Fundamentos de

Big Data

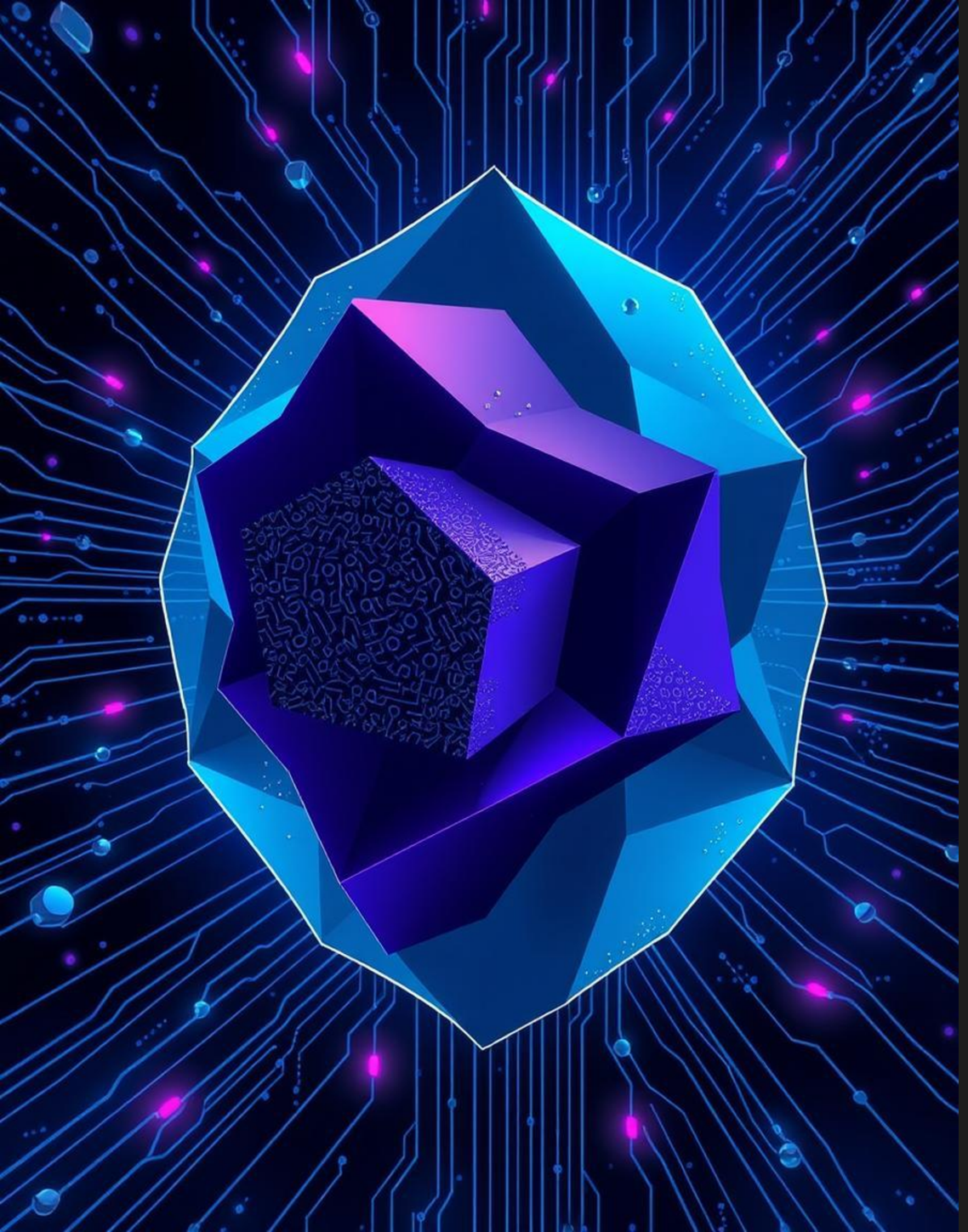
Sesión 5

¿Qué es MLlib?

Biblioteca de Machine Learning de Apache Spark.

Diseñada para Big Data y procesamiento distribuido.

Soporta tareas como clasificación, regresión, clustering y reducción de dimensionalidad.



Caracteristicas de MLlib

- ☑ **Escalabilidad:** Optimizado para entornos distribuidos
- ☑ **Alto rendimiento:** Computación distribuida sobre Apache Spark
- ☑ **Fácil integración:** Compatible con Python, Scala, Java y R
- ☑ **Conjunto de algoritmos optimizados**
- ☑ **Soporte para DataFrames y RDDs**

Herramientas Complementarias (MLlib Tools)

ML Pipelines: Modelos estructurados en pasos ordenados.

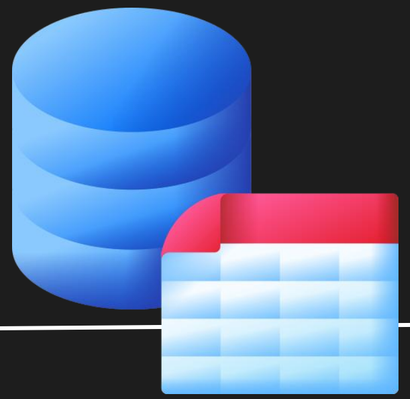
Hyperparameter Tuning: Cross-Validation y Grid Search

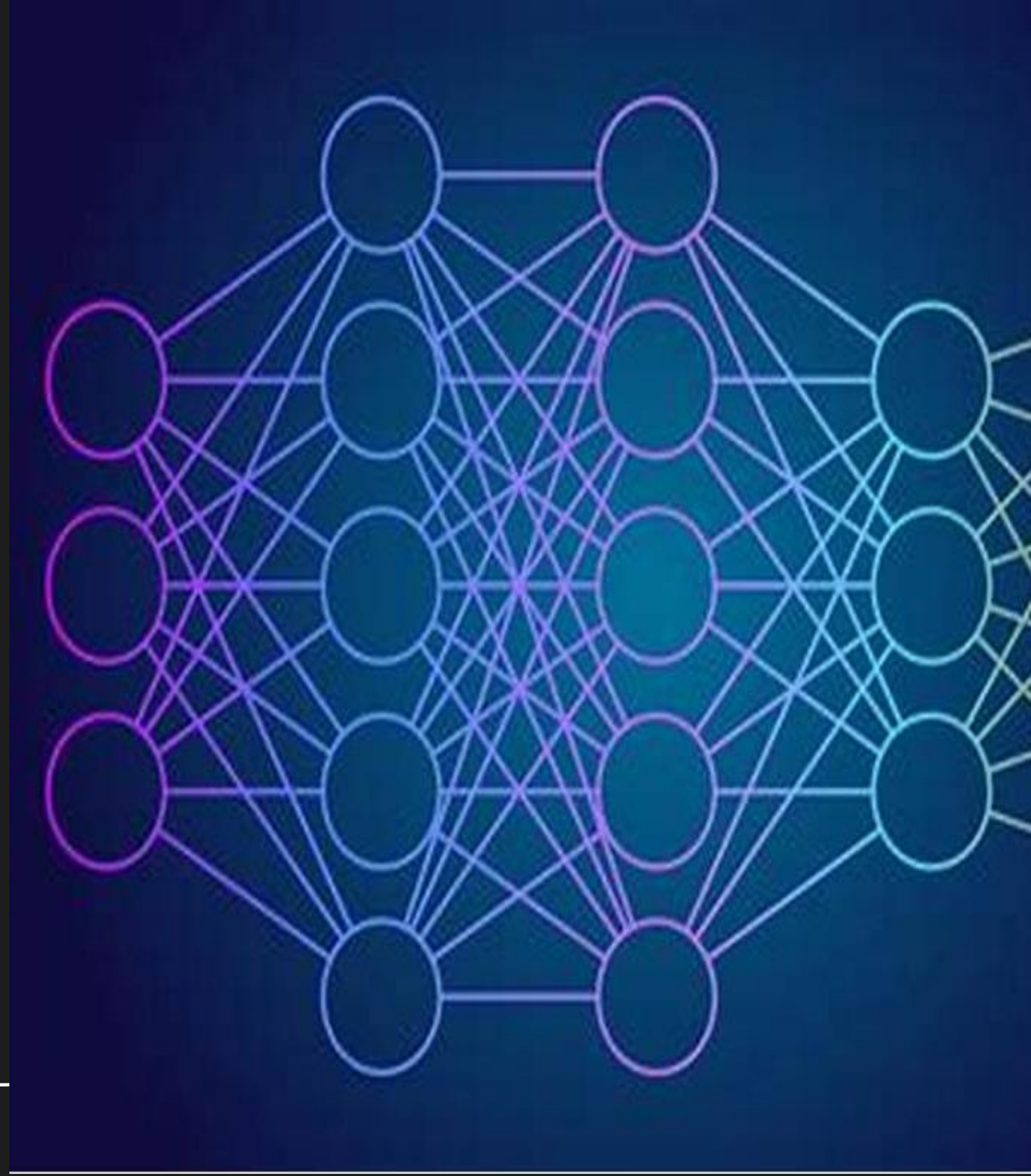
Persistencia de modelos: Guardar y reutilizar modelos entrenados.

Estructuras de Datos en MLlib

RDDs (Resilient Distributed Datasets)
•Estructura inmutable y distribuida
•Operaciones paralelas

DataFrames
•Optimización superior a los RDDs
•Manipulación de datos con SQL

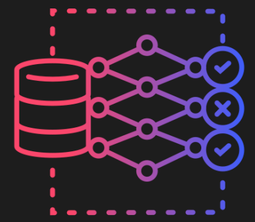




Algoritmos de Machine Learning en MLlib

Algoritmos Supervisados

- Regresión Lineal y Logística
- Árboles de Decisión y Random Forest
- Support Vector Machines (SVMs)
- Naïve Bayes



Algoritmos No Supervisados

- Clustering con K-Means
- LDA para modelado de temas
- PCA para reducción de dimensionalidad
- Filtrado Colaborativo

Implementación de Algoritmos Supervisados

Para entrenar modelos supervisados con MLlib, se siguen estos pasos generales:

1. **Carga de datos:** Se leen los datos en un DataFrame de Spark.
2. **Preprocesamiento:** Se convierten los datos en el formato adecuado, generalmente con *VectorAssembler*.
3. **División del dataset:** Se divide en **train** y **test**.
4. **Entrenamiento del modelo:** Se ajusta el modelo con el conjunto de entrenamiento.
5. **Evaluación:** Se mide el desempeño en el conjunto de prueba con métricas apropiadas.
6. **Predicción:** Se usan nuevos datos para obtener predicciones.

```python
from pyspark.sql import SparkSession
from pyspark.ml.classification import LogisticRegression
from pyspark.ml.feature import VectorAssembler

# Crear sesión de Spark
spark = SparkSession.builder.appName("MLlib_Supervised").getOrCreate()

# Cargar datos en un DataFrame
data = spark.read.csv("datos.csv", header=True, inferSchema=True)

# Transformar características en un solo vector
assembler = VectorAssembler(inputCols=["feature1", "feature2"], outputCol="features")
data = assembler.transform(data)

# Dividir en train y test
train, test = data.randomSplit([0.8, 0.2])

# Crear modelo de regresión logística
lr = LogisticRegression(featuresCol="features", labelCol="label")

# Entrenar modelo
model = lr.fit(train)

# Realizar predicciones
predictions = model.transform(test)

# Mostrar resultados
predictions.select("label", "prediction").show()
```

Implementación de Algoritmos No Supervisados

Los algoritmos no supervisados siguen un flujo similar, pero sin una variable objetivo. Se centran en **agrupar datos o reducir dimensiones**.

- ✓ Similar a los supervisados, pero sin etiquetas
- ✓ Modelos para descubrir patrones en los datos
- ✓ Ejemplo: **Clustering con K-Means**
- ✓ Reducción de dimensionalidad con PCA

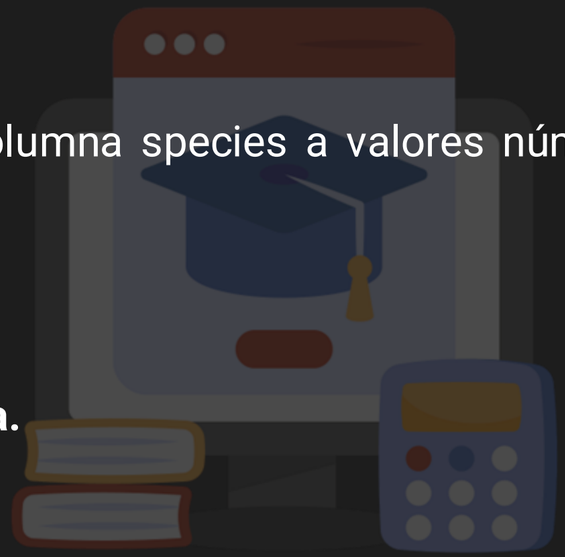
```python
from pyspark.ml.clustering import KMeans

# Crear modelo K-Means con 3 clusters
kmeans = KMeans(featuresCol="features", k=3)

# Entrenar modelo
model = kmeans.fit(train)

# Realizar predicciones de clusters
clusters = model.transform(test)

# Mostrar resultados
clusters.select("features", "prediction").show()
```

Actividad Práctica Guiada

Objetivo: Entrenar un modelo de Regresión Logística con Mllib en PySpark utilizando el dataset Iris, aplicando preprocesamiento, entrenamiento y evaluación del modelo.

Requisitos:

1. **Dataset:** dataset Iris.
2. **Preprocesamiento de datos:** Convertir columna species a valores números. Convertir características en un solo vector para Mllib.
3. **División en entrenamiento y prueba.**
4. **Entrenar el modelo de Regresión Logística.**
5. **Realizar predicciones.**
6. **Evaluar el modelo.**



El detalle de la actividad se encuentra en la guía de estudio de la sesión.

Preguntas

Sección de preguntas

Fundamentos de

Big Data

Continúe con las actividades