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 views SpecObj and PhotoObj, create first model for classification of celestial objects, create second model for classification of spectral type of star and select most accurate machine learning algorithms for tasks.
Project is split into two parts, 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.

1. Celestial object type classifier

Data collection

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	Columns
1	Redshift	1
2	J2000 Right Ascension and Declination, location of stars	2
3	Astronomical magnitude system, 5 bands of telescope	5
4	De Vaucouleurs fit b/a, describes ellipticity of object	5
5	De Vaucouleurs fit scale radius or the effective radius	5
6	Adaptive fourth moment of object	5
7	Petrosian radius	5
8	Petrosian flux	5
9	Petrosian magnitude	5
10	PSF magnitude	5
11	Spectrum projected onto filters	5

Data collection

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.

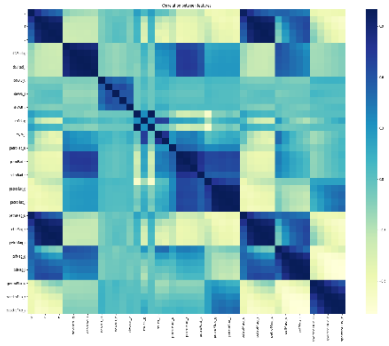

Figure 1 Correlation plot before feature removal

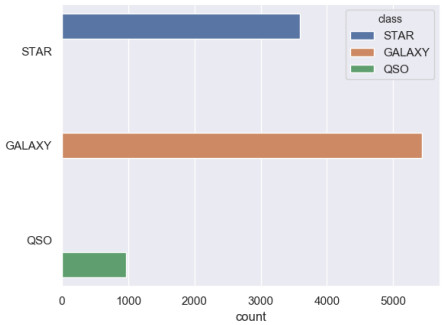

Figure 2 Celestial object count

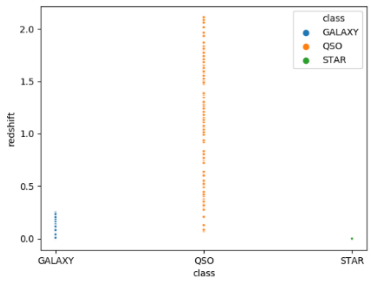

Figure 3 Redshift and celestial object type relationship

Figure 4 visualizes data set after using PCA, dimensionality was reduced to 2 components. Blue are stars, orange are quasars, green are galaxies. Blue and orange are forming two lines close to each other, but they can be separated. Green points are all spread out. It visualizes that it is possible to classify celestial objects using selected features. Total 12 features remained in data set and they are described in Table 2.

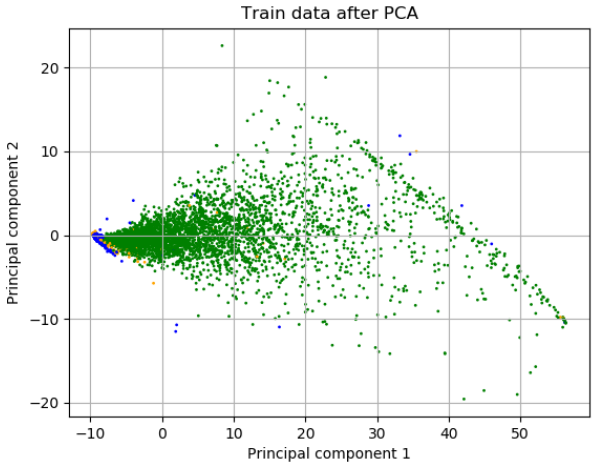

Figure 4 PCA with 2 components for celestial type classifier

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

Training models

Algorithms used for supervised training: Decision Tree, Random Forest, AdaBoost, KNN, SVM and Multilayer Perceptron (2 hidden layers with 100 neurons and 1000 iterations). For unsupervised training: K-means and DBSCAN. Data was split 30/70 for training and test data. While training model 5-fold cross-validation was used and reported accuracy is mean accuracy.

Evaluation

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 4 shows original classes and looks very different, so clustering is not accurate.

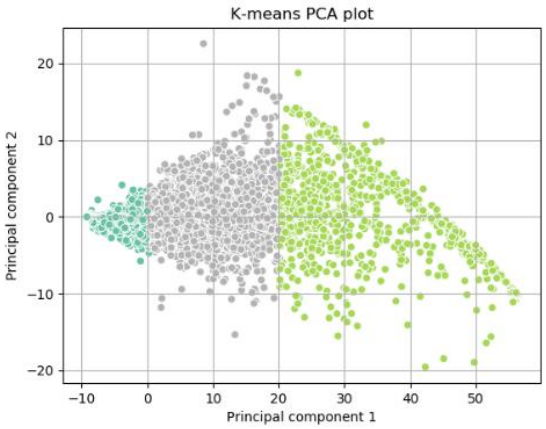

Figure 5 K-means clustering result after PCA

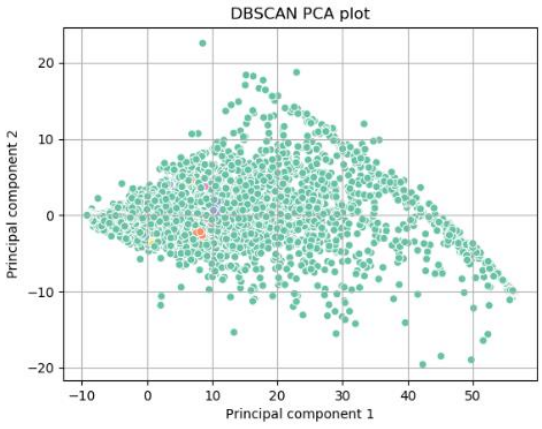

Figure 6 DBSCAN clustering result after PCA

For supervised learning celestial object classifier was trained on 7000 objects and results are in Table 3. MLP offers the same accuracy as shallow methods, but is very slow, so it is not worth using it. 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	0.287408
2	KNN	0.982999	0.353325
3	Random Forest	0.991142	0.519779
4	AdaBoost	0.943297	3.304573
5	SVM	0.983286	1.840783
6	MLP	0.983713	20.478998

Selected model for celestial object classification is Random Forest and further validations are based on that. Figure 7 shows test data classification results, it looks like Figure 4, but not exactly. Galaxies are spread out the same. Stars and quasars have different shapes but are in the same region. Table 5 shows confusion matrix and Table 6 classification report.

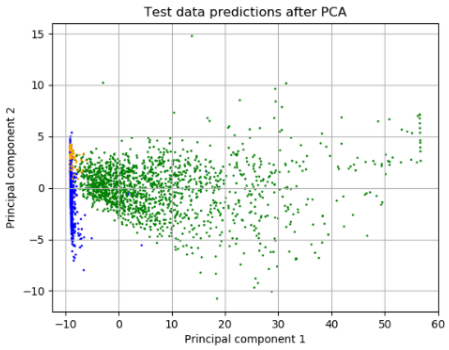

Figure 7 Test data result for celestial object type classifier after PCA

Table 4 Confusion matrix for celestial object classifier

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

Table 5 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

2. Star spectral type classifier

Data collection

Table 1 describes selected features. Total 52 features were selected. 'Columns' means that feature is split between N columns.

Table 6 Selected features for star spectral type classifier

Nr	Description of feature	Columns
1	Redshift	1
2	J2000 Right Ascension and Declination, location of stars	2
3	(B-V) color	1
4	Effective temperature	1
5	log10(gravity)	1
6	Metallicity ([Fe/H])	1

Data collection

For star spectral type, Figure 8 describes Pearson correlation plot where color, temperature, metallicity and log of gravity are correlated, likely derived values, so they will be removed. Figure 9 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.

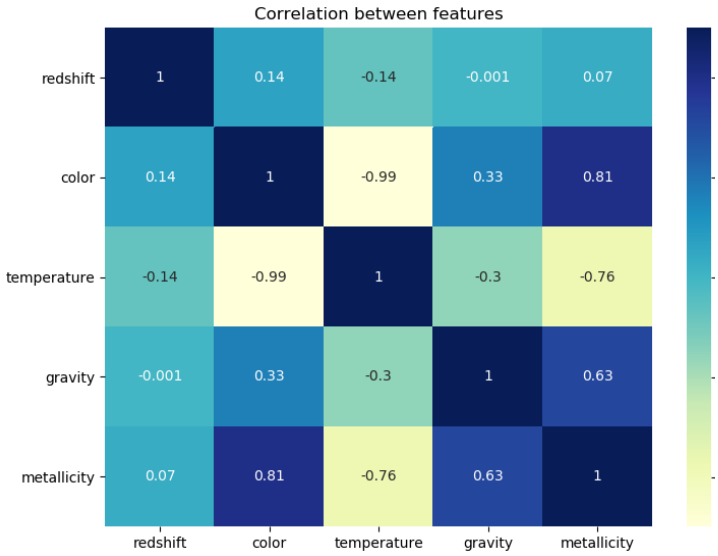

Figure 8 Star spectral type feature correlation

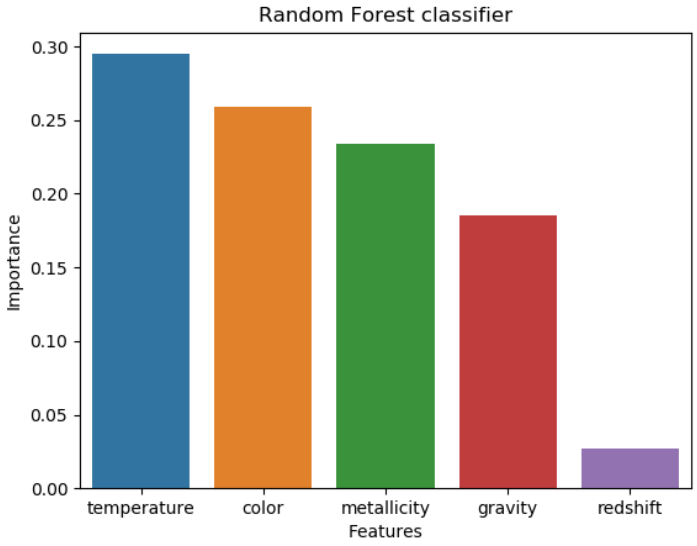

Figure 9 Star spectral type feature importance

Figure 10 shows relationship between spectral type and temperature. It is filtered visualization, where filter is minimum 100 stars in each spectral type, so it would be possible to visualize. On the plot there are 31 unique spectral types and 158 unique temperatures. In complete training data set there are 92 unique spectral types and 254 unique temperature values. Currently, it is impossible to say if it is possible correctly classify stars, so we'll try it in next step.

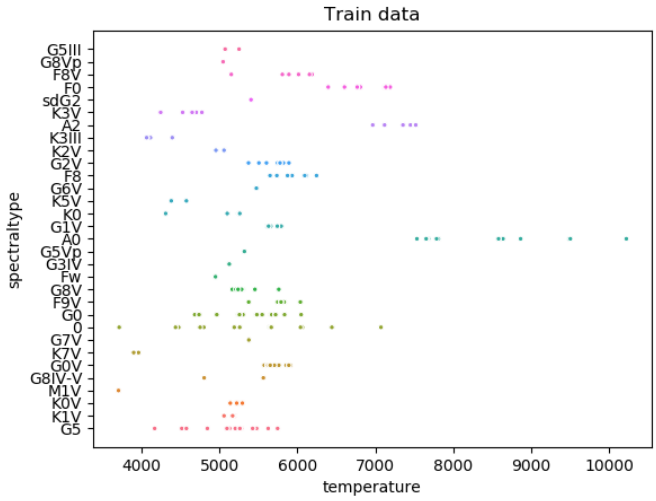

Figure 10 Star spectral type and temperature relationship

Table 2 describes features used in star classifier. Only temperature remains.

Table 7 Features remain in data after feature selection

Nr	Description of feature	Used in model	N
1	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 100 neurons and 1000 iterations). For unsupervised training: K-means and DBSCAN. Data was split 30/70 for training and test data. While training model 5-fold cross-validation was used and reported accuracy is mean accuracy.

Evaluation

Star spectral type classifier was trained on ~6100 stars and results are in Table 8. Results tell that decision tree and random forest are fast and accurate for current task.

Table 8 Star spectral type classifier cross-validation result

No	Algorithm	Accuracy	Time (s)
1	Decision Tree	0.978090	0.165846
2	KNN	0.963582	0.297680
3	Random Forest	0.975956	0.332062
4	AdaBoost	0.224313	7.547534
5	SVM	0.976602	1.588649
6	MLP	0.066284	72.604804

Figure 11 shows filtered visualization (spectral types with over 100 stars in them). Also, hard to tell if classification is correct, but reported model accuracy is very high. I think it possible to predict so accurately because there are limited number of unique temperatures and some spectral types have only one unique star/temperature, so decision tree works very well in this case. If temperature numbers were continuously spread out, then I don't think models would be so accurate.

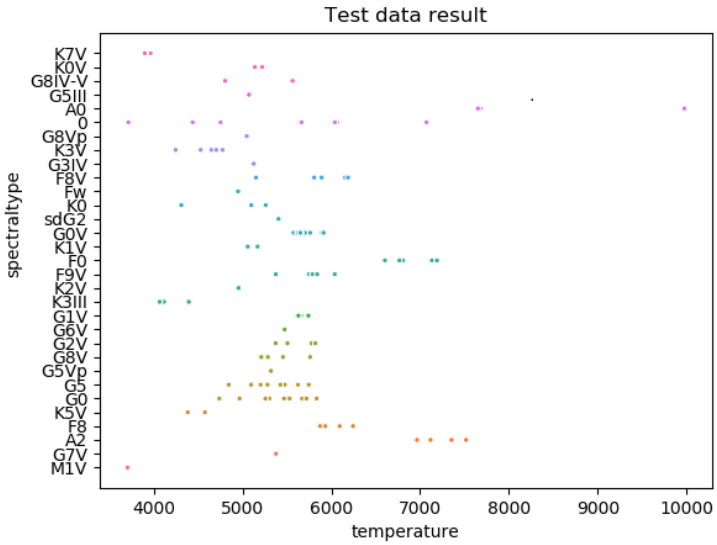

Figure 11 Test data result for star spectral type classifier

Best model for star spectral type classification was Decision Tree and further validations are based on that. Complete confusion matrix and classification report for star classification are difficult to present because there are 91 unique spectral types, so only weighted average result is displayed in Table 9.

Table 9 Classification report for star type classifier

	precision	recall	F1-score	support
avg	0.98	0.98	0.98	2570

Conclusion

Two models were created based on data from SDSS database.

First model (classification of celestial object type) building started with selecting a lot of features that seemed to have potential impact on classification. After drawing correlation plot, it was apparent that a lot of features are correlated and to be removed. Other features had close to no impact on accuracy, so they were removed too. So, my guess was mostly wrong. After training model, decision trees and random forest were most accurate and fastest algorithms for given task. AdaBoost can classify two classes, so for three classes it is not reliable. MLP performed good, but it is too slow and not worth using. Biggest impact on celestial object type had redshift and it is in accordance with Hubble's law, which states that redshift of objects in space is proportional with their distance from Earth. It means that stars are closest, galaxies are at average distance, quasars are brightest objects and very far away. Measuring redshift of those objects is the way of measuring distance and therefore a way to classify them. Therefore, classification works very well here. Best model for celestial object type classification is Random Forest, because it has highest accuracy.

Second model (classification of star spectral type) building started with selecting features for classification, but selected features were correlated and probably derived from each other, so they were removed. Redshift had almost no impact on classification of stars, because their distances are all similar, so only temperature remained for training. After training model, AdaBoost and MLP performed very poorly. MLP performed poorly because there were a lot of classes and little stars in them (some classes had one star in them), so it suffered from lack of data. Best model for star spectral type is Decision Tree, because it is most accurate. Predicting spectral type of star using temperature is in accordance with Hertzsprung-Russell diagram (main-sequence stars), which is used to categorize stars by their temperature, color, luminosity and absolute magnitude.

Unsupervised algorithms K-means and DBSCAN are using centroids and density for clustering. Classification in this project does not depend on density or centroids of data, so DBSCAN and K-means clustered celestial objects a lot differently than humans and results are not comparable.

There are a lot of improvements to be made. Starting with getting to know SDSS database better and understand what each feature exactly is and how it is derived. Need to study a lot of astrophysics to get theoretical background on how to analytically derive relationships between features. Then it would be possible to pick more features for classification and increase accuracy of model. Initial data is extracted using image recognition, so it would be interesting to use classification on actual images. Also extracting and deriving values like redshift, metallicity, temperature etc. from images would be great challenge, but all that is outside scope of this project.

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