Instructions on how to compile and execute the application.

The programming language used to develop the application is Python, so to use SPARK it is necessary to import PYSPARK. Python is an interpreted language so no compilation is needed. The format of the input is:

```
main.py [-h] [--LR] [--GLR] [--RF] [--GBT] [--FM] [--sample SAMPLE] path2data
```

The application needs to be called with the path2data argument. This option represents the path to a folder containing the csv files.

You can tell the application which models to use by the following optional arguments:

- --LR: Use Linear Regression algorithm
- --GLR: Use Generalized Linear Regression algorithm
- --RF: Use Random Forest algorithm
- --GBT: Use Gradient Boost Tree algorithm
- --FM: Use Factorization Machines Regression algorithm

If no model is selected, the linear regression model is chosen by default.

With --sample option, the user can specify the percentage of the data to use in the application.

To get more info about the meaning of the arguments and the proper usage of the application use the [-h] argument:

```
This application will try to learn to predict the arrival delay from data of fligths
usage: main.py [-h] [--LR] [--GLR] [--RF] [--GBT] [--FM] [--sample SAMPLE] path2data

positional arguments:
  path2data        Path to a folder (between quotes) containing the data as csv files

optional arguments:
  -h, --help       show this help message and exit
  --sample SAMPLE  Select just a fraction of the data, input between 0 and 1

Models:
  Select a specific models, if none is selected linear regression will be applied

  --LR             Use Linear Regression algorithm to predict the arrival delay of a flight
  --GLR            Use Generalized Linear Regression algorithm to predict the arrival delay of a flight
  --RF             Use Random Forest algorithm to predict the arrival delay of a flight
  --GBT            Use Gradient Boost Tree algorithm to predict the arrival delay of a flight
  --FM             Use Factorization Machines Regression algorithm to predict the arrival delay of a flight
```

The application can be executed by calling the python interpreter or by using spark-submit command:

```
> python.exe main.py [-h] [--LR] [--GLR] [--RF] [--GBT] [--FM] [--sample SAMPLE] path2data

> .\spark-submit main.py [-h] [--LR] [--GLR] [--RF] [--GBT] [--FM] [--sample SAMPLE] path2data
```

Instructions on where to place the input data.

As mentioned in the previous section, the application needs a path to a folder with the data in csv format.

The application checks if the folder exists, if it is not empty and if it contains any csv files. If all the conditions are met, the program reads all the csv files and stores them in a dataframe.

The path can be anywhere as long as the application can reach it, during the development of the application a folder on the same level as the script was used to store the input data.