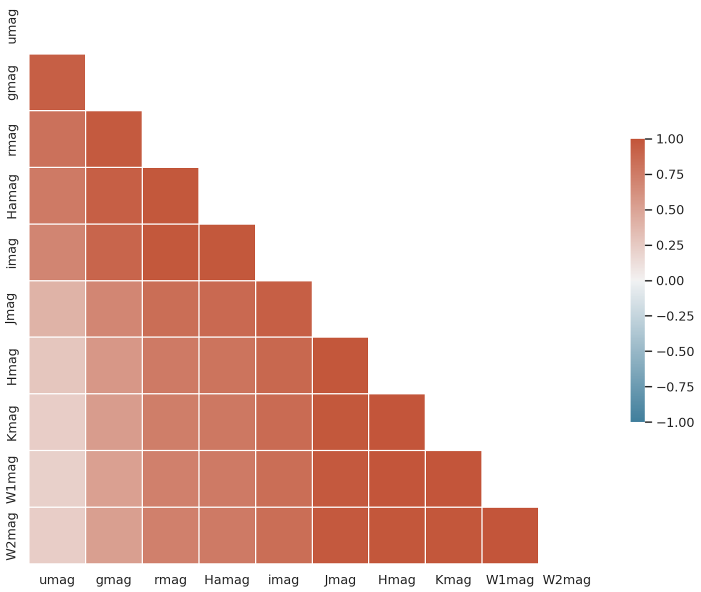

Figure 1: Correlation between features

1 Dataset

The Aidelman dataset contains 3.300.000 samples and each sample is labeled as an Hα emitting object (EM) and/or a Be star (Be). Each sample includes measurements from 10 different channels: u, g, r, Hα, i, J, H, K, W1 and W2.

2 Experiments

2.1 EM classification

In order to establish a baseline performance for classifying EM objects, we split the dataset 80/20, obtaining training and test subsets, and we train three common models using SKLearn:

- Random Forest, with a max depth of 6
- 3-layer neural network, with (10,10) hidden states
- Gradient Boosting, with 300 tree estimators

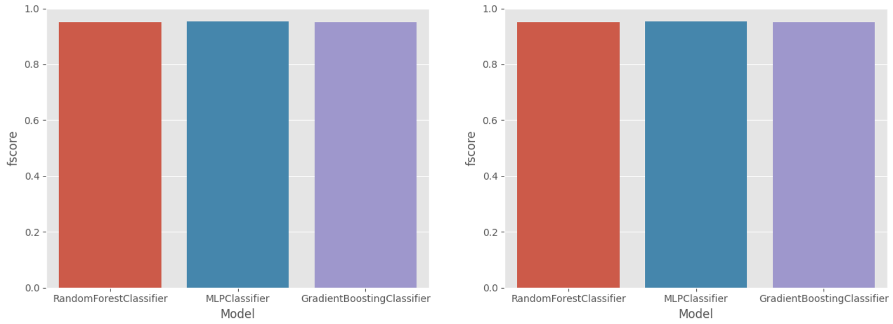

Figure 2: F-Score for train (left) and test (right) sets when classifying stars as EM or not.

```
Train:                         Model    fscore  precision    recall
0      RandomForestClassifier  0.951191   0.908095  0.998581
1               MLPClassifier  0.954169   0.919153  0.991958
2  GradientBoostingClassifier  0.951439   0.911145  0.995461
Test:                         Model    fscore  precision    recall
0      RandomForestClassifier  0.951387   0.908369  0.998682
1               MLPClassifier  0.954212   0.919330  0.991846
2  GradientBoostingClassifier  0.951507   0.911367  0.995345
```