Report on Higgs Boson Machine Learning Project

**Members:**

* Sergio Khalil 202101286
* Malak Mehrez 202006608
* Tamer AlBouz 202150560

This report provides a detailed overview of a Jupyter notebook that contains well-documented code for a machine learning project on Higgs Boson particle data that’s made of 600 thousand examples.

**Data Exploration**

The notebook starts by importing necessary libraries such as pandas, numpy, scikit-learn, keras, and tensorflow. It then loads the Higgs Boson dataset from a Google Drive folder and assigns column names to the data.

The notebook then proceeds to explore the data by checking its first few rows, retrieving its information and basic statistics, and checking its shape.

**Data Preparation**

During the data exploration phase, it was discovered that columns 8 and 21 have mixed data types although they should be float64. The values of these columns were checked for unexpected values that lead to having mixed data types. For column 8, the solution was to convert the string objects to numeric values using pandas' to\_numeric function. For column 21, the solution was to convert the string objects to numeric values using pandas' to\_numeric function and coerce any non-numeric values to NaN.

The data is then cleaned by removing training examples with NaN values.

**Model Selection**

The notebook then splits the dataset into training (80%) and testing (20%) sets using scikit-learn's train\_test\_split function. It also shows how to use only high-level features by selecting the last 7 columns of the dataset but their usage lead to a lower accuracy (0.73) in the neural network model and almost no effect on the other models.

The notebook then trains several machine learning models on the data, including linear regression, logistic regression, decision trees, XGBoost, and a neural network.

**Model Optimization**

The models are optimized by adjusting their hyperparameters. For example, the XGBoost model's hyperparameters such as n\_estimators, max\_depth, and learning\_rate are adjusted to improve its performance. The neural network model's architecture is also adjusted by changing the number of layers and units per layer.

**Model Evaluation**

The models are evaluated using metrics such as mean squared error and accuracy. The linear regression model achieved a mean squared error of 0.22. The logistic regression model achieved an accuracy of 0.64. The decision tree model also achieved an accuracy of 0.64. The XGBoost model achieved an accuracy of 0.7315. The neural network model achieved an accuracy of 0.7512.

**Conclusion**

In summary, this page is a well-documented Jupyter notebook that walks through the process of loading, exploring, cleaning, modeling and evaluating Higgs Boson particle data using various machine learning techniques. The notebook provides a clear and concise overview of the steps involved in building machine learning models for this type of data and highlights some of the challenges that can arise during the data preparation phase.

**GitHub Repository:** [**https://github.com/malakm2002/HIGGS\_Classification.git**](https://github.com/malakm2002/HIGGS_Classification.git)