MACHINE LEARNING REPORT

1. INTRODUCTIOIN

根据以上数据分析作为基础，我们将进一步对数据进行机器学习，以达到探明更深层次规律以及产生实际运用的目的。机器学习分为监督学习、无监督学习和强化学习三类，本项目仅涉及监督学习和无监督学习部分。

在无监督学习方面，我们会使用属于聚类算法的 KMeans 进行聚类分析；在监督学习方面，我们会使用多种典型分类算法进行内容分类，并分析它们之间的优劣属性，另外，我们会使用线性回归，岭回归和随机森林算法对数据进行回归预测。

Based on the above data analysis, we will further machine learning the data to achieve the purpose of exploring deeper laws and practical applications. Machine learning is divided into supervised learning, unsupervised learning and intensive learning. This project only involves supervised learning and unsupervised learning.

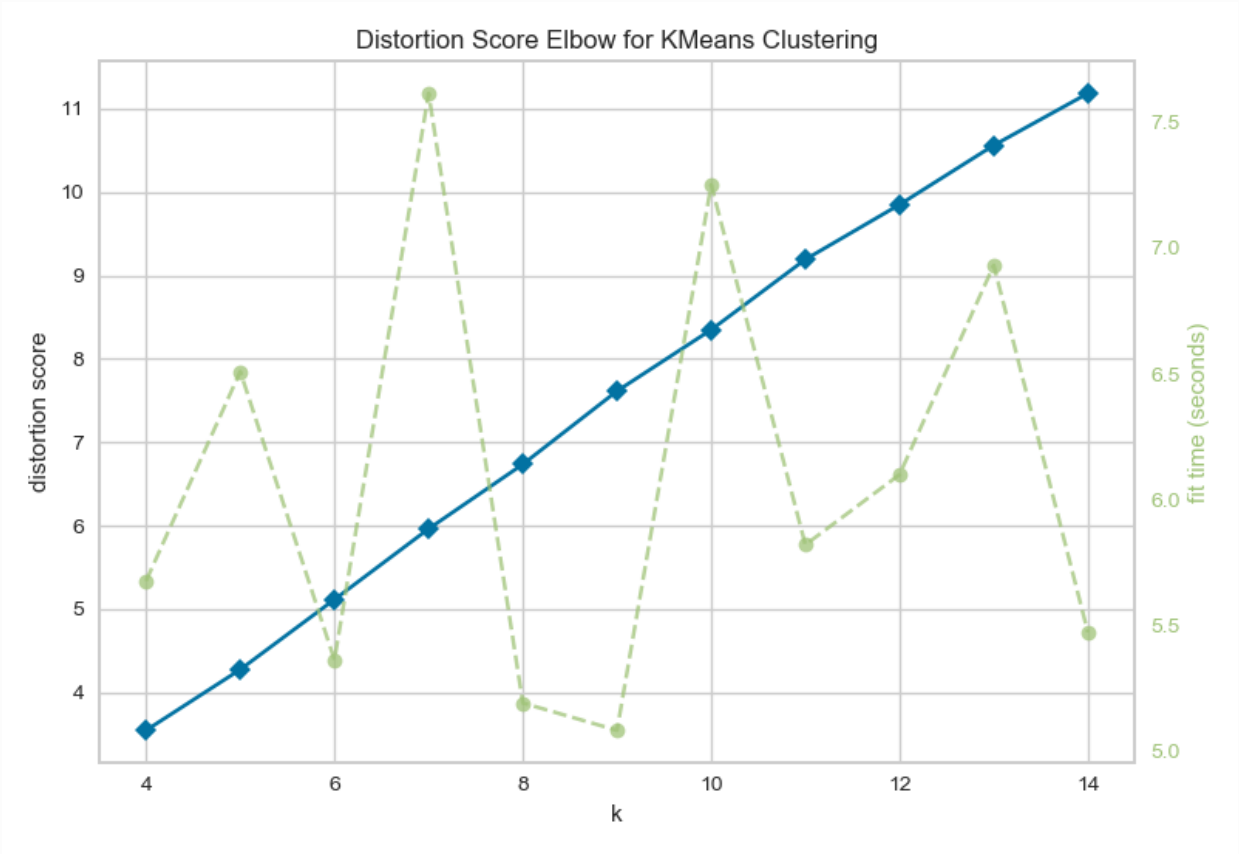
In unsupervised learning, we use KMeans, which is a clustering algorithm, for cluster analysis. In supervised learning, we use a variety of typical classification algorithms to classify content and analyze the pros and cons between them. In addition, we Linear regression, ridge regression and random forest algorithms are used to predict the data.

Sur la base de l'analyse des données ci-dessus, nous approfondirons l'apprentissage automatique des données afin d'explorer des lois plus en profondeur et des applications pratiques. L'apprentissage automatique est divisé en apprentissage supervisé, apprentissage non supervisé et apprentissage intensif.Ce projet ne comprend que l'apprentissage supervisé et l'apprentissage non supervisé.

Dans l’apprentissage non supervisé, nous utilisons KMeans, un algorithme de classification, pour l’analyse par grappes. Dans l’apprentissage supervisé, nous utilisons une variété d’algorithmes de classification typiques pour classifier le contenu et analyser les avantages et les inconvénients entre eux. Les algorithmes de régression linéaire, de régression de crête et de forêt aléatoire sont utilisés pour prédire les données.

2. UNSUPERVISED LEARNING - CLUSTERING

1) Find the Most Suitable K Value



为了使用 KMeans 方法进行聚类，我们需要预先了解一个合适的 K 值，以决定将数据分为多少簇。我们使用 Elbow Method 对 4 - 15 范围内的 K 值进行了计算，得到每个 K 值对应的 Distortion Score 和 Fit Time。对于同一个 K 值而言，我们需要 Distortion Score 尽可能小且 Fit Time 尽可能低，所以我们可以选择 7 作为簇数量。

In order to cluster using the KMeans method, we need to know a suitable K value in advance to decide how many clusters to divide the data into. We use the Elbow Method to calculate the K values in the range 4 - 15 to get the Distortion Score and Fit Time for each K value. For the same K value, we need the Distortion Score to be as small as possible and the Fit Time to be as low as possible, so we can choose 7 as the number of clusters.

Pour regrouper à l'aide de la méthode KMeans, nous devons connaître à l'avance une valeur K appropriée afin de déterminer le nombre de grappes dans lesquels diviser les données. Nous utilisons la méthode Elbow pour calculer les valeurs K dans la plage 4-15 afin d'obtenir le score de distorsion et le temps d'adaptation pour chaque valeur K. Pour la même valeur K, nous avons besoin que le score de distorsion soit le plus petit possible et que le temps d'adaptation soit le plus bas possible, afin de pouvoir choisir 7 comme nombre de clusters.

2) Do the cluster

我们将 TF-IDF 矩阵作为训练依据传入 KMeans 方法，由此得到一个模型。我们存储训练出的模型，以便在之后直接调用模型而不是再次花费时间进行训练。

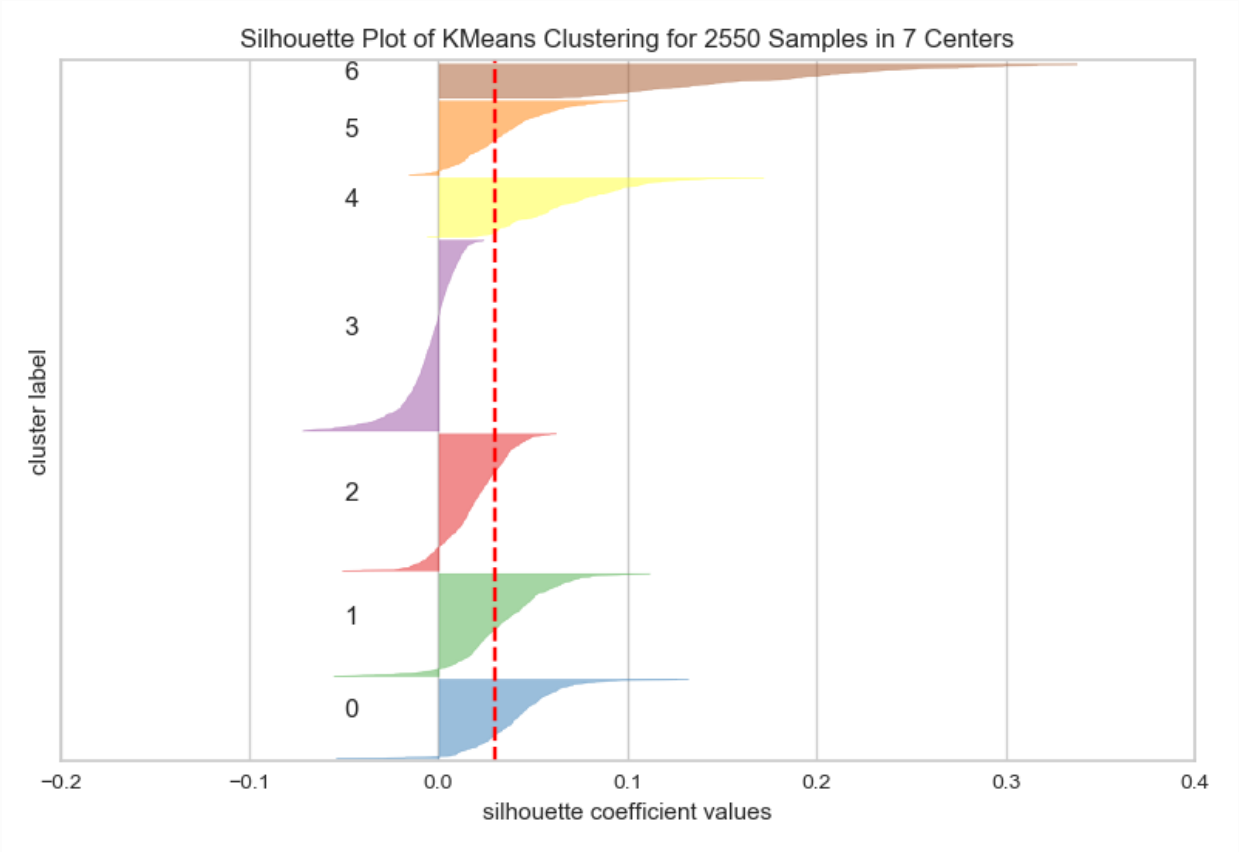
We pass the TF-IDF matrix as a training basis to the KMeans method, and thus get a model. We store the trained models so that we can call the model directly instead of spending time again.

Nous passons la matrice TF-IDF comme base d'apprentissage à la méthode KMeans et obtenons ainsi un modèle. Nous stockons les modèles formés de manière à pouvoir appeler le modèle directement au lieu de passer du temps à nouveau.

```
tfidf = TfidfVectorizer(max_df=0.8, stop_words='english')
tfidf_matrix = tfidf.fit_transform(replaceList)
km = KMeans(n_clusters=7).fit(tfidf_matrix)

joblib.dump(km, modelURL)
```

3) Calculate the Sihouette Value



Sihouette Value 是用来检测 KMeans 分类好坏程度的一种系数。我们通过可视化看到，被分为 7 类的数据项中，分为第 7 项的数据比较准确，第 4 类数据数值出现了较大规模的负数，结果并不良好。总体而言，聚类结果可以接受。

Sihouette Value is a factor used to detect how well KMeans are classified. We can see through visualization that among the data items classified into 7 categories, the data classified into the 7th item is more accurate, and the value of the 4th type data has a larger negative number, and the result is not good. Overall, the clustering results are acceptable.

La valeur de Sihouette est un facteur utilisé pour détecter la qualité de la classification des KM. Nous pouvons voir à travers la visualisation que parmi les données classées en 7 catégories, les données classées dans le septième sont plus précises et que la valeur des données de type 4 a un nombre négatif plus grand et que le résultat n'est pas bon. Dans l'ensemble, les résultats du regroupement sont acceptables.

3. SUPERVISED LEARNING - CLASSIFICATION

1. Find out input and output

我们使用上一步 clustering 得到的结果作为新的参数，称为「主题类别」，它代表了视频所属的主题类别。另外，我们根据直方图，均衡分割了评论量、观看量和持续时间三个属性，将它们从连续变量变成了离散变量，从而可以在classification 中使用它们。我们决定使用主题、评论量和持续时间作为输入量，得到一个关于观看量的分类。

We use the result of clustering in the previous step as a new parameter called "topic category", which represents the topic category to which the video belongs. In addition, we divide the three attributes of comment quantity, view quantity and duration according to the histogram, and change them from continuous variables to discrete variables, so that they can be used in classification. We decided to use the subject, comment volume, and duration as input to get a categorization of views.

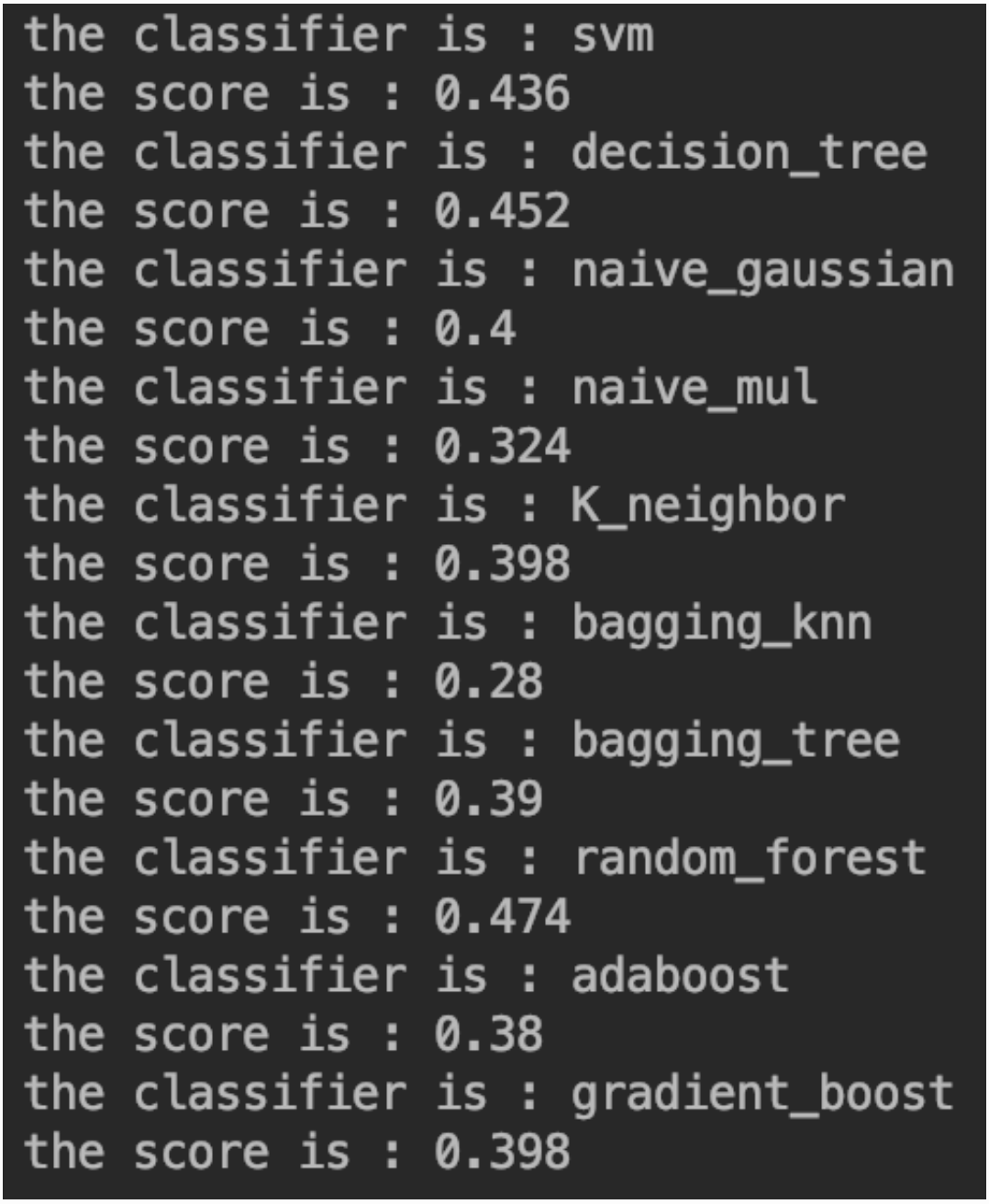
Nous utilisons le résultat de la classification à l'étape précédente en tant que nouveau paramètre appelé "catégorie de sujet", qui représente la catégorie de sujet à laquelle la vidéo appartient. De plus, nous divisons les trois attributs quantité de commentaire, quantité de vue et durée en fonction de l'histogramme, et les modifions de variables continues à variables discrètes, afin de pouvoir les utiliser dans la classification. Nous avons décidé d'utiliser le sujet, le volume de commentaire et la durée comme entrées pour obtenir une catégorisation des vues.

2. Select classification algorithm

我们使用 2000 余个数据中的 500 条作为测试集。通过训练集，我们使用不同算法得到分类模型结果，并对这些结果进行了评分。结果如下：

We used 500 of the more than 2,000 data as a test set. Through the training set, we used different algorithms to obtain the classification model results and scored the results. The results are as follows:

Nous avons utilisé 500 des plus de 2 000 données comme ensemble de test. Au cours de la formation, nous avons utilisé différents algorithmes pour obtenir les résultats du modèle de classification et les noter. Les résultats sont les suivants:



从评分我们可以看到，作为集成方法的 RandomForest 评分最高，而同样作为集成方法的 BaggingKnn 算法则得分最低。因此，我们的分类决定使用 RandomForest 算法。

From the scoring we can see that the RandomForest score is the highest for the integration method, and the BaggingKnn algorithm, which is also the integration method, has the lowest score. Therefore, our classification decided to use the RandomForest algorithm.

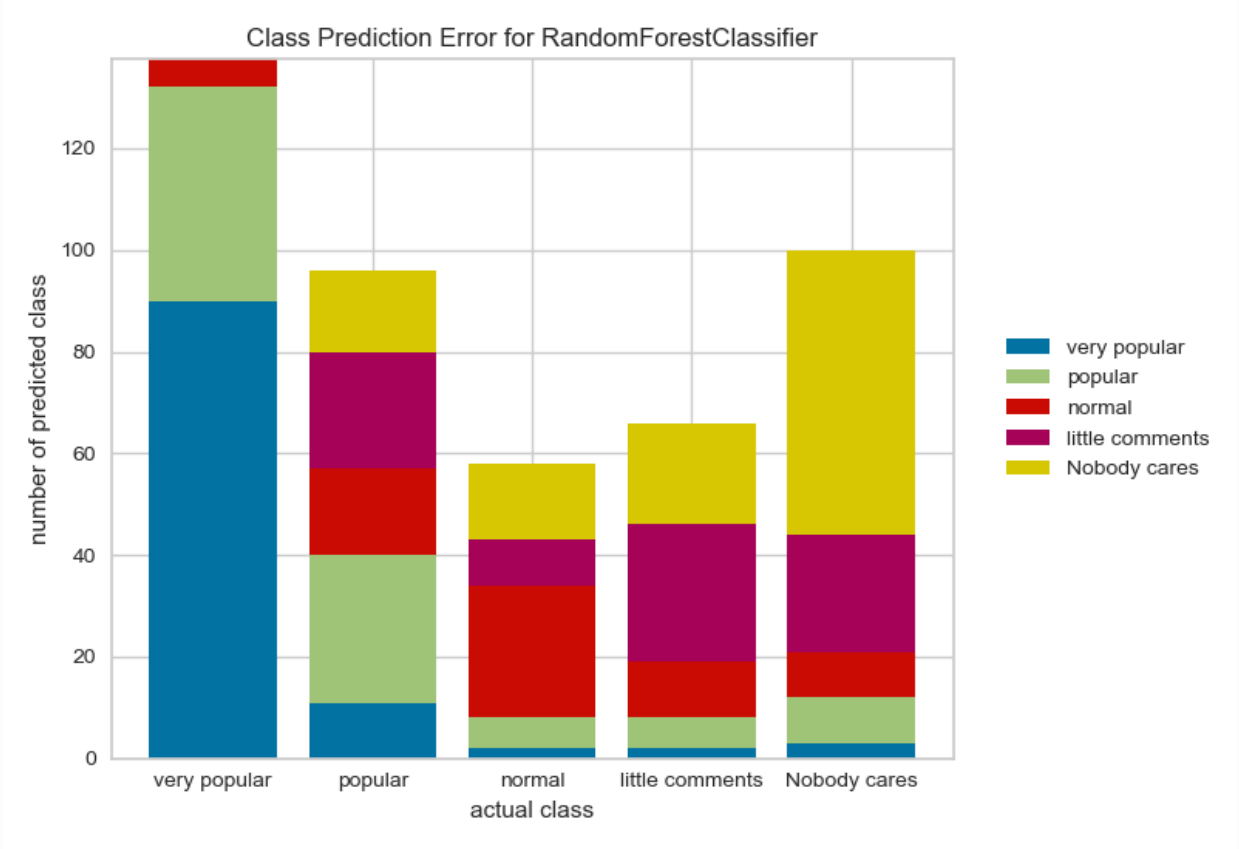
La notation indique que le score RandomForest est le plus élevé pour la méthode d'intégration et que l'algorithme de BaggingKnn, qui est également la méthode d'intégration, a le score le plus bas. Par conséquent, notre classification a décidé d'utiliser l'algorithme RandomForest.

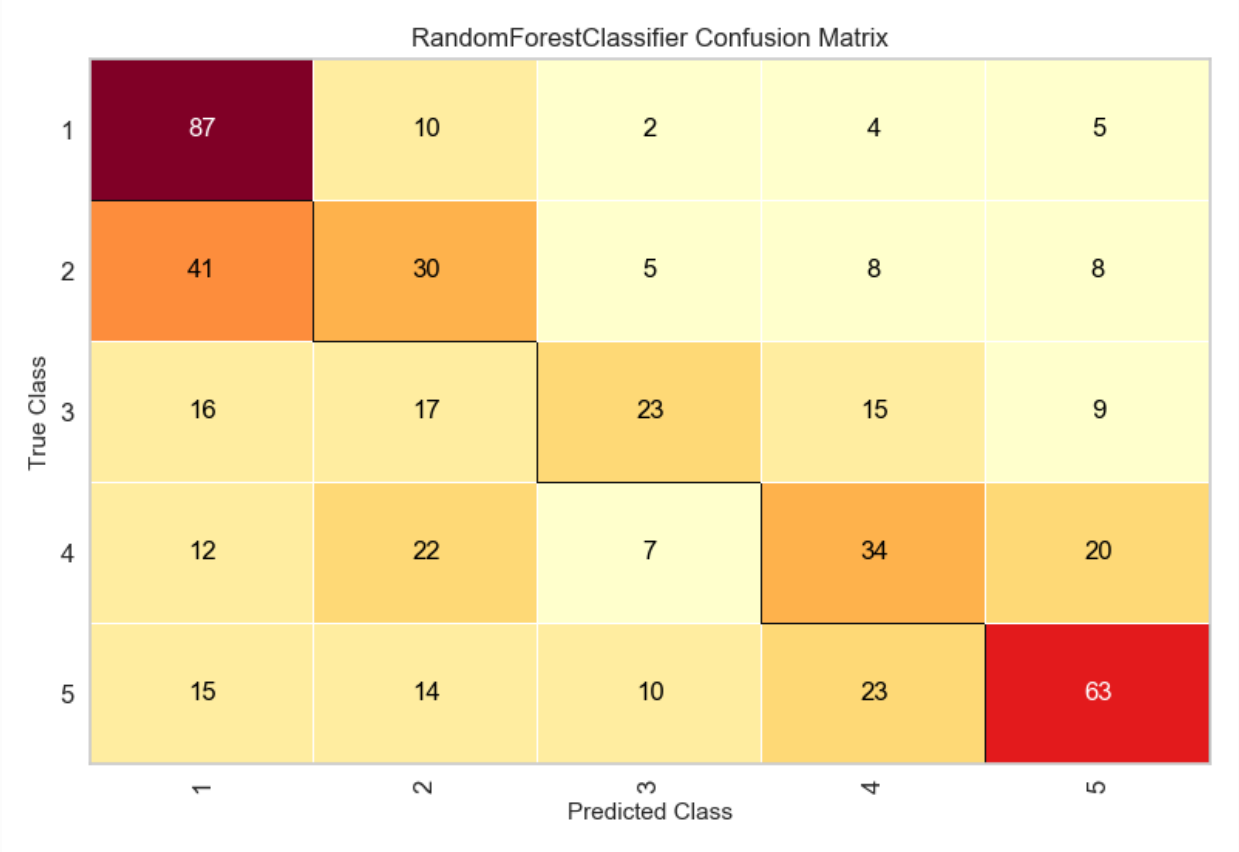
3. Test results

我们使用 RandomForest 算法进行分类，之后为了进行结果检验，我们得到这种算法分类过后得到的 Prediction Error Figure 和 Confusion Matrix 。

We use the RandomForest algorithm for classification, and then for the result test, we get the Prediction Error Figure and Confusion Matrix obtained by this algorithm classification.

Nous utilisons l'algorithme RandomForest pour la classification, puis pour le test de résultat, nous obtenons la figure d'erreur de prédiction et la matrice de confusion obtenues par cette classification d'algorithme.





结果可以看出，本次分类结果不能算是十分完美，但是基本完成了分类任务，对于测试集数据能够进行大致区分。

我们猜测问题出现在相关性不强。在讨论 Regression 时我们会看到，本次数据集中的几个数据之间没有很强烈的相关性，因此其预测结果并不能达到具有强相关性的数据集那么完美。

As a result, it can be seen that the classification result is not perfect, but the classification task is basically completed, and the test set data can be roughly distinguished.

We suspect that the problem is not relevant. When we discussed Regression, we saw that there is not a strong correlation between the data in this dataset, so the prediction results are not as perfect as the datasets with strong correlation.

En conséquence, on peut constater que le résultat de la classification n'est pas parfait, mais la tâche de classification est essentiellement terminée et les données de l'ensemble de tests peuvent être distinguées de manière approximative.

Nous soupçonnons que le problème n'est pas pertinent. Lorsque nous avons discuté de la régression, nous avons constaté qu'il n'existait pas de forte corrélation entre les données de cet ensemble de données, de sorte que les résultats de la prévision ne sont pas aussi parfaits que ceux des ensembles de données présentant une forte corrélation.

4. SUPERVISED LEARNING - REGRESSION

1. Looking for relevance

我们使用了 Heatmap 来表述各数字属性之间的关系。

We used Heatmap to represent the relationship between the various numeric attributes.

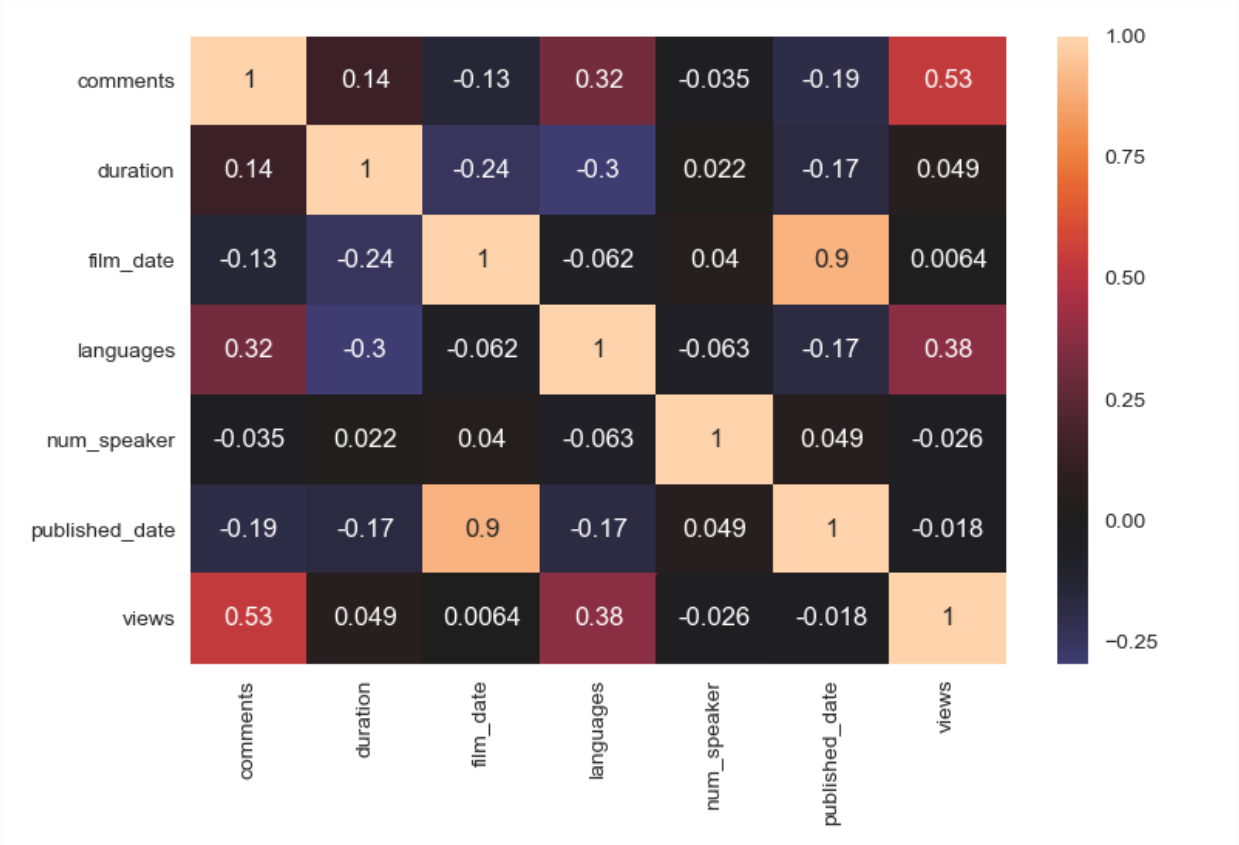
Nous avons utilisé Heatmap pour représenter la relation entre les différents attributs numériques.

```
seaborn.heatmap(data.corr(), center=0, annot=True)
mp.show()
```

得到了如下结果。

The following results were obtained.

Les résultats suivants ont été obtenus.



我们可以看出，除了 views 和 comments 之间有较明显的正相关（0.53）之外，其他属性之间并无太大强相连性。这也证明了我们可能并不能得出完美的预测结果，但是我们可以尝试去进行回归。我们最终使用了 duration, language 和 views 来对 comments 进行回归预测。

We can see that there is not much strong correlation between other attributes except for the obvious positive correlation (0.53) between views and comments. This also proves that we may not be able to produce perfect predictions, but we can try to go back. We ended up using duration, language and views to make regression predictions on comments.

Nous pouvons constater qu'il n'y a pas beaucoup de corrélation forte entre les autres attributs, à l'exception de la corrélation positive évidente (0,53) entre les vues et les commentaires. Cela prouve également que nous ne sommes peut-être pas en mesure de produire des prévisions parfaites, mais nous pouvons essayer de revenir en arrière. Nous avons fini par utiliser la durée, le langage et les vues pour faire des prédictions de régression sur les commentaires.

2. Try different regression algorithms

为了得到相对好的回归结果，我们尝试了三种算法：线性回归，岭回归和随机森林算法。它们的结果如下：

In order to get relatively good regression results, we tried three algorithms: linear regression, ridge regression and random forest algorithm. Their results are as follows:

Afin d'obtenir des résultats de régression relativement bons, nous avons essayé trois algorithmes: la régression linéaire, la régression de crête et l'algorithme de forêt aléatoire. Leurs résultats sont les suivants:

```
MSE: 14752.617541618034
MSE(Calculate MSE according to the formula): 14752.617541618047
RMSE: 121.46035378516743


MSE: 6954.128185800001
MSE(Calculate MSE according to the formula): 6954.128185799998
RMSE: 83.39141554021013


MSE: 14752.619583607015
MSE(Calculate MSE according to the formula): 14752.61958360701
RMSE: 121.46036219115689
```

我们可以看到，位于中间的随机森林算法得到的 RMSE 偏差最小，而线性回归和岭回归的结果并无太大差别，距离随机森林算法的结果有较大误差。因此，我们决定在项目中使用随机森林算法进行回归预测。

然而，相比于数据的平均数 186.17 而言，RMSE 高达 83，这是不太好的结果。所以我们考虑试试其他算法分析。

We can see that the random forest algorithm in the middle obtains the smallest RMSE deviation, while the results of linear regression and ridge regression are not much different, and the results from the random forest algorithm have large errors. Therefore, we decided to use the random forest algorithm for regression prediction in the project.

However, the RMSE is as high as 83 compared to the average of 186.17, which is not a good result. So we consider trying other algorithmic analysis.

Nous pouvons voir que l'algorithme de forêt aléatoire au milieu obtient l'écart RMSE le plus petit, alors que les résultats de régression linéaire et de régression de crête ne sont pas très différents et que les résultats de l'algorithme de forêt aléatoire ont de grandes erreurs. Par conséquent, nous avons décidé d'utiliser l'algorithme de forêt aléatoire pour la prédiction de régression dans le projet.

Cependant, le RMSE est aussi élevé que 83 par rapport à la moyenne de 186,17, ce qui n'est pas un bon résultat. Nous envisageons donc d'essayer d'autres analyses algorithmiques.

3. Deviation analysis

我们使用可视化组件 Yellowbricks 进行另一种方式的误差分析。我们选取了 Lasso 和 Ridge 方法进行误差分析。

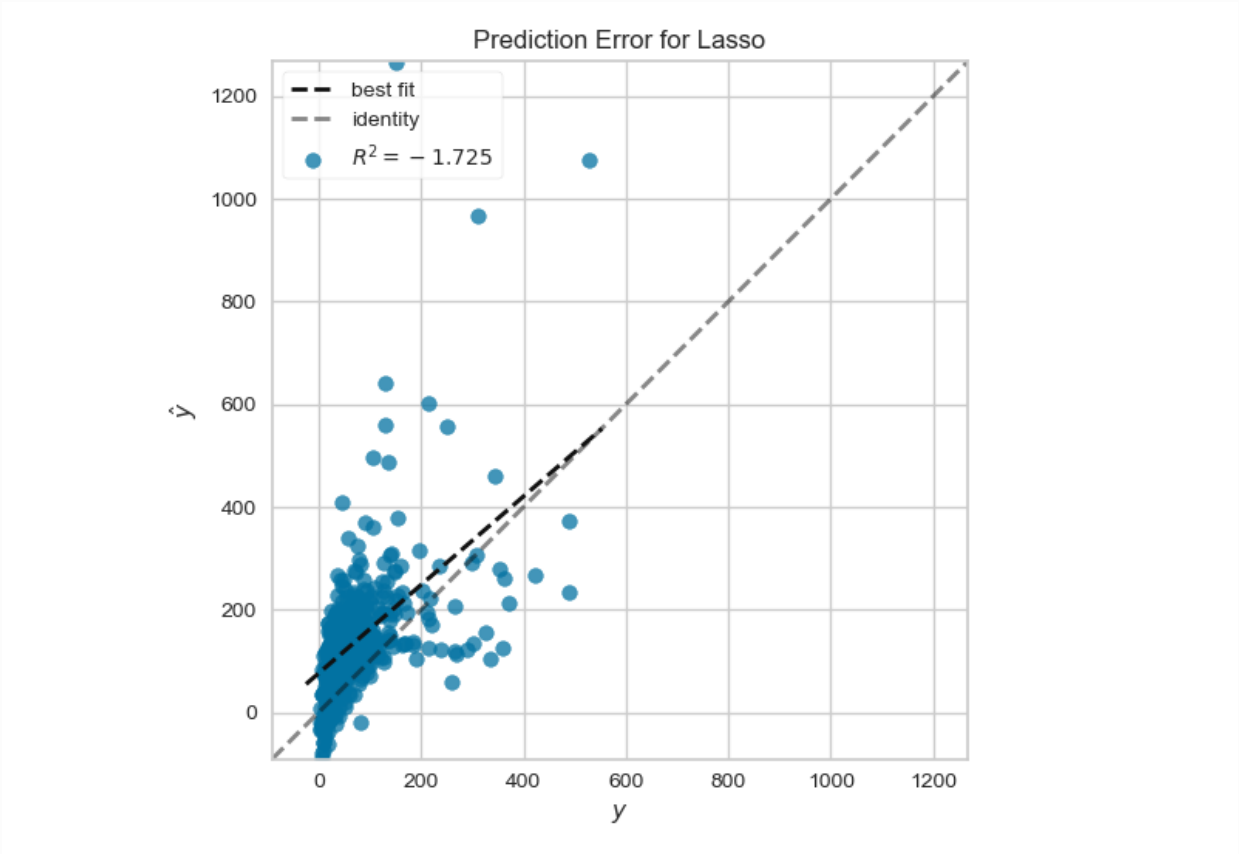
We use the visual component Yellowbricks for another way of error analysis. We chose the Lasso and Ridge methods for error analysis.

Nous utilisons le composant visuel Yellowbricks pour une autre manière d'analyser les erreurs. Nous avons choisi les méthodes de Lasso et Ridge pour l'analyse d'erreur.

Lasso 是针对稀疏数据进行的有偏估计，可以在稀疏数据下尽可能降低 MSE 的值。

Lasso is a biased estimate of sparse data that minimizes the value of MSE under sparse data.

Lasso est une estimation biaisée de données rares qui minimise la valeur de la MPE sous des données rares.



我们通过 Lasso 算法进行的回归分析得到上面的图像。我们的数据几乎聚集在 45 度角的直线附近，但是预测值都比真实值偏大了一点点。这张图可以看出，使用 Lasso 算法比较适合我们的稀疏、无关数据，但是因为它是有偏估计，我们不能运用在真实预测上。

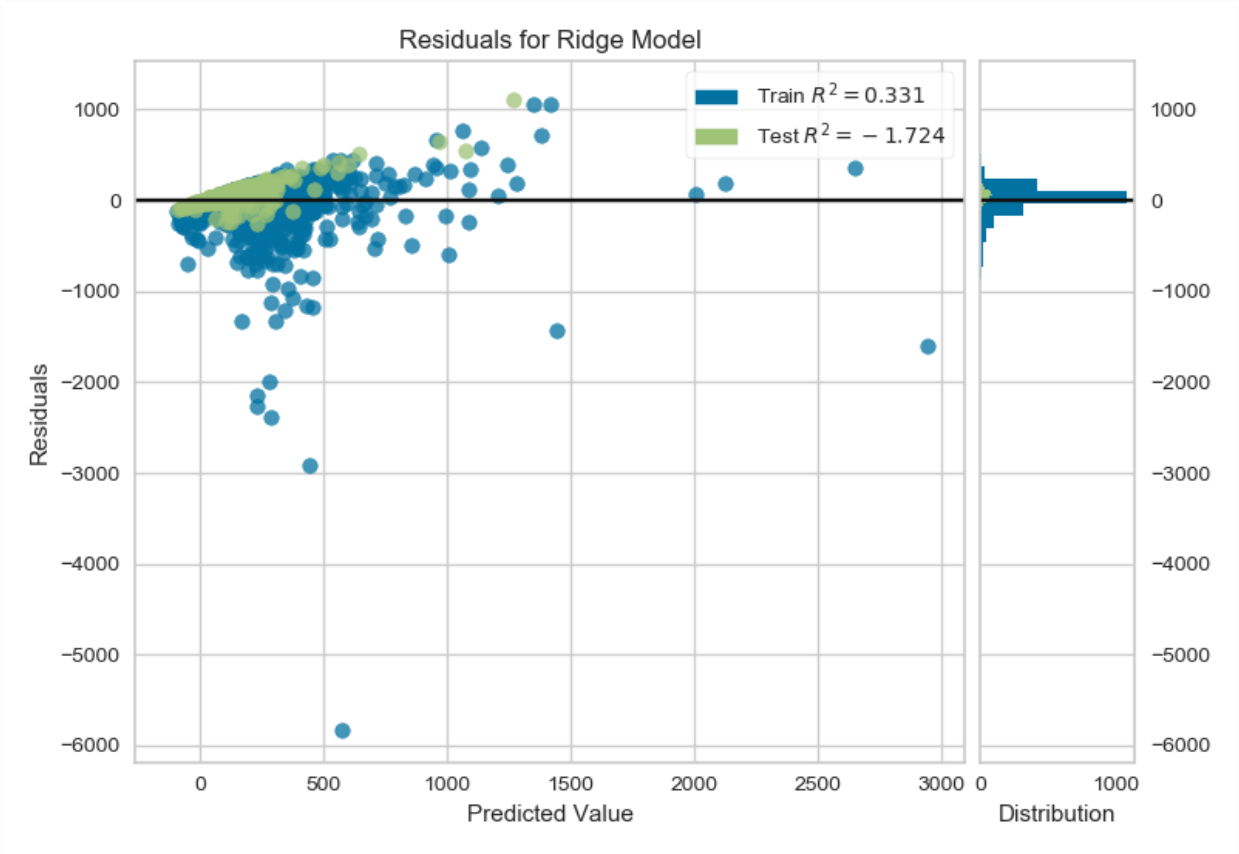
We obtained the above image by regression analysis performed by the Lasso algorithm. Our data is almost concentrated near a straight line at a 45-degree angle, but the predicted values are a little larger than the true value. As you can see from the figure, the Lasso algorithm is more suitable for our sparse, irrelevant data, but because it is a biased estimate, we can't use it in real prediction.

Nous avons obtenu l'image ci-dessus par analyse de régression effectuée par l'algorithme de Lasso. Nos données sont presque concentrées près d'une ligne droite à un angle de 45 degrés, mais les valeurs prédites sont un peu plus grandes que la valeur réelle. Comme vous pouvez le voir sur la figure, l'algorithme de Lasso convient mieux à nos données éparses et non pertinentes, mais comme il s'agit d'une estimation biaisée, nous ne pouvons pas l'utiliser dans des prédictions réelles.

我们另外使用 Ridge 算法进行了误差分析。

We also used the Ridge algorithm for error analysis.

Nous avons également utilisé l'algorithme Ridge pour l'analyse d'erreur.



我们的训练数据是蓝色的点，而测试数据是绿色的点。这张图显示了测试数据和训练数据之间的 residuals。我们看到，大部分测试数据的 residuals 都聚集在 0 附近，说明我们的训练模型比较贴合实际。

Our training data is a blue point, and the test data is a green point. This figure shows the residuals between the test data and the training data. We see that the residuals of most of the test data are clustered around 0, indicating that our training model is more realistic.

Nos données d'entraînement sont un point bleu et les données de test, un point vert. Cette figure montre les résidus entre les données de test et les données d'apprentissage. Nous constatons que les résidus de la plupart des données de test sont regroupés autour de 0, ce qui indique que notre modèle de formation est plus réaliste.