**Apache Spark – Mllib MLlib (librairie Machine Learning)**

Performed by : Amir Amemi

Houcem Gharbi

**1. Installation et Configuration d'Apache Spark**

**- Téléchargement d'Apache Spark et de Python et Java depuis le site officiel.**

<https://spark.apache.org/downloads.html>

A screenshot of a computer

Description automatically generated

<https://github.com/steveloughran/winutils/tree/master/hadoop-2.8.1>

A screenshot of a computer

Description automatically generated

<https://www.oracle.com/in/java/technologies/downloads/#jdk21-windows>

A screenshot of a computer

Description automatically generated

<https://www.python.org/downloads/>

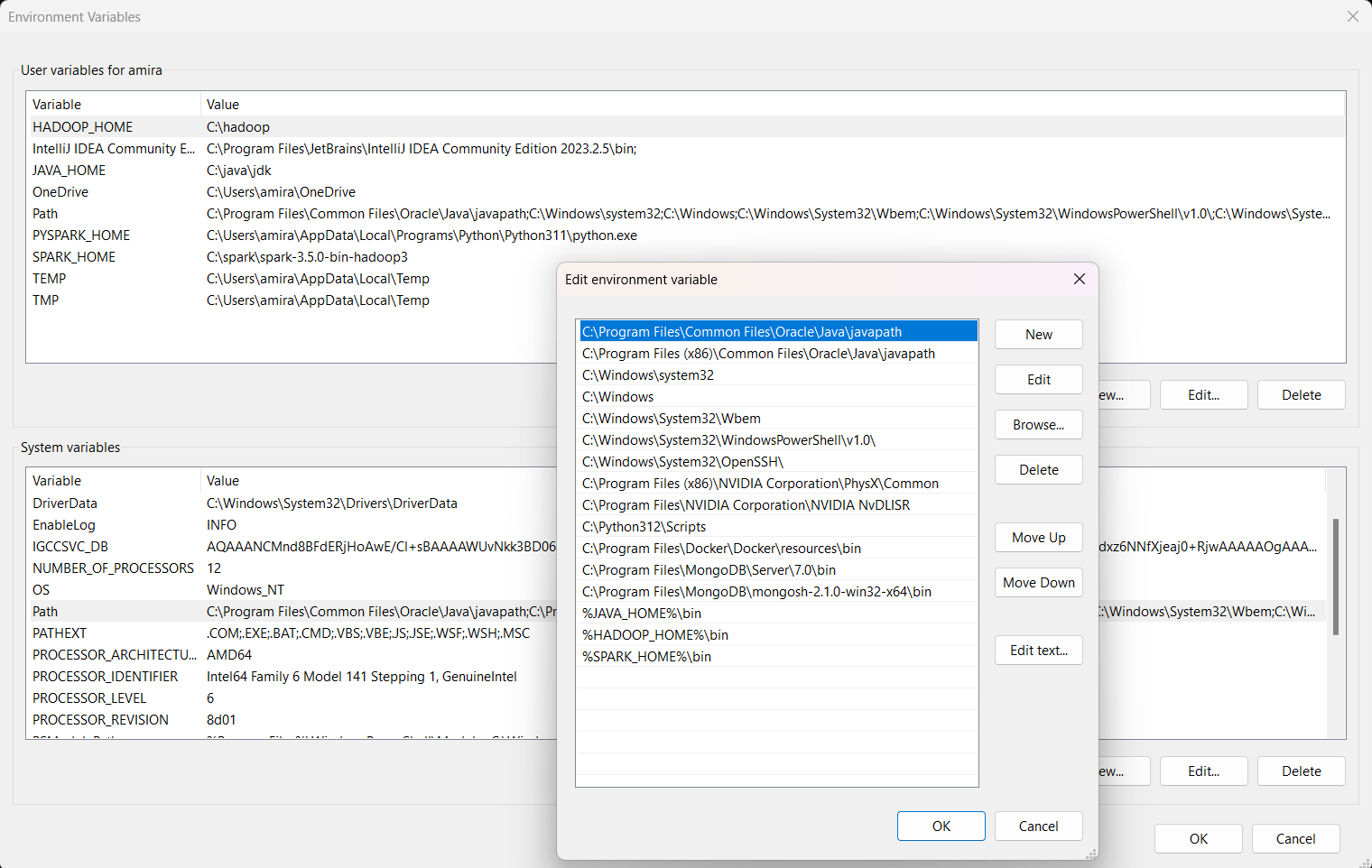
A screenshot of a computer

Description automatically generated

**- Extraction des fichiers et configuration des variables d'environnement.**

A screenshot of a computer

Description automatically generated



HADOOP\_HOME- C:\hadoop

JAVA\_HOME- C:\java\jdk

SPARK\_HOME- C:\spark\spark-3.5.0-bin-hadoop3

PYSPARK\_HOME- C:\path\python\..\python.exe

%SPARK\_HOME%\bin

%HADOOP\_HOME%\bin

%JAVA\_HOME%\bin

**- Vérification de l'installation avec l'exécution d'exemples simples.**

A screenshot of a computer

Description automatically generated

A computer screen shot of a black screen

Description automatically generated

**2. Préparation de l'Environnement**

# Create a Spark session

spark = SparkSession.builder.appName("CustomerSegmentation").getOrCreate()

**3. Chargement des Données**

# Load customer data

data = spark.read.csv("customer\_data.csv", header=True, inferSchema=True)

**4. Création de Caractéristiques**

# Feature engineering: Combine relevant features into a single vector

feature\_columns = ["TotalPurchaseAmount", "FrequencyOfPurchases", "AveragePurchaseValue"]

assembler = VectorAssembler(inputCols=feature\_columns, outputCol="features")

data = assembler.transform(data)

**5. Choix et Configuration du Modèle MLlib**

# Train a KMeans model

kmeans = KMeans(k=3, seed=42)

**6. Entraînement du Modèle**

model = kmeans.fit(data)

**7. Prédictions sur de Nouvelles Données**

# Make predictions

predictions = model.transform(data)

**8.Conclusion**

MLlib dans Apache Spark offre une solution puissante pour le développement de modèles de machine learning sur des données massives. En suivant ces étapes, vous pouvez mettre en œuvre avec succès des composants MLlib dans votre environnement Spark.