Prédiction de la qualité du vin

Master 2 Big Data et Fouille de Données

Université Paris 8 Vincennes Saint-Denis

Réaliser par : **Rolih DANGBO**

21/01/2020

Sommaire

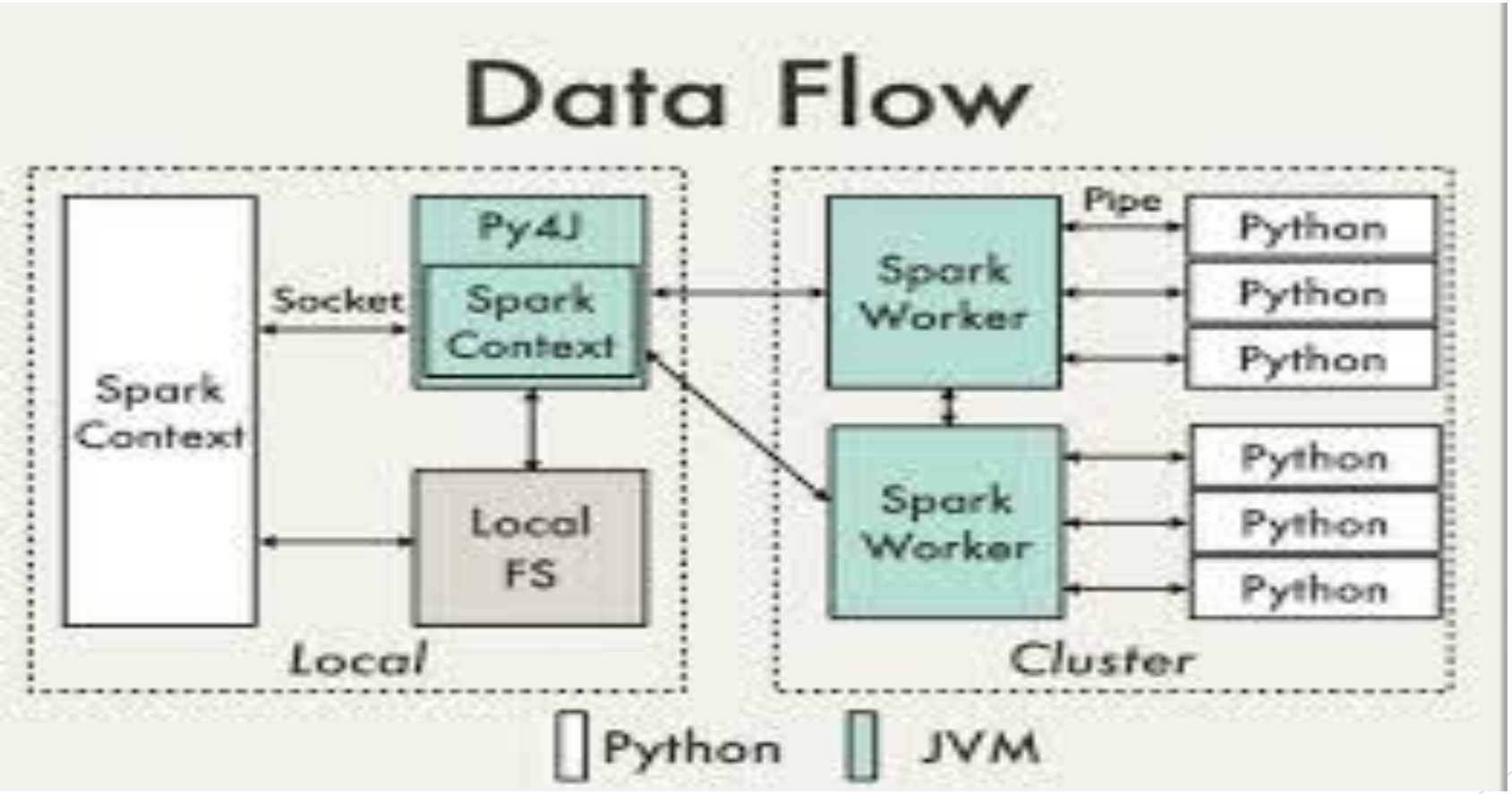
Objectifs

- Collecte des données
- Exploration des données
- Utilisation d’algorithmes de machine sur les données avec PySpark
- Expérimentations et résultats

Problématique

Comment Prédire la qualité du vin avec des algorithmes de machine learning supervisé?

Architecture de Spark



Description des données

- Nombre de lignes : 4898
- Nombre de colonnes : 12

Description des données

- fixed acidity : colonne qui décrit la teneur d'acidité fixe dans le vin.
- volatile acidity : colonne qui décrit la teneur d'acidité volatile dans le vin.
- citric acid : colonne qui décrit la teneur d'acide citrique contenu dans le vin.
- residual sugar : colonne qui décrit la teneur de sucre contenu dans le vin.
- chlorides : colonne qui décrit la teneur en chlorure de sodium contenu dans le vin.
- free sulfure dioxyde : colonne décrivant la teneur en dioxyde de souffre libre contenu dans le vin.
- total sulfure dioxyde : colonne décrivant la teneur en dioxyde de souffre total contenu dans le vin.

Description des données (suite)

- density : colonne décrivant la densité du vin.
- pH : colonne décrivant le ph du vin.
- sulphates : colonne décrivant la teneur en sulphates de sodium contenu dans le vin.
- alcohol : colonne décrivant la teneur d'alcool contenu dans le vin.
- quality : colonne décrivant la qualité du vin. Cette colonne va servir comme label pour la prédiction de la qualité du vin (moins bon à très bon).

Algorithmes utilisés

- Régression Linéaire
- Random Forest

Expérimentations et résultats

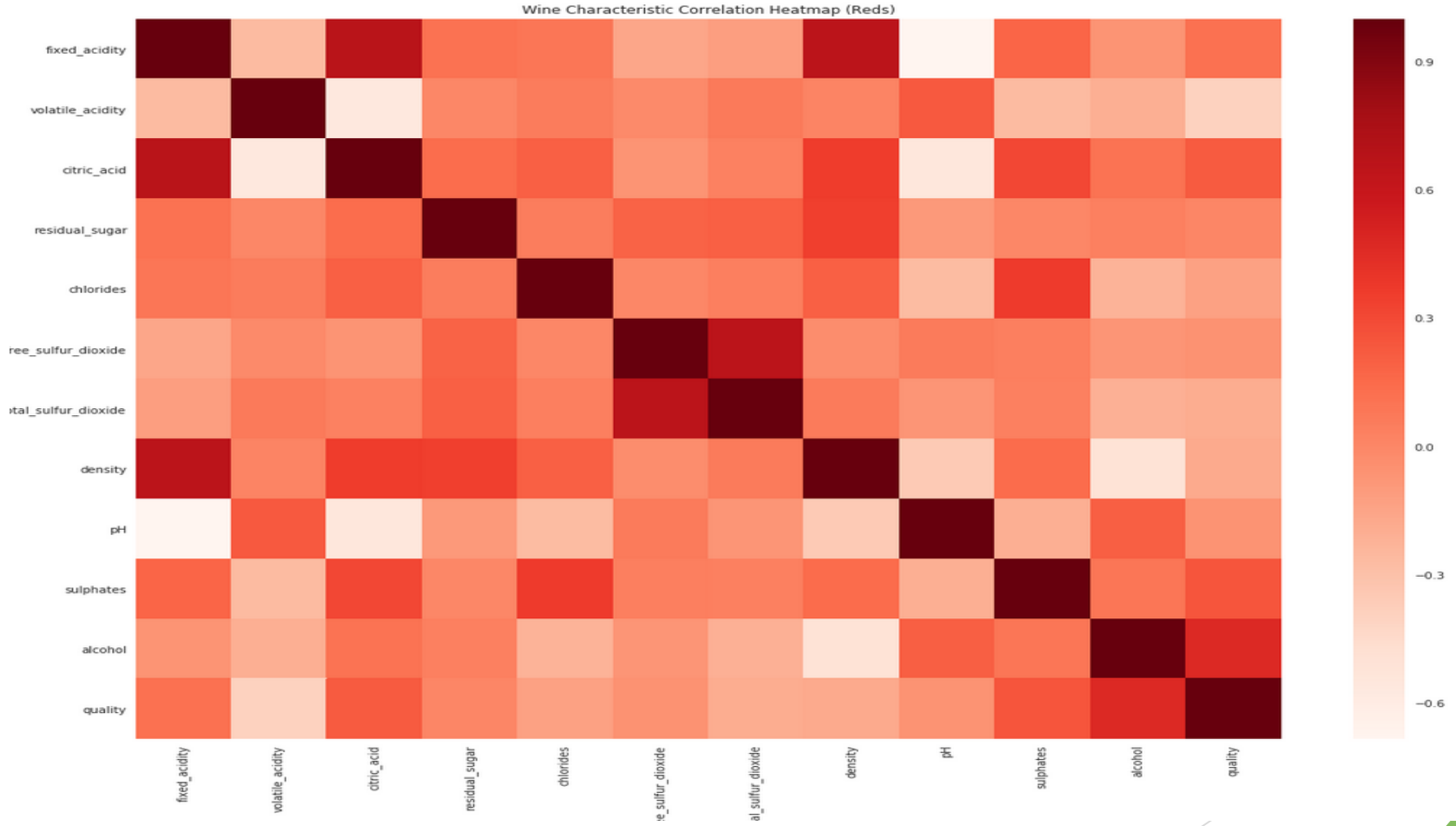
- Test de corrélations entre les variables

▸ (1) Spark Jobs

feature:	corcoef	cc_log	covar
========:	=======	======	=====
alcohol:	0.445	0.404	0.424
density:	-0.404	-0.372	-0.001
residual sugar:	-0.278	-0.255	-1.316
pH:	0.227	0.203	0.030
total sulfur dioxide:	-0.193	-0.169	-7.554
volatile acidity:	-0.181	-0.209	-0.017
chlorides:	-0.172	-0.157	-0.003
sulphates:	0.103	0.105	0.009
free sulfur dioxide:	-0.084	-0.062	-1.263
citric acid:	0.022	0.027	0.002

Expérimentations et résultats

- Test de corrélations entre les variables



Expérimentations et résultats

- Expérimentations et résultats

Régression Linéaire	Random Forest
Accuracy : 85 %	Accuracy : 64 %

Conclusion

- Montée en compétences sur l’utilisation des algorithmes de machine learning avec PySpark
- Evaluation des modèles