d'un état à un autre et que la situation n'est pas homogène sur l'ensemble du territoire américain. On peut noter que le nombre de districts par était n'est pas toujours équivalent.

2.2 La base de données "Product"

Nous allons partir du même principe que pour la base "district", où nous allons analyser cette base de données en s'appuyant sur les statistiques descriptives pour essayer de tirer le maximum d'informations essentielles à la compréhension de l'état de l'apprentissage numérique aux USA.

La commande "head" , nous fournit le résultat suivant:

Figure 9: Aperçu de la base de données product

	LP ID	URL	Product Name	Provider/Company Name	Sector(s)	Primary Essential Function
0	13117	https://www.splashmath.com	SplashLearn	StudyPad Inc.	PreK-12	LC - Digital Learning Platforms
1	66933	https://abcmouse.com	ABCmouse.com	Age of Learning, Inc	PreK-12	LC - Digital Learning Platforms
2	50479	https://www.abcya.com	ABCya!	ABCya.com, LLC	PreK-12	LC - Sites, Resources & Reference - Games & Si...
3	92993	http://www.aleks.com/	ALEKS	McGraw-Hill PreK-12	PreK-12; Higher Ed	LC - Digital Learning Platforms
4	73104	https://www.achieve3000.com/	Achieve3000	Achieve3000	PreK-12	LC - Digital Learning Platforms

Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

A partir des données produits, nous 6 colonnes présentant les variables qui sont:

LP ID : L'identifiant unique du produit

URL : Lien Web vers le produit spécifique

Nom du produit : Nom du produit spécifique

Nom du fournisseur/de l'entreprise : Nom du fournisseur du produit

Secteur(s) : Secteur de l'éducation où le produit est utilisé

Fonction essentielle primaire : La fonction de base du produit. Il y a deux couches d'étiquettes ici. Les produits sont d'abord étiquetés dans l'une de ces trois catégories : LC = Learning & Curriculum, CM = Classroom Management, et SDO = School & District Operations. Chacune

de ces catégories comporte plusieurs sous-catégories avec lesquelles les produits ont été étiquetés.

Le tableau des statistiques descriptives de la base de données “Product” est présenté ci-dessous:

Tableau 2 : Statistiques descriptives

	LP ID	URL	Product Name	Provider/Company Name	Sector(s)	Primary Essential Function
count	372.000000	372	372	371	352	352
unique	NaN	372	372	290	5	35
top	NaN	https://artsandculture.google.com/	Math Playground	Google LLC	PreK-12	LC - Digital Learning Platforms
freq	NaN	1	1	30	170	74
mean	54565.795699	NaN	NaN	NaN	NaN	NaN
std	26247.551437	NaN	NaN	NaN	NaN	NaN
min	10533.000000	NaN	NaN	NaN	NaN	NaN
25%	30451.000000	NaN	NaN	NaN	NaN	NaN
50%	53942.500000	NaN	NaN	NaN	NaN	NaN
75%	77497.000000	NaN	NaN	NaN	NaN	NaN
max	99916.000000	NaN	NaN	NaN	NaN	NaN

Source: MEZIANI.S et MORTEUIL.L à l’aide du logiciel Python

Pour les données Product, nous avons 372 lignes avec 6 colonnes. Cette base de données contient également des valeurs manquantes en raison de la suppression des données pour maximiser l’anonymat de ces dernières.

L’identifiant unique du produit a une étendue assez importante avec un minimum de 10533 et un maximum de 99916.

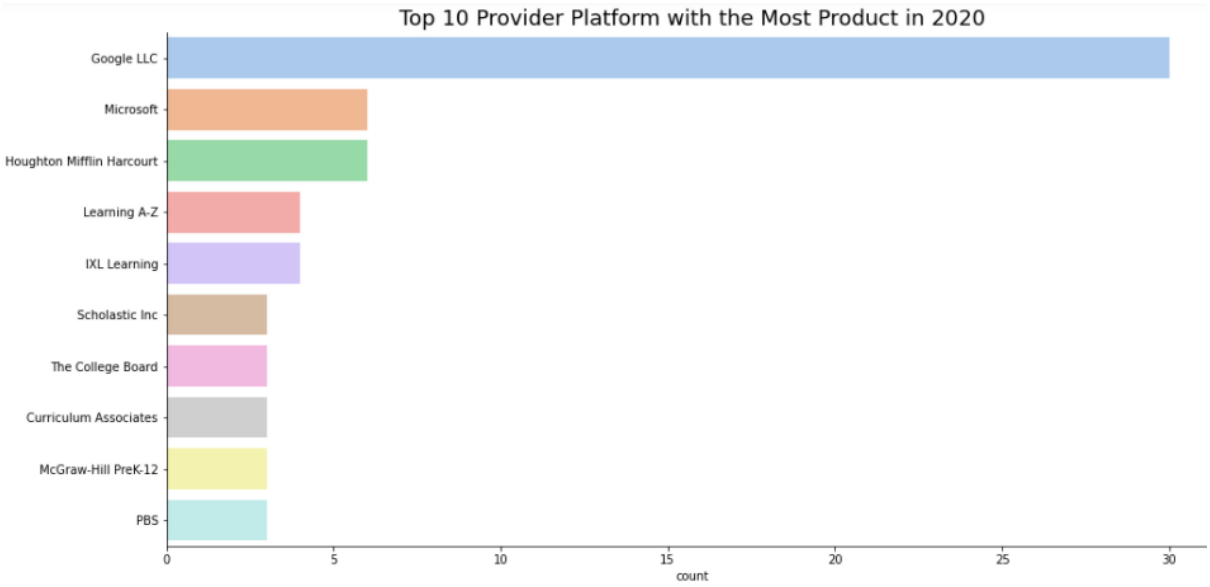
Nous pouvons aussi vérifier les valeurs manquantes dans chaque colonne:

```
Number of missing Values in every column:
LP ID                          0
URL                            0
Product Name                   0
Provider/Company Name          1
Sector(s)                     20
Primary Essential Function    20
dtype: int64
```

On constate que les trois premières variables ne contiennent pas de valeurs manquantes et les trois dernières ont un nombre réduit de valeurs manquantes.

Nous pouvons représenter graphiquement le Top 10 des fournisseurs de plateformes ayant le plus de produits en 2020.

Figure 10: Top 10 des fournisseurs de plateformes ayant le plus de produits en 2020

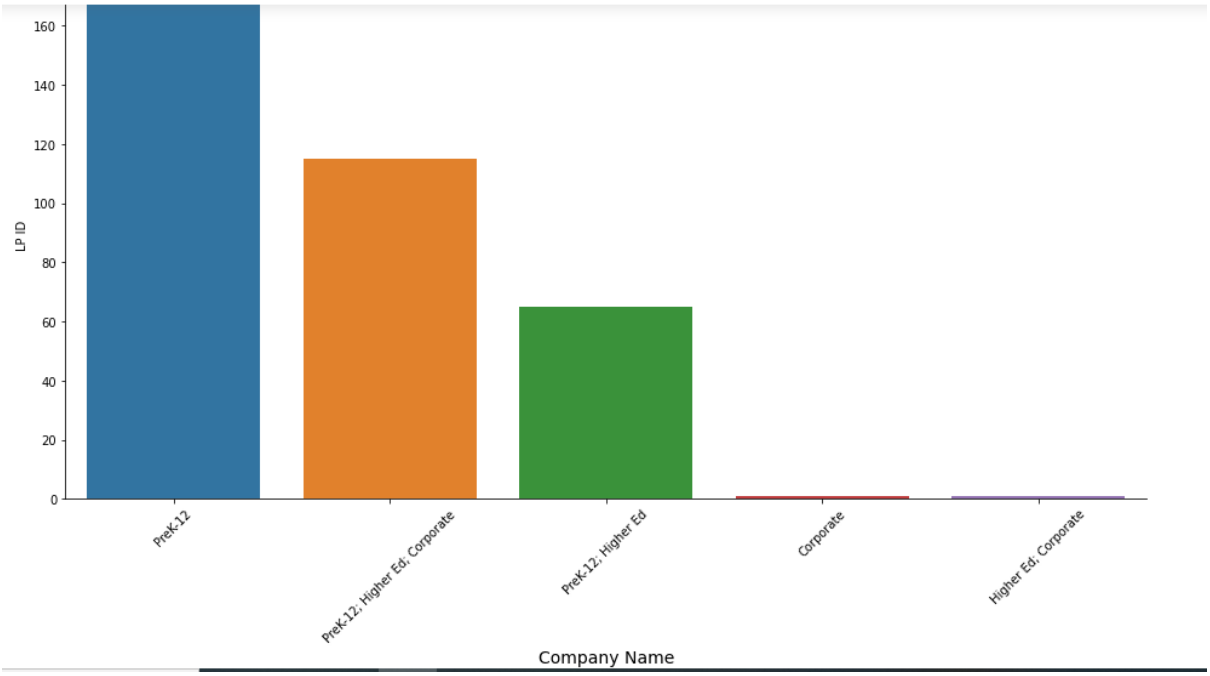


Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

A partir de la figure ci-dessus, on constate que Google LLC est le fournisseur/société qui a le plus de produits, avec 30 produits, suivi de Microsoft et Hooughton Mifflin Harcourt à égalité avec 6 produits. Nous pouvons voir qu'aucun fournisseur/société n'a plus de 10 produits, à part Google.

Nous représentons ainsi, les secteurs des plateformes les plus fréquentés

Figure 11: Les secteurs des plateformes les plus fréquentés



Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

D'après le graphique ci-dessus, nous remarquons que le secteur PreK-12 est le secteur de plateforme le plus fréquenté dans cet ensemble de données. Suivi respectivement par le secteur corporate et Higher Id qui sont des secteurs de l'enseignement supérieur et qui ne sont pratiquement pas fréquentés. En effet, le secteur PreK-12 est destiné aux élèves de la première à la douzième année.

2.3 La base de données "Engagement"

Les données d'engagement sont agrégées au niveau du district scolaire, et chaque fichier représente les données d'un district scolaire. Le nom de fichier à 4 chiffres représente le district_id qui peut être utilisé pour établir un lien avec les informations sur le district dans district_info. Le lp_id peut être utilisé pour établir un lien avec les informations sur les produits dans product_info.

Cet ensemble de données est composé des informations suivantes :

time : date en "YYYY-MM-DD".

lp_id : L'identifiant unique du produit

pct_access : Pourcentage d'élèves du district qui ont au moins un événement de chargement de page d'un produit donné et un jour donné.

engagement_index : Total des événements de chargement de page par millier d'élèves pour un produit donné et un jour donné.

Figure 12: Aperçu de la base de données Engagement

	time	district_id	lp_id	pct_access	engagement_index
0	2020-01-27	8815	32213.0	100.00	3000.00
1	2020-02-25	8815	90153.0	33.33	2666.67
2	2020-02-25	8815	99916.0	0.00	NaN
3	2020-02-25	8815	28504.0	0.00	NaN
4	2020-02-25	8815	95731.0	33.33	333.33

Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

Nous constatons que les variables de la base de données engagement sont toutes numériques. Elle contient des valeurs manquantes marquées par "NAN".

Tableau 3: Statistiques descriptives

	district_id	lp_id	pct_access	engagement_index
count	2.232419e+07	2.232365e+07	2.231074e+07	1.694578e+07
mean	5.237180e+03	5.470879e+04	5.042399e-01	1.676063e+02
std	2.644058e+03	2.647069e+04	3.180568e+00	1.682223e+03
min	1.000000e+03	1.000300e+04	0.000000e+00	1.000000e-02
25%	2.956000e+03	3.085100e+04	0.000000e+00	3.700000e-01
50%	4.929000e+03	5.500700e+04	2.000000e-02	1.920000e+00
75%	7.675000e+03	7.766000e+04	9.000000e-02	1.365000e+01
max	9.927000e+03	9.999100e+04	1.000000e+02	2.130455e+05

Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

Le tableau ci-dessus résume les statistiques descriptives de base de la base de données "engagement", soit la moyenne, le minimum, le maximum, les quartiles, etc des variables. Cette base de données continent des valeurs manquantes qui sont résumées dans la figure ci-après:

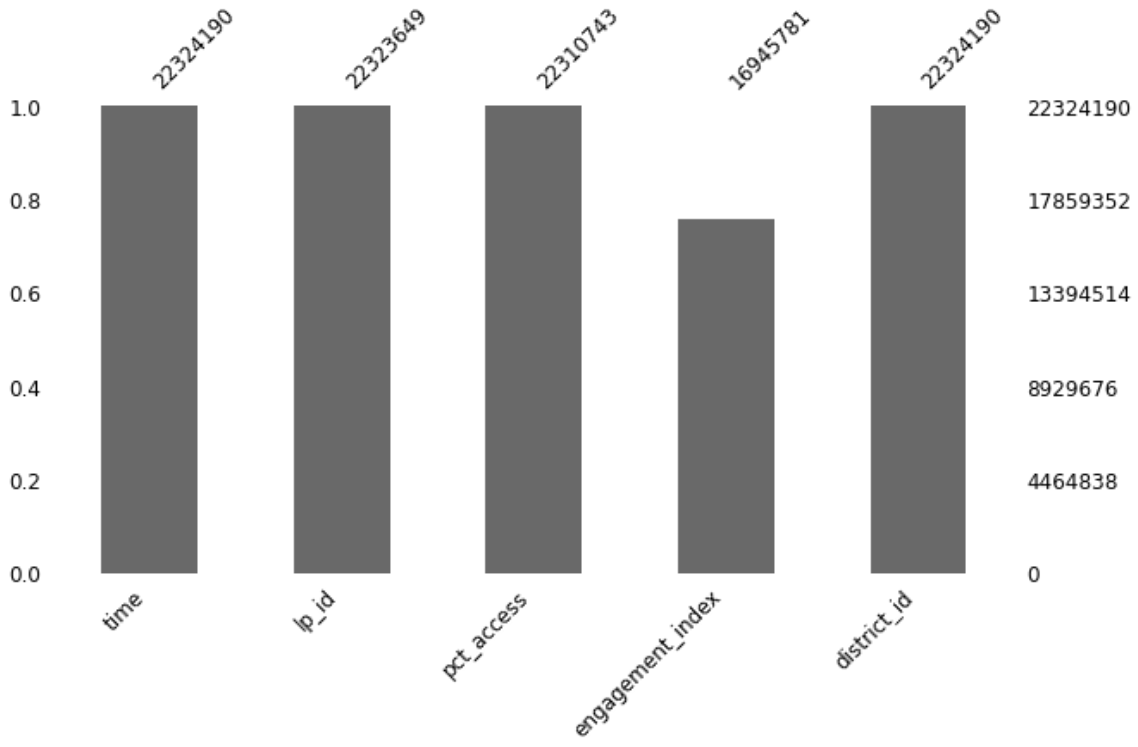
```
Number of missing Values in every column:
time                      0
district_id               0
lp_id                   541
pct_access            13447
engagement_index    5378409
dtype: int64
```

On constate que les deux variables district_id et time ne contiennent pas de valeurs manquantes.

2.4 Nettoyage des trois bases de données des valeurs manquantes

Afin d'éviter l'obtention de mauvais résultats, dû à des valeurs manquantes ou à des erreurs de saisie, nous allons procéder au nettoyage de nos bases de données.

Figure 13: La base de données après le nettoyage des valeur manquantes



Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

Comme on peut le constater sur le graphique précédent, nos trois bases de données contiennent toutes les trois des valeurs manquantes (NAN) et ces dernières peuvent interférer dans l'ajustement des résultats. Pour cela, nous allons procéder au nettoyage de ces bases de

données en utilisant la commande "**dropna**" sous python. En exécutant cette commande, on constate qu'effectivement les dimensions de nos bases de données baissent.

3 Modélisation de l'état de l'apprentissage numérique aux USA en 2020

Il est important de préciser que cette compétition ne comporte pas de modélisation par le machine learning. En effet, la majorité des participants ont effectué des analyses de base de statistiques descriptives pour visualiser justement l'état de l'apprentissage numérique aux USA durant la crise sanitaire liée à la COVID19.

Toutefois, en étudiant nos bases de données, nous avons réfléchi au fait que nous pouvons tirer un panier de variables qui peuvent justement expliquer une autre variable dépendante et pouvoir appliquer par conséquent un ensemble de méthodes de machine learning.

3.1 ARBRE DE DÉCISION

Un arbre de décision est un outil d'aide à la décision ou d'exploration de données qui permet de représenter un ensemble de choix sous la forme graphique d'un arbre. C'est une des méthodes d'apprentissage supervisé les plus populaires pour les problèmes de classification de données. Il existe deux principaux types d'arbre de décision : les arbres de régression qui permettent de prédire une valeur numérique, ils cherchent à maximiser la variance inter-classes et les arbres de classification qui permettent de prédire à quelle classe la variable de sortie appartient.[1]

Dans notre cas , nous allons plutôt nous intéresser aux arbres de régression pour essayer de créer des sous-ensemble dont la variable à expliquer, qui est engagement_index, soient les plus dispercés possible.

Pour cela, nous avons pensé d'abord à créer une nouvelle base de données avec seulement les variables qui nous intéressent. Nous nous retrouvons avec notre variable à expliquer et 4 variables explicatives (le nombre de variables explicatives va augmenter une fois le recodage des variables effectué, ce que nous expliquons juste après). Ensuite, nous avons supposé que la variable "engagement_index" qui représente le total des événements de chargement de page par millier d'élèves pour un produit donné et un jour donné comme une variable à expliquer qui peut être expliquée par l'ensemble de ces variables:

[1] https://dataanalyticspost.com/Lexique/arbre-de-decision/ , consulté le 18/11/2021.

pct_access: Cette variable peut expliquer en partie la variable d'intérêt,du fait qu'elle est une sous-composante de cette dernière.

locale : nous pouvons penser au fait que la zone peut jouer un rôle sur les évènements de chargement de page par milliers d'élèves car les zones rurales sont connues par une activité moins animée par rapport aux villes, ajoutant à cela le manque de réseau notamment dans les zones moins développées des USA.

pct_black/hispanic : Nous avons vu dans la figure n°6 que le pourcentage d'élèves des districts identifiés comme étant noirs ou hispaniques représentent moins de 20% du total des élèves, Par conséquent, cette variable est également une composante de la variable à expliquer et nous l'intégrons pour voir si elle a un impact significatif ou pas.

pct_free/reduced : Le pourcentage d'élèves des districts éligibles pour un déjeuner gratuit ou à prix réduit peut jouer d'une manière positive sur la variable dépendante car ces élèves auront plus d'avantages et moins de frais notamment pour payer l'internet par exemple ce qui laisse charger plus de pages dans un jour donné.

La variable "pct_access" est quantitative , tandis que les trois autres sont qualitatives avec plusieurs modalités. Pour cela, nous avons transformé chaque modalité (classe) de ces trois dernières en variables binaires pour pouvoir tracer l'arbre de classification. Pour cela, nous avons du modifier le type des variables pour les passer en catégories. Par la suite , nous avons créé une nouvelle base de données qui regroupe toutes les classes des variables et la variable quantitative "pct_access" et celle à expliquer (engagement_index) pour avoir au final une base de dimension suivante: 16945781 observations en ligne avec 16 variables en colonnes.

```
DF_dummies1 = pd.get_dummies(districts_engagement_data["pct_black/hispanic"], prefix="B/H", prefix_sep='_')
DF_dummies2 = pd.get_dummies(districts_engagement_data["locale"], prefix="Loc", prefix_sep='_')
DF_dummies3 = pd.get_dummies(districts_engagement_data["pct_free/reduced"], prefix="Free", prefix_sep='_')
```

Avant de passer à la modélisation, nous avons séparé le jeu de données en deux échantillons : le jeu "train" d'apprentissage et le jeu "test" .

En appelant les commandes nécessaires sous python,et en utilisant notamment la librairie "sklearn" , on obtient les résultats suivants:

Figure 14: La classification des variables explicatives selon leur importance

```
print(arbre1.score(X_train,y_train))
print(arbre1.score(X_test,y_test))
```

```
0.7784475489793808
0.5368047695966744
```

```
poids_variables = arbre1.feature_importances_
```

```
tab = pd.DataFrame([poids_variables])
tab.columns = df1.drop(['engagement_index'], axis=1).columns
tab
```

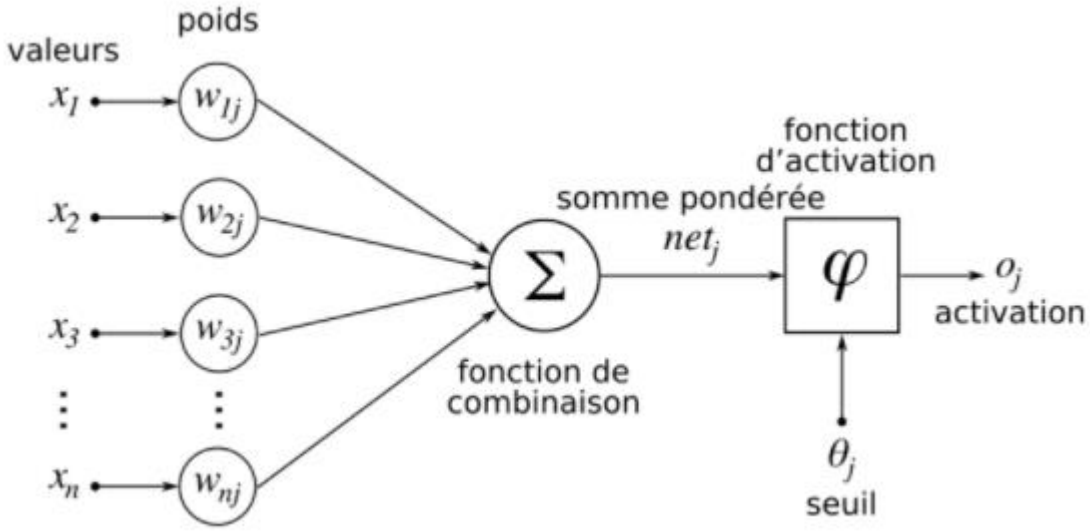
	pct_access	B/H_[0, 0.2[	B/H_[0.2, 0.4[	B/H_[0.4, 0.6[	B/H_[0.6, 0.8[	B/H_[0.8, 1[	Loc_City	Loc_Rural	Loc_Suburb	Loc_Town	Free_[0, 0.2[	Free_[0.2, 0.4[	Free_[0.4, 0.6[	Free_[0.6, 0.8[	Free_[
0	0.978982	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.018248	0.00277	0.0	0.0	0.0	0.0	

Source: MEZIANI.S et MORTEUIL.L à l'aide du logiciel Python

A partir de la figure ci-dessus, nous constatons que notre arbre de décision classe nos 15 variables explicatives selon leur importance dans l'explication des variations de la variable 'engagement_index'. On remarque que la variable "pcr-access" est la plus importante car elle explique 97,8% de la réussite de notre modèle. Toutes les autres modalités des trois autres variables explicatives que nous avons intégrées n'expliquent pratiquement pas la variable engagement_index. Avec ces résultats, nous avons une qualité d'ajustement de notre régression appliquée sur le jeu train qui est de l'ordre de 77,8 % et qui est un résultat satisfaisant en le comparant au score du jeu de données test soit 53,68%. En effet, le modèle de régression a réalisé le meilleur arbre de décision sur les données train après s'être entraîné sur ce même jeu de données plusieurs fois.

3.2 RÉSEAUX DE NEURONES

Les réseaux de neurones sont des imitations simples des fonctions d'un neurone dans le cerveau humain pour résoudre des problématiques d'apprentissage de Machine Learning. Ils s'avèrent plus performants que les techniques de régression pour des tâches de Machine Learning. Nous pouvons résumer son architecture dans la figure ci-dessous:



Un réseau de neurones peut prendre des formes différentes selon l'objet de la donnée qu'il traite et selon sa complexité et la méthode de traitement de la donnée. De plus, il est appliqué dans plusieurs domaines comme la reconnaissance d'images, le filtrage des données, prédiction des données, etc.[2]

Les architectures de réseaux neuronaux peuvent être divisées en 4 grandes familles :

- Réseaux de neurones Feed fowarded
- Réseaux de neurones récurrent (RNN)
- Réseaux de neurones à résonance
- Réseaux de neurones auto-organisés

Dans notre cas, le réseau de neurone adapté à nos données appartient aux réseaux de neurones Feed fowarded et qui est le MLP (Multi Layer Perceptron Regressor) qui nécessite des entrées qui se transforment en sortie mais il ne peut y avoir de retour en arrière.

Dans un premier temps, nous avons testé plusieurs critères sur notre régression MLP et nous avons choisi de garder ceux qui rendaient le meilleur résultat. Nous avons choisi de lui affecter 3 couches cachées de 30 neurones et une fonction d'activation de type 'relu' (f(x) = max(0, x)). Sa tolérance d'optimisation est de 0.01, ce qui signifie que lorsque le score ne s'améliore pas de 0.01 sur 10 itérations consécutives alors la convergence est considérée comme atteinte et l'apprentissage s'arrête.

[2] https://www.juripredis.com/fr/blog/id-19-demystifier-le-machine-learning-partie-2-les-reseaux-de-neurones-artificiels , consulté le 18/11/2021.

Avec nos deux jeux de données “train” et “test” , nous avons appliqué la méthode des réseaux de neurones MLP et nous avons obtenu ce résultat:

```
MLP1 = MLP_regressor.fit(X_train,y_train)
print(MLP1.score(X_train,y_train))
print(MLP1.score(X_test,y_test))
```

```
0.6816699485597435
0.5257904675320686
```

Le réseau de neurone appliqué sur nos données nous fournit un score qui est égal à 68,16% sur le jeu de données train ce qui signifie que certaines de nos variables explicatives permettent d’expliquer une partie des variations de la variable engagement_index. Le score du jeu test est de 52,6%.

3.3 RÉGRESSION LINÉAIRE MULTIPLE

Nous disposons de 15 variables explicatives et une variable à expliquer. Pour celà, nous avons pensé à revenir aux modélisations de base du machine learning, soit la régression linéaire multiple qu’on a appliqué sur le jeu de données train et test.

```
Linear1 = Linear_regressor.fit(X_train,y_train)
print(Linear1.score(X_train,y_train))
print(Linear1.score(X_test,y_test))
```

```
0.6206714804631507
0.5042017440226683
```

```
coef=Linear1.coef_
```

```
tab = pd.DataFrame([coef])
tab.columns = df1.drop(['engagement_index'], axis=1).columns
tab
```

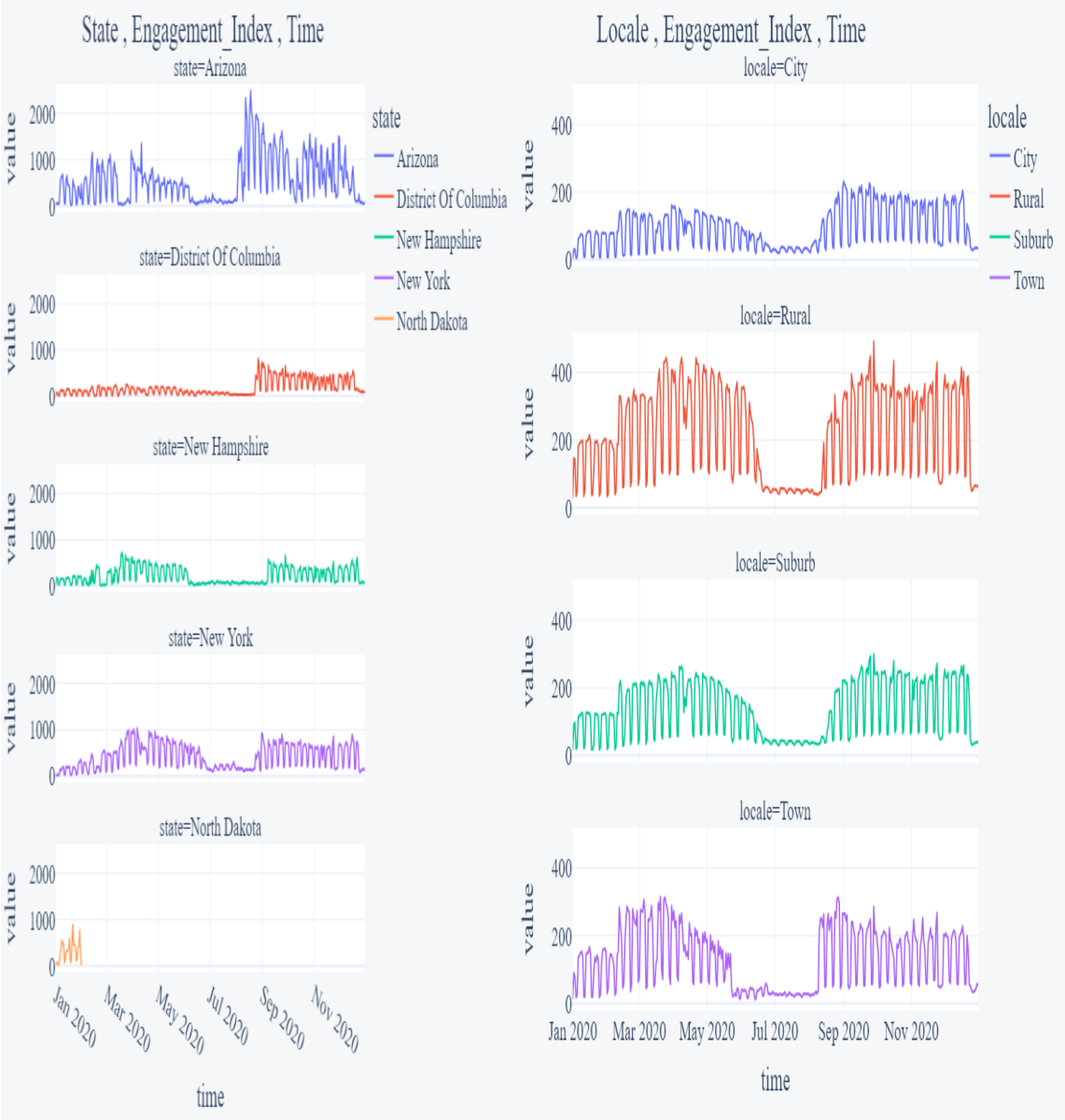
	pct_access	B/H_[0, 0.2[	B/H_[0.2, 0.4[	B/H_[0.4, 0.6[	B/H_[0.6, 0.8[	B/H_[0.8, 1[	Loc_City	Loc_Rural	Loc_Suburb	Loc_Town	Free_[0, 0.2[	Free_[0.2, 0.4[	Free_[0.4, 0.6[	Free_
0	380.656409	25.873562	18.05681	155.16823	80.405495	-91.00081	55.439166	9.738704	-16.73511	140.060527	-10.420501	-42.326233	-99.290854	-78.828

A partir de la figure ci-dessus, nous pouvons dire que les résultats de cette régression sont proches des deux réalisées précédemment. En effet, le score du jeu de données train est plus important en le comparant au jeu test , soit 62% et 50,4% respectivement. La variable pct_access qui est la variable qui explique le plus notre variable à expliquer parmis nos variables explicatives à une influence positive sur la variable à expliquer.

3.4 MODÉLISATION PAR SÉRIE TEMPORELLE

Comme nous avons la variable “time” dans notre data frame, qui est en format journalier ; nous avons tracé plusieurs graphiques de séries temporelles représentant notamment la variable “engagement_index” en fonction du temps dans les différentes zones et des Etats. Nous avons pensé à une procédure que nous avons tenté sur python mais qui n’a pas marché. Cette dernière consiste à tracer une droite séparant les données de la série temporelle en valeurs supérieures et inférieures à la moyenne ou à la médiane pour pouvoir récupérer deux jeux de données et appliquer un autre type de réseau de neurones.

Figure 15: La série temporelle de la variable engagement_index



Source: MEZIANI.S et MORTEUIL.L à l’aide du logiciel Python

3.5 COMPARAISON DES RÉSULTATS OBTENUS

Dans cette partie, nous allons dresser une comparaison des résultats que nous avons obtenus dans les trois estimations effectué dans le point précédent:

Tableau 4 : Comparaison des résultats

Méthode	Arbre de régression	Réseau de neurone	Régression OLS
Score jeu train	77,8 %	68,16%	62%
Score jeu test	53,68%	52,6%	50,4%

Source: MEZIANI.S et MORTEUIL.L

En analysant ce tableau, on constate que les scores obtenus pour les trois méthodes que ce soit pour le jeu test ou train sont proches les uns des autres avec un score plus élevé du jeu train (ce qui est normal). La meilleure estimation est l'arbre de régression avec un score de 77.8%pour le jeu train et 53,68% pour le jeu test..

4 CONCLUSION ET DISCUSSIONS DES RÉSULTATS

Les effets de la pandémie de COVID-19 et les restrictions de mouvement et la fermeture des établissements d'enseignement et des lieux de travail qui en découlent ont montré que sans accès à Internet, il est impossible de participer à une grande partie de la vie sociale et économique. De plus, la pandémie du virus Covid-19 a stimulé l'utilisation des plateformes d'apprentissage numérique et l'étude à domicile, obligeant chaque étudiant à l'utiliser pour obtenir une éducation partout dans le monde notamment aux USA.

Après avoir traiter les données disponibles sur l'apprentissage numérique aux Etats Unis, et après avoir visualisé les principales graphiques, nous pouvons tirer les conclusions suivantes:

La région suburbaine continent le plus de districts scolaires dans la plupart des Etats des USA. Dans la majorité des cas, le pourcentage des étudiants afroaméricains ou hispanoaméricains représente moins de **20%** des étudiants de ces districts. De plus, Google LLC est le fournisseur qui a le plus de produits.

New York, l'État où l'indice d'engagement est le plus élevé avec un score de 7,2, est également l'État où les dépenses par élève sont les plus élevées (18 000 USD). Alors que le pourcentage d'étudiants éligibles à la gratuité (32%) et à la réduction (52%) n'est pas directement lié à l'indice d'engagement.

Après le nettoyage de nos bases de données et après avoir fusionné les deux bases d'engagement et de district, nous avons effectué une estimation en utilisant la méthode de l'arbre de régression, un réseau de neurone MLP et une régression linéaire multiple pour expliquer la variable "engagement _index" . Les résultats obtenus montrent que le total d'événements de charge de pages par milliers d'élèves dépend fortement du nombre d'élèves qui ont chargé des pages effectivement dans le district (ce qui semble normal, les deux variables sont possiblement très corrélées).

La pandémie de COVID-19 a provoqué la plus grande perturbation des systèmes éducatifs de l'histoire, dont ont pâti la quasi-totalité des élèves et des enseignants de la planète, dans les écoles maternelles et primaires, les collèges et lycées, les établissements d'enseignement et de formation techniques et professionnels, les universités, etc.

La capacité de faire face aux fermetures d'écoles a été plus ou moins grande selon le niveau de développement, avec de très de fortes disparités : ainsi, au cours du deuxième trimestre 2020, 86% des élèves du primaire ont cessé complètement d'être scolarisés dans les pays à faible indice de développement humain, contre seulement 20 % dans les pays à indice de développement humain très élevé. Pour cela, l'accès à l'internet est devenu un besoin fondamental à satisfaire et qui devrait être envisagé de toute urgence avant d'envoyer des enfants en formation à distance. Comme le montrent les données de l'année scolaire 2020/21 aux USA, cette recommandation a été suivie par plusieurs districts scolaires. Cette recommandation devrait être considérée comme une meilleure pratique par les décideurs politiques à l'avenir pour aider à créer une base équitable pour l'apprentissage à distance.

5 LISTE DES TABLEAUX ET DES FIGURES

6 TABLE DE MATIERES