Spark Project

Important Notes

To run on a **local machine**, the SparkSession initializer must include the `.master("local")` line.

To run on the **hadoop cluster**, this line must be removed/commented out **before generating the JAR**.

Installing the program

The project includes configuration code for use with IntelliJ, including an easy way to compile and export the program to a JAR for use with a hadoop cluster.

Code for Logistic Regression can be found in `spark.logisticregression.LRMain`

Code for Decision Tree can be found in `spark.decisiontree.DTMain`

Static functions shared between the two classes can be found in `spark.HelperFunctions`

To generate a JAR file with IntelliJ

❗ ***Make sure `.master(local)` is commented out***

`Build -> Build Artifacts -> SparkProject -> Build`

This will generate a JAR file in `out/artifacts/SparkProject` named `SparkProject.jar`

Running on the hadoop cluster

`input` should be the directory where the data file `data.kdd` can be found in your hdfs.

`SparkProject.jar` should be the name of the jar file which contains both the `LRMain` and `DTMain` classes.

Logistic Regression

```
spark-submit --class spark.logisticregression.LRMain --master yarn --deploy-mode cluster SparkProject.jar input
```

Decision Tree

```
spark-submit --class spark.decisiontree.DTMain --master yarn --deploy-mode cluster SparkProject.jar input
```