Introduction

This report describes what has been done to create project. Code repository: https://gitlab.cs.ttu.ee/totahv/iti8565
Bigger images are available in code repository under directory */project/img.* SQL queries in the file /project/*queries.sql*
Sloan Digital Sky Survey (SDSS) is a database that holds imaging and spectroscopic data of night sky. Data is collected from 2.5 m wide-angle optical telescope located in New Mexico, US and covers 35% of the sky. Data used in current project was collected in July 2017, it is available through SDSS website and can be queried using SQL statements.

Framing the problem

Goals of project: explore Sloan Digital Sky Survey (SDSS) database, create classification model for celestial objects, create classification model for spectral type of star and select most accurate machine learning algorithms for tasks.
Searching answers for two problems:
1. What features are necessary to predict type of celestial object. Three available types are: quasars, galaxies and stars.
2. What features are necessary to predict spectral type of a star. There are a lot of spectral types, for example A0, G5 etc.

Data collection

Features have been explored and selected from two views of the SDSS database: SpecObj and PhotoObj. Table 1 describes selected features. Total 52 features were selected. 'Columns' means that feature is split between N columns.

Table 1 Selected features

Nr	Description of feature	Used in model	Columns
1	Redshift	Celestial and star spectral type classifiers	1
2	J2000 Right Ascension and Declination, location of stars	Celestial and star spectral type classifiers	2
3	Astronomical magnitude system, 5 bands of telescope	Celestial type classifier	5
4	De Vaucouleurs fit b/a, describes ellipticity of object	Celestial type classifier	5
5	De Vaucouleurs fit scale radius or the effective radius	Celestial type classifier	5
6	Adaptive fourth moment of object	Celestial type classifier	5
7	Petrosian radius	Celestial type classifier	5
8	Petrosian flux	Celestial type classifier	5
9	Petrosian magnitude	Celestial type classifier	5
10	PSF magnitude	Celestial type classifier	5
11	Spectrum projected onto filters	Celestial type classifier	5
12	(B-V) color	Star spectral type classifier	1
13	Effective temperature	Star spectral type classifier	1
14	log10(gravity)	Star spectral type classifier	1
15	Metallicity ([Fe/H])	Star spectral type classifier	1

Data preparation

For celestial type classifier Pearson correlation plot was drawn and a lot of features were correlated, so some were removed. Figure 1 describes correlations before removal. Figure 2 shows counts of each celestial object type in data set. Data is biased since there are a lot more galaxies than stars and quasars. Figure 3 describes that redshift has biggest impact on determining celestial object type, but some other features remain in model, because they increase accuracy.

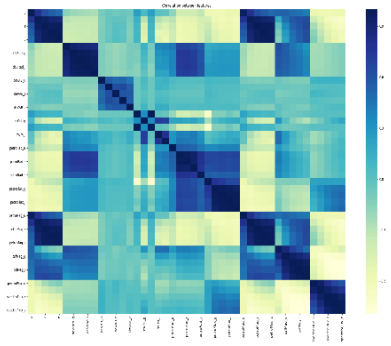

Figure 1 Correlation plot before feature removal

Figure 2 Celestial object count

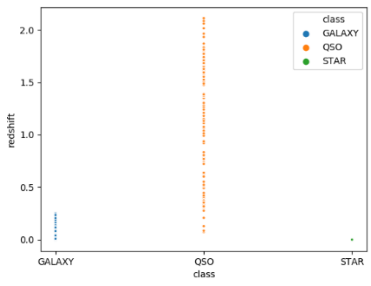

Figure 3 Redshift and celestial object type relationship

Figure 4 visualizes pair plot between selected features for celestial type classifier. Figure 5 visualizes data set after using PCA, dimensionality was reduced to 2 components. Blue are stars, green are galaxies and orange are quasars. It visualizes that it is possible to classify celestial objects using selected features.

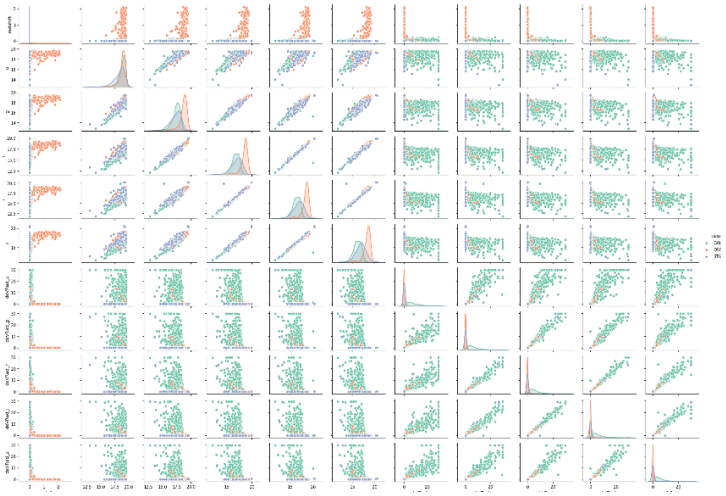

Figure 4 Pair plot of features for celestial type classifier

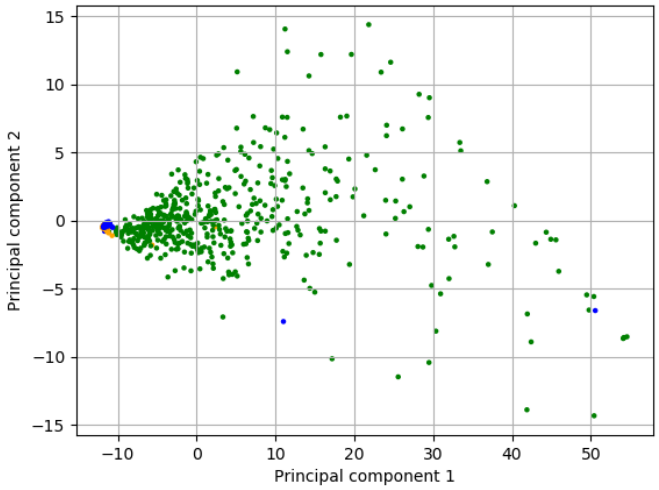

Figure 5 PCA with 2 components for celestial type classifier

For star spectral type, Figure 6 describes Pearson correlation plot where color, temperature, metallicity and log gravity are correlated, likely derived values, so they will be removed. Figure 7 uses random forest classifier to visualize importance of selected features in tree model. Redshift has almost no impact on spectral type of star and will also be excluded. Figure 8 shows relationship between spectral type and temperature and it can be concluded that one feature (temperature) is enough to predict star spectral type.

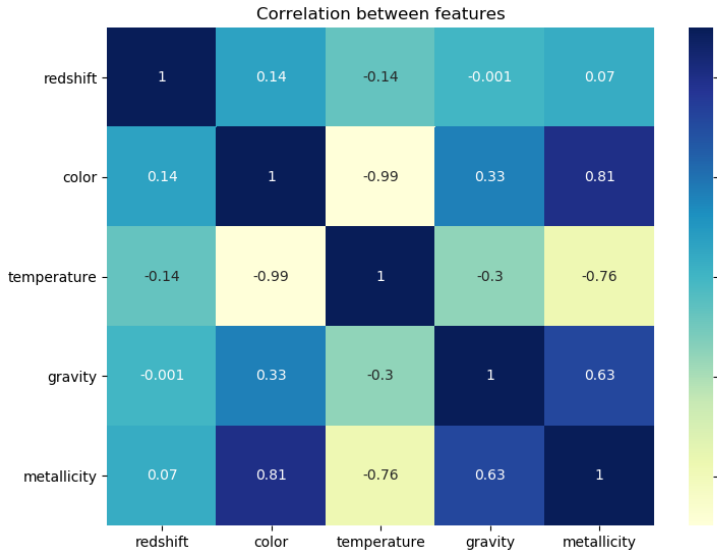

Figure 6 Star spectral type feature correlation

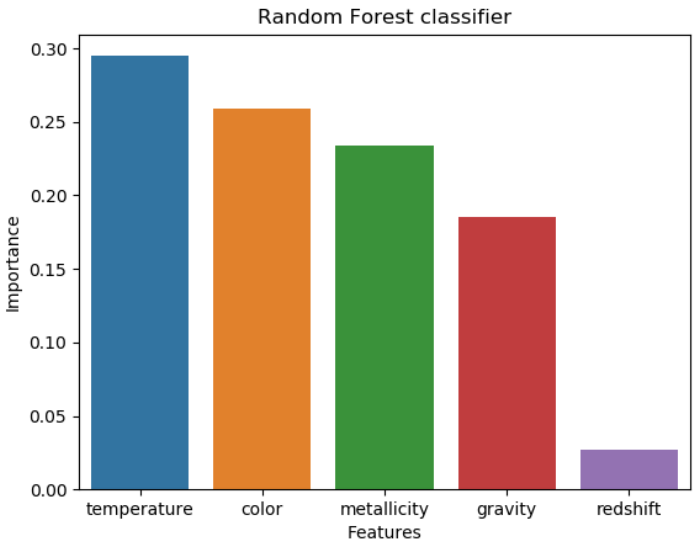

Figure 7 Star spectral type feature importance

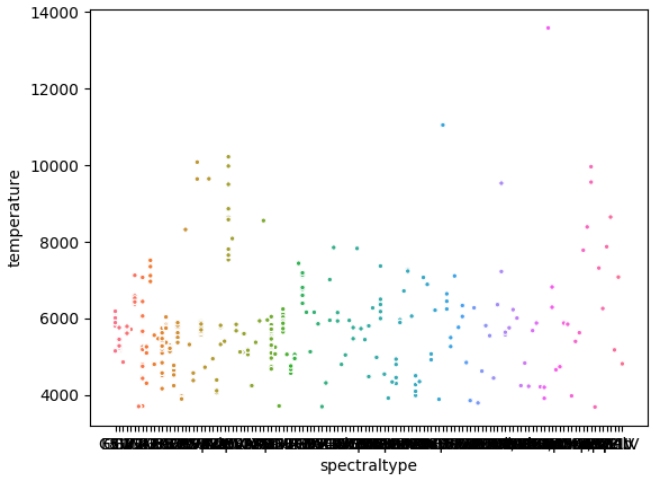

Figure 8 Star spectral type and temperature relationship

Total 15 features remained in data set and they are described in Table 2.

Table 2 Features remain in data after feature selection

Nr	Description of feature	Used in model	N
1	Redshift	Celestial type classifier	1
2	Astronomical magnitude	Celestial type classifier	5
3	Effective radius	Celestial type classifier	5
5	Effective temperature	Star spectral type classifier	1

Training models

Algorithms used for supervised training: Decision Tree, Random Forest, AdaBoost, KNN, SVM and Multilayer Perceptron (2 hidden layers with 1000 neurons and 1000 iterations). For unsupervised training: K-means and DBSCAN. Data was split 30/70 for training and test data. For training model 5-fold cross-validation was used and reported accuracy is mean accuracy.

Evaluation

Celestial object classifier was trained on 7000 objects and results are in Table 3. Star spectral type classifier was trained on ~6000 stars and results are in Table 4. Results tell that decision tree and random forest are fast and accurate for current task.

Table 3 Celestial object type classifier cross-validation result

No	Algorithm	Accuracy	Time (s)
1	Decision Tree	0.987000	0.310476
2	KNN	0.982999	0.523816
3	Random Forest	0.991142	0.762329
4	AdaBoost	0.943297	4.311504
5	SVM	0.983286	2.411332
6	MLP	0.987285	893.894166

Table 4 Star spectral type classifier cross-validation result

No	Algorithm	Accuracy	Time (s)
1	Decision Tree	0.978090	0.195611
2	KNN	0.963582	0.445589
3	Random Forest	0.975956	0.414152
4	AdaBoost	0.224313	9.242380
5	SVM	0.976602	1.965495
6	MLP		

For unsupervised learning used K-means and DBSCAN. Applied them on celestial object clustering. Figure 9 and Figure 10 describe clustering results after PCA. Figure 5 shows original classes and looks different, so clustering is not accurate.

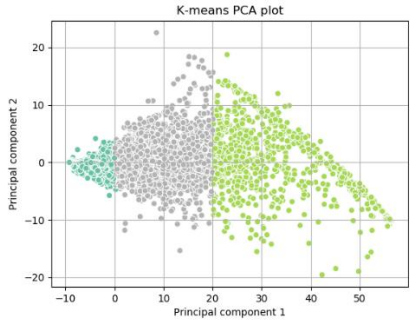

Figure 9 K-means clustering result after PCA

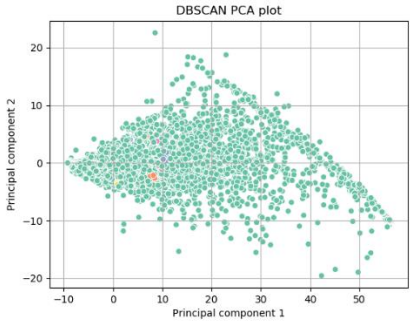

Figure 10 DBSCAN clustering result after PCA

Best model for celestial object classification is Random Forest and further validations are based on that. Figure 11 shows test data classification results, it looks very similar to Figure 5. Table 5 shows confusion matrix and Table 6 classification report.

Figure 11 Test data result for celestial object type classifier

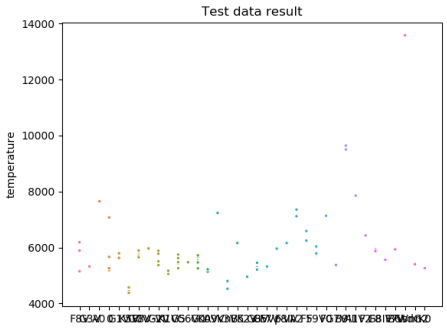

Figure 12 Test data result for star spectral type classifier

Table 5 Confusion matrix for celestial object classifier

	GALAXY	QSO	STAR
GALAXY	1563	8	4
QSO	12	286	0
STAR	4	0	1123

Table 6 Classification report for celestial object classifier

	precision	recall	F1-score	support
GALAXY	0.99	0.99	0.99	1575
QSO	0.97	0.96	0.97	298
STAR	1.00	1.00	1.00	1127

Table 7 Classification report for star type classifier

	precision	recall	F1-score	support
avg	0.99	0.98	0.98	28565

Best model for star spectral type classification was also Random Forest and further validations are based on that. Figure 12 shows test data classification result and it looks very similar to Figure 8. Complete confusion matrix and classification report are difficult to present because there are 141 unique spectral types, so only weighted average result is displayed in Table 7.

Conclusion

Two models were created based on data from SDSS database.

First model is using Random Forest to predict celestial object type using redshift, effective radius and telescope 5-band spectrum. Results are in accord with Hubble's law, which states that redshift of objects in space is proportional with their distance from Earth.

Second model is using Decision Tree to predict stars spectral type and after working out data only one feature (temperature) is enough for prediction, which is in accord with Hertzsprung-Russell diagram (main-sequence stars), which is used to categorize stars by their temperature, color, luminosity and absolute magnitude.

Unsupervised algorithms DBSCAN and K-means failed to classify celestial object type the way humans do.

Used libraries

NumPy, Pandas, Matplotlib, seaborn, scikit-learn

Sources

[1] SDSS http://skyserver.sdss.org/dr15/en/tools/search/sql.aspx
[2] Kaggle https://www.kaggle.com/apoorvakesarwani/sloan-classification-algorithms
[3] Kaggle https://www.kaggle.com/karnar95/is-that-a-star-galaxy-nah-it-s-quasarrr
[4] Photometric system https://en.wikipedia.org/wiki/Photometric_system
[5] Effective radius https://en.wikipedia.org/wiki/Effective_radius