DARIA KUDOIAR
dariakudoiar@gmail.com

QUESTION 1. SVM classification for Real Data: Asteroids Data Set

STEP 1

Practical goal of the original task was determining the highest risk of an asteroid impact and how impact predictions change over time. In addition, which possible asteroid impact would be the most devastating, given the asteroid's size and speed. Originally the Asteroids data was collected by NASA's Near-Earth Object Program at the Jet Propulsion Laboratory (California Institute of Technology) along with Impact risk data (which we are not going to use in our classification task).

Collected Asteroids data represents 15636 cases and 15 columns (14 features per case and 1 target column).

There are 8 original Asteroid`s group (classes). We will keep three of them.

Name of class (Asteroid`s group)	**Size of class**
Apollo Asteroid	6940
Amor Asteroid	5918
Apollo Asteroid (Hazard)	1520
Aten Asteroid	987
Aten Asteroid (Hazard)	155
Amor Asteroid (Hazard)	99
Apohele Asteroid	11
Apohele Asteroid (Hazard)	5

DARIA KUDOIAR
dariakudoiar@gmail.com

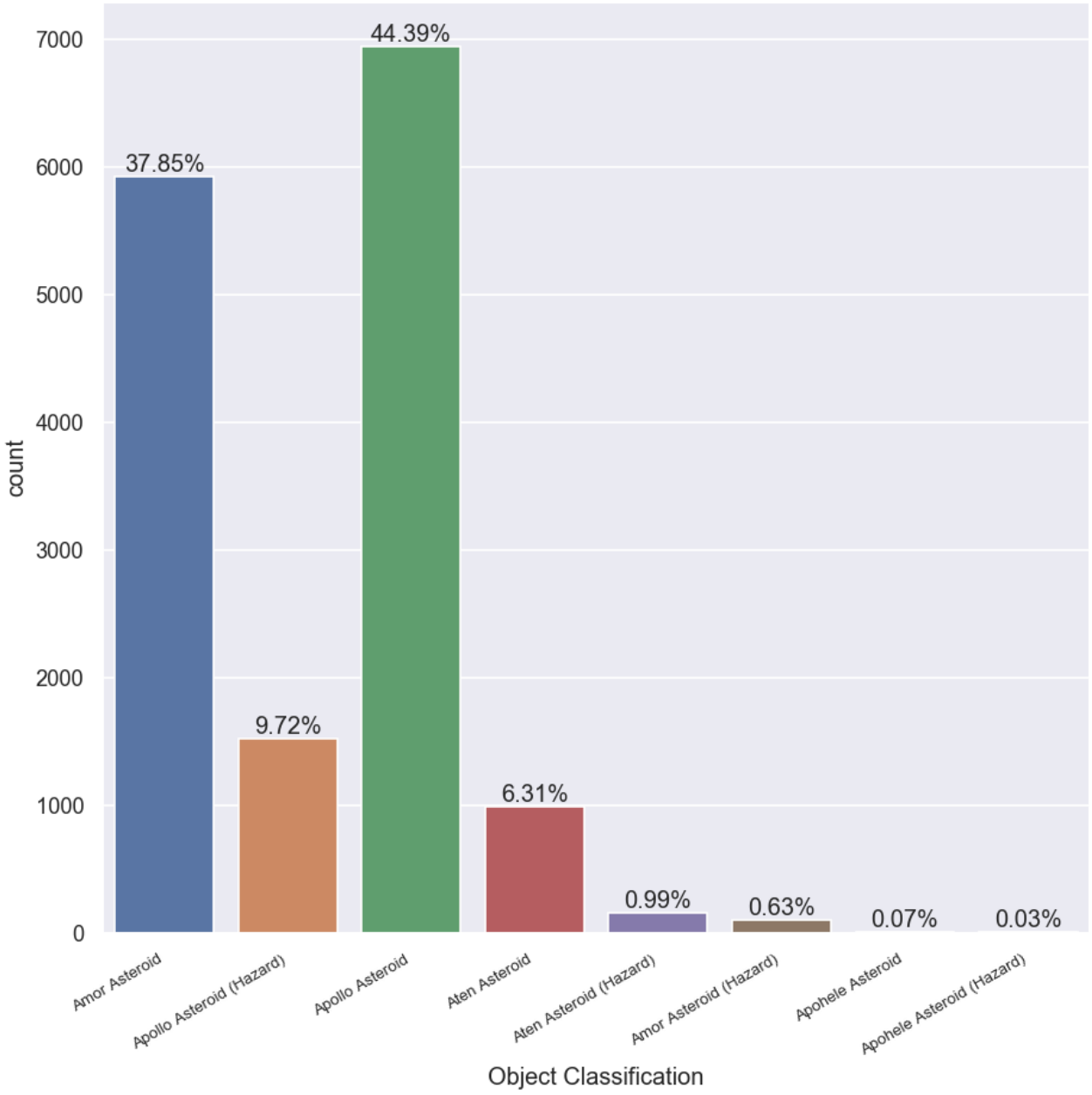

Figure 1: Count plot for Original Data Set

DARIA KUDOIAR
dariakudoiar@gmail.com

Data columns descriptions

Name of feature	**Meaning**	**Discrete or Continuous feature**
Object Name	Unique name of asteroid (Catalogue name)	discrete (number of unique values: 15635)
Object Classification	Classification of asteroid by their characteristic, Asteroid`s group (target column)	continuous
Epoch (TDB)	Moment in time used as a reference point for the epoch of osculation of the orbital elements of asteroid	continuous
Orbit Axis (AU)	Orbital characteristic (a line perpendicular to the plane of the orbit at the foci of that orbit)	continuous
Orbit Eccentricity	Dimensionless parameter that determines the amount by which its orbit around another body deviates from a perfect circle	continuous
Orbit Inclination (deg)	Measures the tilt of an object's orbit around a celestial body. It is expressed as the angle between a reference plane and the orbital plane or axis of direction of the orbiting object	continuous
Perihelion Argument (deg)	The point of closest approach between the orbiting body (e.g. a planet) and the focus	continuous

DARIA KUDOIAR
dariakudoiar@gmail.com

Node Longitude (deg)	One of the orbital elements used to specify the orbit of an object in space. It is the angle from a reference direction, called the origin of longitude, to the direction of the ascending node, measured in a reference plane	continuous
Mean Anomaly (deg)	The fraction of an elliptical orbit's period that has elapsed since the orbiting body passed periapsis, expressed as an angle which can be used in calculating the position of that body in the classical two-body problem	continuous
Perihelion Distance (AU)	The point in the orbit of an asteroid nearest to the sun. It is the opposite of aphelion, which is the point farthest from the sun.	continuous
Aphelion Distance (AU)	The point in the orbit of an asteroid farthest from the sun	continuous
Orbital Period (yr)	Time an asteroid takes to complete one orbit around another object	continuous
Minimum Orbit Intersection Distance (AU)	The distance between the closest points of the osculating orbits of two bodies	continuous
Orbital Reference	Orbital reference number	continuous
Asteroid Magnitude	Visual magnitude an observer would record if the asteroid were placed 1 Astronomical Unit (au) away, and 1 au from the Sun and at a zero-phase angle	continuous

DARIA KUDOIAR
dariakudoiar@gmail.com

STEP 2 Reduced Data Set

First, we drop the column 'Object Name' which represents the unique (Catalogue) name of Asteroid. This column has 15635 distinct values and doesn't seem to correlate with other features. We leave all the rest of features described above for classification task.

We randomly choose a sample of 2 largest classes: Apollo and Amor and we use the whole subset of Aten Asteroids to create a 3rd class. We get Reduced Data Set with 5137 cases and 13 features per each case. We convert target column to numeric, such that we get CL1 = Apollo =1, CL2 = Amor = 2, CL3 = Aten =3. To get correct proportion between the classes we need to UP sample minority class. CL3(Aten) has the smallest size compared to the other two classes. To get the correct proportion we replicate the under-represented class. Up-sampling is the process of randomly duplicating observations from the minority class in order to reinforce its signal. We follow next steps:

1. First, we separate observations from each class into different DataFrames.
2. Next, we resample the minority class with replacement, setting the number of samples to match that of the majority class.
3. Finally, we combine the up-sampled minority class DataFrame with the original majority class DataFrame.

We get new Reduced Data Set ready for further transformations to implement classification task with 6250 cases and 13 features per each case.
To get an idea how the data set look like we implement PCA analysis in order to to visualize data and see how well classes separate at this point. We keep 3 Principal components to get a 3D representation of original data. White, Red and Black points represent different classes. Since PCA is not necessarily helping segment or separate the data, which we see in the plots below. It is clear that in a 3-dimensional space, the classes are not separated, but there are some areas where each Asteroid type somewhat grouped together. However, there are no 3 nice clusters and we can say that the data is not linearly separable

DARIA KUDOIAR
dariakudoiar@gmail.com

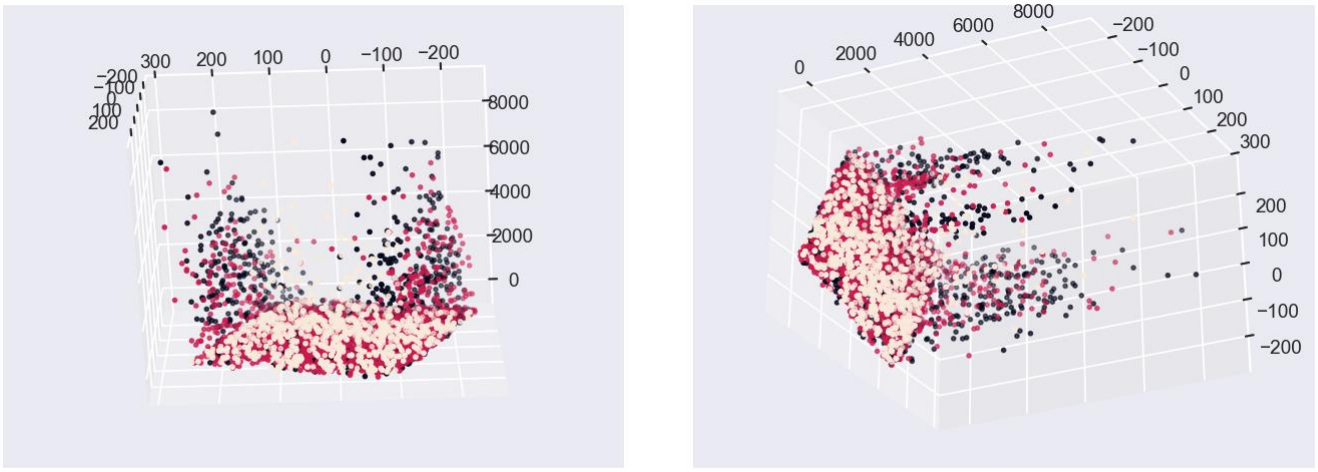

Figure 2: Visualization of Data set with PCA

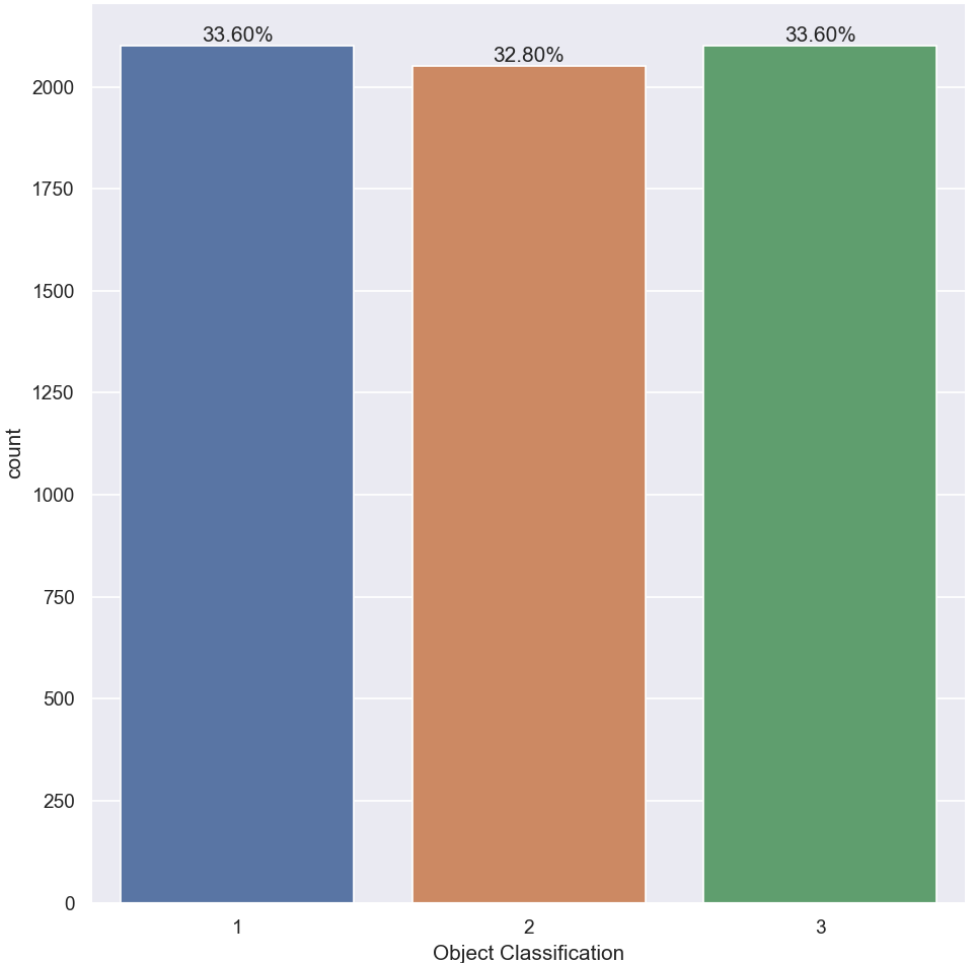

Figure 3: Count plot for Reduced Data Set

DARIA KUDOIAR
dariakudoiar@gmail.com

We calculate Mean and Standard Deviation for each feature within each class.

Name of feature	CL1		CL2		CL3	
	MEAN	SD	MEAN	SD	MEAN	SD
Epoch (TDB)	57172.18	1380.45	57449.16	1007.99	57500.77	1015.98
Orbit Axis (AU)	1.72	0.51	2.03	0.50	0.89	0.08
Orbit Eccentricity	0.49	0.18	0.41	0.14	0.33	0.16
Orbit Inclination (deg)	11.92	10.98	14.04	11.32	12.66	10.49
Perihelion Argument (deg)	184.55	101.61	177.28	105.72	190.69	110.89
Node Longitude (deg)	173.45	104.40	175.44	103.04	171.40	105.88
Mean Anomaly (deg)	171.52	120.24	170.51	116.34	185.07	100.49
Perihelion Distance (AU)	0.81	0.19	1.14	0.08	0.61	0.17
Aphelion Distance (AU)	2.63	1.03	2.93	0.98	1.18	0.14
Orbital Period (yr)	2.32	1.04	2.96	1.09	0.85	0.11
Minimum Orbit Intersection Distance (AU)	0.06	0.08	0.18	0.11	0.06	0.06
Orbital Reference	15.45	25.16	22.47	33.54	18.02	31.03
Asteroid Magnitude	23.36	3.15	21.61	2.64	23.76	2.76

For visual representation of the graphically depicting groups of features within each class we use boxplots of means and standard deviations. We had to drop the most explicit outlier – “Epoch” feature that has range 9263. Box plot helps to understand how data is spread out and is also used for detect outliers in data set.

DARIA KUDOIAR
dariakudoiar@gmail.com

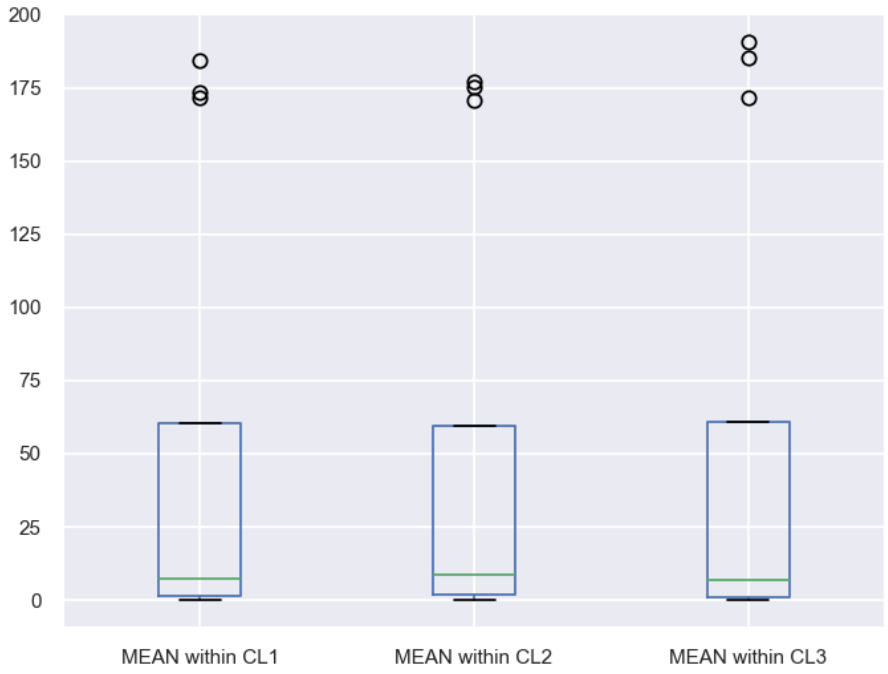

Figure 4: Boxplots of the Mean Values of Features within each Class

We have somewhat equally spread out means of features within each class and it seems that within each class we have three ‘outliers’ with values ~200. Boxplots for all three classes are somewhat identical, all three skewed right and it may indicate the high variation among the mean of features somewhere between 5 to 55.

DARIA KUDOIAR
dariakudoiar@gmail.com

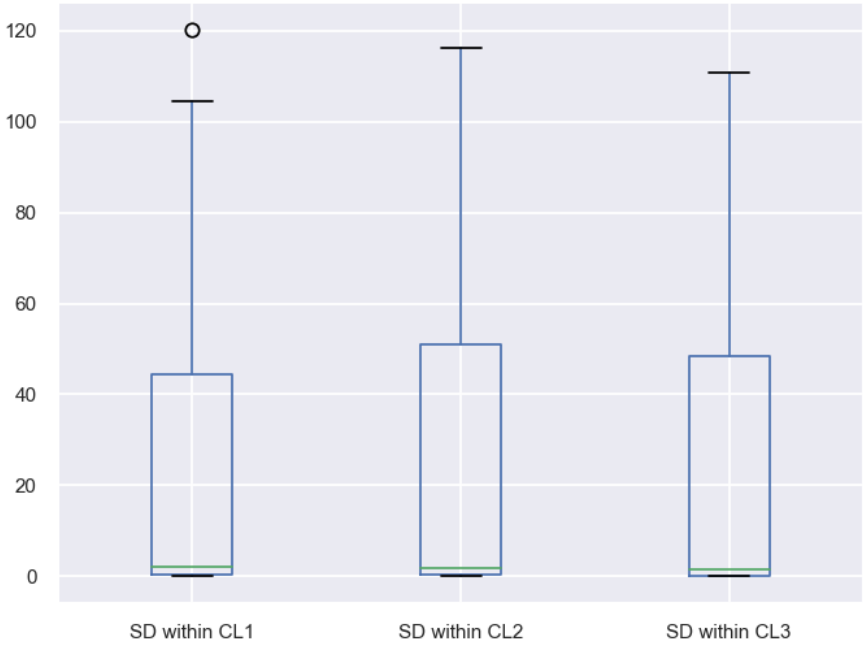

Figure 5: Boxplots of the Standard deviations of features within each Class

We see that standard deviations within CL2 is slightly more spread out compared to the other two classes CL1 and CL3. We see outlier for standard deviation within class 1 with value ~120. All three boxplots comparatively tall and skewed right.

To explore correlation of features we create a correlation heatmap. The highest correlation has values that close to 1 or – 1, we should look at the lightest and the most intense colors on the heatmap. We see that Orbit Axis and Aphelion Distance, Orbital Period and Aphelion Distance, Orbital Period and Orbit Axes, Orbit Eccentricity and Aphelion Distance have the strongest correlation. However, we will keep all the features for further analysis.

DARIA KUDOIAR
dariakudoiar@gmail.com

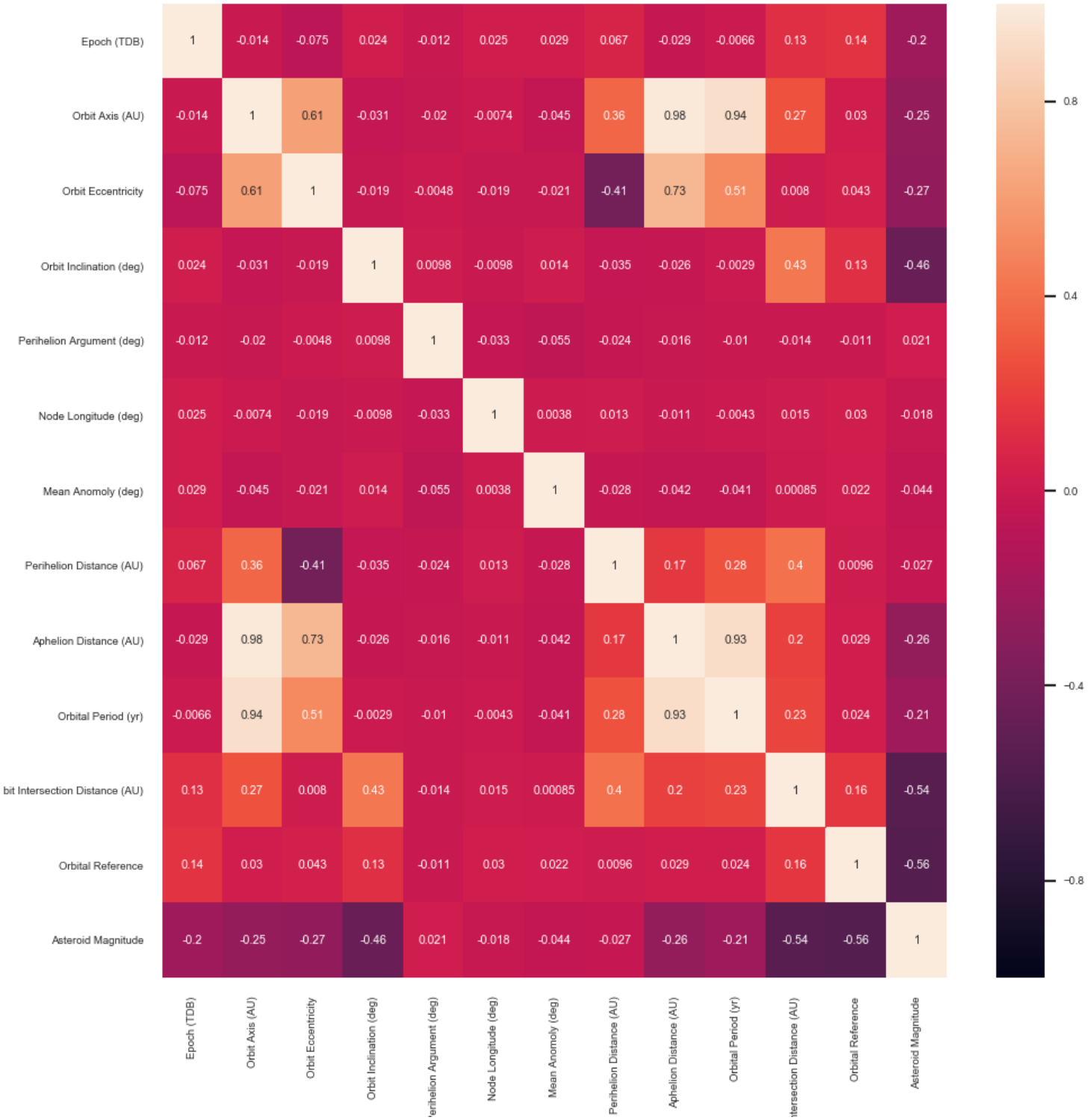

Figure 6: Feature Correlation Heatmap

DARIA KUDOIAR
dariakudoiar@gmail.com

STEP 3

We center and rescale data to get mean = 0 and standard deviation = 1. To get the right proportion for each class within train and test set, we split separately each class into train and test set and then concatenate all three train set and all three test set with three classes into one train set and one test set. Shape of sets for CL1: train – (1680; 13), test – (420,13). Shape of sets for CL2: train – (1640; 13), test – (410,13). Shape of sets for CL3: train – (1680; 13), test – (420,13).

We get the final train set with size 5000 cases and 13 columns, test set with 1250 cases and 13 columns.

QUESTION 2. SVM classification by Radial Kernel

We select a Radial or RBF Kernel, given by the formula:

$$K(x, y) = e^{-\gamma \|x-y\|^2}$$

Hyperplane 'RBF' uses a nonlinear hyper-plane to separate the data and in this case, we expect to have high accuracy of classification task for our data set.

STEP 1 Optimization of parameters

We optimize parameters Cost (c) and Gamma for Radial Kernel with CL1vsCL2, and then re-evaluate best parameters with CL1vsCL3 and do it again with CL2vsCL3. The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points.

We use GridSearchCV function in Scikit-Learn Python library. This function found the best parameter by running the model multiple times with fixed cross validation value (cv = 10). When 10-fold cross validation is done we can see different score in each iteration. This happens because when we use train_test_split method, the dataset get split in random manner into testing and

training dataset. Thus it depends on how the dataset got split and which samples are training set and which samples are in testing set. This is the reason we got different accuracy score. We fix list of parameters: Cost = [5, 7, 10, 15] and Gamma = [0.01, 0.1, 0.25, 0.5]. We run again SVM with best parameters on Train and Test set for each of the pair of classes and compare performances, confusion matrices and ratios of support vectors. We take all the values of C into account and check out the accuracy score with kernel as Radial and we conclude the best parameters for further analysis are:

Cost = 15 and Gamma = 0.1 With these parameters we get percentages of correct prediction for Train set: 99.9% and Test set: 98.8%, and 0.15 as a ratio of support vectors.

QUESTION 3. SVMs for multiclass classification with Radial Kernel

We use the best parameters to implement SVM classification for multiclass problem. We run SVM1 to classify CL1 vs (not CL1), SVM2 to classify CL2 vs (not CL2), SVM3 to classify CL3 vs (not CL3)

To perform SVM mentioned above, we need to change the target labels such that we get "binary" classification. To run SVM1 and classify CL1 vs (not CL1), which means we search for classification CL1 vs CL2∪CL3 we need to transform target column to get CL1 vs CL0. In this case we convert target values as CL1 = 1, CL2 = 0 and CL3 = 0. We follow the same procedure for SVM2 to classify CL2 vs (not CL2), where CL2 = 1 and CL1 = 0 and CL3 = 0. Also, SVM3 to classify CL3 vs (not CL3), where CL3 = 1 and CL1 = 0, CL2 = 0. We train model on TRAIN set and then its ready to implement algorithm on unseen before data set – Test set.

DARIA KUDOIAR
dariakudoiar@gmail.com

SVM1 to classify CL1 vs (not CL1)

Table1: Number and Ratio of Support vectors on Training set

Number S of support vectors for Radial Kernel	1385
Ratio of support vectors for Radial Kernel	0.28

Table 2: Classification Accuracy and Classification Errors, Confidence Intervals for CA and CE

	Percentage of Accuracy	95% Confidence Interval for Accuracy	Percentage of Classification Errors	95% Confidence Interval for Classification Errors
Training Set	93.7%	[92.08; 94.37]	6.3%	[4.95; 7.65]
Test Set	92.8%	[92.08; 93.52]	7.2%	[5.77; 8.63]

We compute confusion matrices for Training and Test set and converted it in terms of frequencies. Each column of the matrix corresponds to a predicted class and each row of the matrix corresponds to an actual class. The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions.

CL 1 represents Class 1 (Apollo Asteroid) and CL 0 represent either Class 2 (Amor Asteroid) or Class 3 (Aten Asteroid).

Table 3: Converted confusion matrix for Training data set

Predicted class / True class	CL 0	CL 1
CL 0	97.5%	2.5%
CL 1	13.97%	86.03%

DARIA KUDOIAR
dariakudoiar@gmail.com

Analyze confusion matrix for TRAIN set we can say that SVM algorithm correctly classified 86.03% of Apollo Asteroids and 13.97% of Apollo Asteroids SVM classified incorrectly, SVM made a decision to classify it as either Amor or Aten Asteroids. However, 97.5% of combined CL0 (CL2∪CL3) were classified correctly, while 2.5% of true CL0 were classified as CL1.

Table 4: Converted confusion matrix for Test data set

Predicted class / True class	CL 0	CL 1
CL 0	97.92%	2.08%
CL 1	16.9%	83.10%

Analyze confusion matrix for unseen data. Test set we see that algorithm classified correctly 83.10% of CL1(Apollo Asteroids) and 97.92% of CL0 (Amor or Aten Asteroids). We should take into account that 16.9% of Apollo Asteroids SVM classified as either Amor or Aten, and vice versa 2.08% of Apollo and Aten Asteroids SVM classified as Apollo Asteroids.

DARIA KUDOIAR
dariakudoiar@gmail.com

SVM2 to classify CL2 vs (not CL2)

Table5: Number and Ratio of Support vectors on Training set

Number S of support vectors for Radial Kernel	504
Ratio of support vectors for Radial Kernel	0.1

Table 6: Classification Accuracy and Classification Errors, Confidence Intervals for CA and CE

	Percentage of Accuracy	95% Confidence Interval for Accuracy	Percentage of Classification Errors	95% Confidence Interval for Classification Errors
Training Set	98.98%	[98.7; 99.3]	1.02%	[0.46; 1.58]
Test Set	98.96%	[98.7; 99.2]	1.04%	[0.16; 1.6]

With only 10% of support vectors SVM2 for classification CL2 vs (not CL2) has an accuracy on both Training and Test set is quite high, almost 100%. Also, we see that the confidence intervals on accuracy for Train and Test set overlap, which makes us think that there is no difference in performances.

CL 1 represents Class 1 (Apollo Asteroid) and CL 0 represent either Class 2 (Amor Asteroid) or Class 3 (Aten Asteroid).

Table 7: Converted confusion matrix for Training data set

Predicted class / True class	CL 0	CL 1
CL 0	98.95%	1.05%
CL 1	0.97%	99.03%

DARIA KUDOIAR
dariakudoiar@gmail.com

Percentage of correct prediction for CL0 and CL1 with Train set is 98.95% and 99.03%, respectively. While percentage of misclassification for CL0 and CL1 is around 1% for each of class.

Table 8: Converted confusion matrix for Test data set

Predicted class / True class	CL 0	CL 1
CL 0	98.94%	1.06%
CL 1	1.01%	98.99%

Correct prediction for Test set for CL0 and CL1 is quite close to what we`ve got on Train set. For CL0 it was correctly classified 98.94%, while 1.06% of CL0 were classified as CL1. In addition, 98.99% of CL1 were classified correctly and misclassified 1.01% of CL1 as CL0.

DARIA KUDOIAR
dariakudoiar@gmail.com

SVM3 to classify CL3 vs (not CL3)

Table9: Number and Ratio of Support vectors on Training set

Number S of support vectors for Radial Kernel	270
Ratio of support vectors for Radial Kernel	0.05

Table 10: Classification Accuracy and Classification Errors, Confidence Intervals for CA and CE

	Percentage of Accuracy	95% Confidence Interval for Accuracy	Percentage of Classification Errors	95% Confidence Interval for Classification Errors
Training Set	99.5%	[99.28; 99.68]	0.52%	[0.12; 0.92]
Test Set	98.8%	[98.5; 99.1]	1.2%	[0.6; 1.8]

CL 1 represents Class 1 (Apollo Asteroid) and CL 0 represent either Class 2 (Amor Asteroid) or Class 3 (Aten Asteroid).

Table 11: Converted confusion matrix for Training data set

Predicted class / True class	CL 0	CL 1
CL 0	99.2%	0.8%
CL 1	0	100%

Table 12: Converted confusion matrix for Test data set

Predicted class / True class	CL 0	CL 1
CL 0	98.2%	1.8%
CL 1	0.6%	99.4%

DARIA KUDOIAR
dariakudoiar@gmail.com

QUESTION 4. Combination of the three SVMs to classify all cases

We apply three SVM to classify classes and get respective predicted values and confusion matrices for each of the SVM. We need to test decisions of SVM and make a conclusion about the resulting prediction for each new case x in R^p. We combine the three classifications of x provided by SVM1, SVM2, SVM3. And take to the account their confusion matrices, in order to obtain a terminal classification of x into one of the 3 classes CL1, CL2, CL3.

We create a function which compare combined predictions from each of the SVMs, calculate reliability and calculate weighted votes based on reliability of SVM. More precisely, we compare the three confusion matrices and specifically the diagonals of matrices and work with percentages of correct predictions for classes.

For ex: confusion matrix for SVM1 where CL1 is 1 and (CL2∪CL3) is 0.

Predicted class / True class	CL 0	CL 1
CL 0	97%	0.3%
CL 1	11.1%	89%

We can assume that model made a decision that x is not in the CL1; x is either in CL2 or CL3. In this case we have 0 for CL1, 97×0.5 for CL2, 97×0.5 for CL3.

We follow the same steps for the rest Confusion matrices and have a summary of combinations. We do the same procedure for both Train and Test set and get the confusion matrices as follows:

DARIA KUDOIAR
dariakudoiar@gmail.com

Table 13: Confusion matrix for Resulting prediction on Train set

Predicted class / True class	CL 1	CL 2	CL 3
CL 1	96.3%	2.1%	1.6%
CL 2	1%	99%	0
CL 3	0	0	100%

95% Confidence Interval for CL1: [95.78; 96.82]

95% Confidence Interval for CL2: [98.73; 99.33]

95% Confidence Interval for CL3: [100 ;100]

Table 14: Confusion matrix for Resulting prediction on Test set

Predicted class / True class	CL 1	CL 2	CL 3
CL 1	94.4%	2.1%	3.5%
CL 2	1.1%	98.9%	0
CL 3	0	0	100%

95% Confidence Interval for CL1: [93.13; 95.67]

95% Confidence Interval for CL2: [98.32; 99.48]

95% Confidence Interval for CL3: [100; 100]

From summary table for Train set there are 96% for CL1, 99% for CL2 and 100% for CL3. Similar results provide confusion matrix for Test set: 94.4% votes for CL1, for CL2 is 98.9% and for CL3 is 100%. We can make a decision based on

maximum votes and take into account confidence interval for each of the classes that x is classified in CL3 – Aten Asteroids.

DARIA KUDOIAR
dariakudoiar@gmail.com

QUESTION 3. SVMs for multiclass classification with Polynomial Kernel

In general, the polynomial kernel is defined as:

$$K(x, y) = (a + \langle x, y \rangle)^r$$

where a – constant term, r – degree of polynomial.

In our case we use Polynomial Kernel of degree 2 with constant term a = 1 for SVM classification. For classification with Polynomial Kernel we run optimization process three times: with CL1vsCL2, and then re-evaluate best parameters with CL1vsCL3 and CL2vsCL3.

We use GridSearchCV function in Scikit-Learn Python library as we used with Radial Kernel with fixed cross validation value (cv = 10). We fix list of parameters: Cost = [7, 10, 15, 30]. To make the final conclusion about best parameter Cost we run again SVM with best parameters on Train and Test set for each of the pair of classes and compare performances, confusion matrices and ratios of support vectors. We conclude the best parameter Cost for further analysis is 30. With Cost = 30 we get high accuracy on Train set with all three SVMs with pairs of classes, as 99.9% $\pm$ 0.08% with 0.03 as a ratio of support vectors and on Test set 97.5% $\pm$ 0.6%.

SVM1 to classify CL1 vs (not CL1)

Table15: Number and Ratio of Support vectors on Training set

Number S of support vectors for Polynomial Kernel	400
Ratio of support vectors for Polynomial Kernel	0.08

Table 16: Classification Accuracy and Classification Errors, Confidence Intervals for CA and CE

	Percentage of Accuracy	95% Confidence Interval for Accuracy	Percentage of Classification Errors	95% Confidence Interval for Classification Errors
Training Set	99.56%	[99.38%; 99.74%]	0.44%	[0.07%; 0.81%]
Test Set	97.52%	[97.08%; 98.52%]	2.48%	[1.62%; 3.34%]

DARIA KUDOIAR
dariakudoiar@gmail.com

We get 99.56% accuracy for Train set with 8% of support vectors and 97.52% accuracy for Test set. We 95% confident that the true accuracy for Train set lies in interval [99.38%; 99.74%] and the true accuracy for Test set lies in interval [97.08%; 98.52%]. What is more important that confidence intervals for Train and Test set do not overlap.

CL 1 represents Class 1 (Apollo Asteroid) and CL 0 represent either Class 2 (Amor Asteroid) or Class 3 (Aten Asteroid).

Table 17: Converted confusion matrix for Training data set

Predicted class / True class	CL 0	CL 1
CL 0	99.85%	0.15%
CL 1	1.02%	98.98%

From confusion matrix for Train set we can conclude that 99.85% of CL0 and 98.98% of CL1 actual CL0 and CL1 were correctly recognized, but 0.15% out of all CL0 cases were recognized by the model as CL1, 1.02% out of all CL1 cases were recognized as CL0.

Table 18: Converted confusion matrix for Test data set

Predicted class / True class	CL 0	CL 1
CL 0	98.78%	1.22%
CL 1	4.86%	95.14%

We have next situation for confusion matrix on Test set: compared to the Train set we have higher misclassification percentage for CL1 which is ~5% and 95.14% of CL1 were correctly classified. For CL0 it was correctly classified 98.78% of all cases and model was confused to classify correctly 1.22% of CL0.

DARIA KUDOIAR
dariakudoiar@gmail.com

SVM2 to classify CL2 vs (not CL2)

Table19: Number and Ratio of Support vectors on Training set

Number S of support vectors for Polynomial Kernel	195
Ratio of support vectors for Polynomial Kernel	0.04

Table 20: Classification Accuracy and Classification Errors, Confidence Intervals for CA and CE

	Percentage of Accuracy	95% Confidence Interval for Accuracy	Percentage of Classification Errors	95% Confidence Interval for Classification Errors
Training Set	99.94%	[100; 99.87]	0.06%	[0.02; 0.08]
Test Set	98.72%	[98.41; 99.03]	1.28%	[1.9; 0.66]

Accuracy of SVM2 to classify CL2 vs (not CL2) is 99.94% on Train set and 98.72% on Test set. However, ratios of support vectors is 4%. Polynomial Kernel did a very good job for classification task on Asteroids data set.

CL 1 represents Class 1 (Apollo Asteroid) and CL 0 represent either Class 2 (Amor Asteroid) or Class 3 (Aten Asteroid).

Table 21: Converted confusion matrix for Training data set

Predicted class / True class	CL 0	CL 1
CL 0	99.91%	0.09%
CL 1	0	100%

Accuracy for Train set shows extremely high percentage. There were correctly classified 100% of CL1 but 99.91% of CL0 with misclassification for CL0 with 0.09%.

DARIA KUDOIAR
dariakudoiar@gmail.com

Table 22: Converted confusion matrix for Test data set

Predicted class / True class	CL 0	CL 1
CL 0	98.83%	1.17%
CL 1	1.51%	98.49%

Compared to train set, confusion matrix for Test set shows slightly lower performance. Since model perform classification task on unseen data we expect smaller number of correct classification. But it still shows very high performance, 98.49% of CL1 classified correctly and 1.51% misclassified. In addition, 98.83% out of all cases for CL0 were classified correctly and 1.17% were classified as CL1.

DARIA KUDOIAR
dariakudoiar@gmail.com

SVM3 to classify CL3 vs (not CL3)

Table 23: Number and Ratio of Support vectors on Training set

Number S of support vectors for Radial Kernel	114
Ratio of support vectors for Radial Kernel	0.02

Table 24: Classification Accuracy and Classification Errors, Confidence Intervals for CA and CE

	Percentage of Accuracy	95% Confidence Interval for Accuracy	Percentage of Classification Errors	95% Confidence Interval for Classification Errors
Training Set	100%	[100; 100]	0%	[0; 0]
Test Set	99.12%%	[98.86; 99.38]	0.88%	[0.36; 1.4]

SVM3 has the smallest ratio of support vectors as 2% with highest percentage of accuracy as 100% for Train set and 99.12% on Test set compared to SVM2 and SVM3.

Confusion matrices below along with overall percentage of accuracy for the SVM3 show somewhat the highest percentage of correctly classified classes.

CL 1 represents Class 1 (Apollo Asteroid) and CL 0 represent either Class 2 (Amor Asteroid) or Class 3 (Aten Asteroid).

Table 25: Converted confusion matrix for Training data set

Predicted class / True class	CL 0	CL 1
CL 0	100%	0
CL 1	0	100%

DARIA KUDOIAR
dariakudoiar@gmail.com

Table 26: Converted confusion matrix for Test data set

Predicted class / True class	CL 0	CL 1
CL 0	98.79%%	1.21%
CL 1	0.24%	99.76%

For Training set we have an ideal confusion matrix, where on diagonal we have 100% which means there were correctly classified all cases for CL1 and CL0, and we have zeros above and below diagonal, which means we don't have classification errors. However, confusion matrix for Test set shows moderately less percentage of correct classification with classification errors for CL1 as 0.24% and CL0 as 1.21%.

In conclusion, comparing SVM1 for CL1 vs (not CL1), SVM2 for CL2 vs (not CL2) and SVM3 for CL3 vs (not CL3) with Polynomial basis Kernel, we can say that SVM3 model provides the best accuracy of classification task with the smallest ratio of support vectors as 2% and confidence intervals for true accuracy as (100%; 100%) for Train set and (98.86%; 99.38%) for Test set. However, for all three SVM models we got a very high accuracy with rather small misclassification error. The 'worst' accuracy scenario we`ve got with SVM1 classification task for CL1 vs (not CL1). With 8% of support vectors we got true accuracy for Train set that lies in interval [99.38%; 99.74%] and the true accuracy for Test set that somewhere between [97.08%; 98.52%]. What is more important that confidence intervals for Percentage of accuracy for all three SVM algorithms do not overlap. Which means that there is significant difference in model`s performances and we can be even more confident with our results.

DARIA KUDOIAR
dariakudoiar@gmail.com

QUESTION 4. Combination of the three SVMs to classify all cases with Polynomial Kernel

We repeat the whole preceding procedure that we used with Radial Kernel to classify all cases using the Polynomial Kernel.

We get as the results confusion matrices on Train and Test set:
CL 1 represents Class 1 which is Apollo Asteroid group and CL 2 represents Class 2 which is Amor Asteroid group and Class 3 is Aten Asteroid group.

The goal is to combine predictions of SVMs models to be able to classify all cases in Train and Test set and make a decision based on reliability of SVM.

CL 1 represents Class 1 (Apollo Asteroid) and CL 2 represents Class 2 (Amor Asteroid) and CL3 is Class 3 (Aten Asteroid). The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier.

Table 27: Confusion matrix for Resulting prediction on Train set

Predicted class / True class	CL 1	CL 2	CL 3
CL 1	97.54%	1.56%	0.9%
CL 2	0.73%	99.27%	0
CL 3	0	0	100%

95% Confidence Interval for CL1: [97.11%; 97.97%]

95% Confidence Interval for CL2: [99.03%; 99.51%]

95% Confidence Interval for CL3: [100%; 100%]

DARIA KUDOIAR
dariakudoiar@gmail.com

Table 28: Confusion matrix for Resulting prediction on Test set

Predicted class / True class	CL 1	CL 2	CL 3
CL 1	95.14%	1.85%	3.01%
CL 2	1.26%	98.74%	0
CL 3	0	0	100%

95% Confidence Interval for CL1: [93.95%; 96.33%]

95% Confidence Interval for CL2: [98.12%; 99.36%]

95% Confidence Interval for CL3: [100%; 100%]

From 3×3 confusion matrices for Train set we can see that the major votes for CL3. Moreover, for CL3 we got result as good as it could be, which is 100%. By analyzing summary table for Test set we see an identical situation – 100% votes for CL3. Prediction conclusion that the x case classified as Aten Asteroid group.

DARIA KUDOIAR
dariakudoiar@gmail.com

Conclusion:

The cost for SVM as a competitive approach is much higher computational time, which was needed for the finding of the optimal parameters of the kernel function in particular. The process of model development was time consuming, as well. The goal of the project was to perform classification task with SVM algorithm on real data set.

We implemented Support Vector Machine multiclass classification task on Asteroids data set. Originally data set was not linearly separable so we decided to use Radial basis Kernel and Polynomial basis Kernel. We implement SVM separately for each pair of three classes to optimize parameters and ran SVM algorithm to classify CL_i vs notCL_i. We obtain a terminal classification of x into one of the 3 classes CL1, CL2, CL3.

Both Radial and Polynomial Kernels provide quite high accuracies. We can see that both Kernels did a pretty good job to find hyperplane that is able to distinguish between three classes.

However, take into account accuracy of prediction on Train and Test set along with ratios of support vectors for each SVM run, we can conclude that the best separator hyperplane for real data set Asteroids was defined with Polynomial basis Kernel.