Big Data Project

Hanciu Raluca-Maria

Cuculescu Andrei

Gr. 411

Introduction

For this project we have chosen a data set from Kaggle, called “Rain in Australia”. This dataset contains about 10 years of daily weather observations from many locations across Australia.

RainTomorrow is the target variable to predict. It means -- did it rain the next day, Yes or No? This column is Yes if the rain for that day was 1mm or more.

- Number of Features: 23
- Number of samples: 145460

Some of the features from the dataset are:

- DATE - The date of observation
- LOCATION - The common name of the location of the weather station
- MINTEMP - The minimum temperature in degrees celsius
- MAXTEMP - The maximum temperature in degrees celsius
- RAINFALL - The amount of rainfall recorded for the day in mm
- EVAPORATION - The so-called Class A pan evaporation (mm) in the 24 hours to 9am
- SUNSHINE - The number of hours of bright sunshine in the day.
- WINDGUESTDIR - The direction of the strongest wind gust in the 24 hours to midnight
- WINDGUESTSPEED- The speed (km/h) of the strongest wind gust in the 24 hours to midnight
- WINDDIR9AM - Direction of the wind at 9am

We have imported all the necessary libraries and then we began with the data cleaning.

How our initial data looks like :

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindS
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	WNW	20.0	
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	WSW	4.0	
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	WSW	19.0	
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	E	11.0	
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	NW	7.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
145455	2017-06-21	Uluru	2.8	23.4	0.0	NaN	NaN	E	31.0	SE	ENE	13.0	
145456	2017-06-22	Uluru	3.6	25.3	0.0	NaN	NaN	NNW	22.0	SE	N	13.0	
145457	2017-06-23	Uluru	5.4	26.9	0.0	NaN	NaN	N	37.0	SE	WNW	9.0	
145458	2017-06-24	Uluru	7.8	27.0	0.0	NaN	NaN	SE	28.0	SSE	N	13.0	
145459	2017-06-25	Uluru	14.9	NaN	0.0	NaN	NaN	NaN	NaN	ESE	ESE	17.0	

145460 rows × 23 columns

Data Cleaning

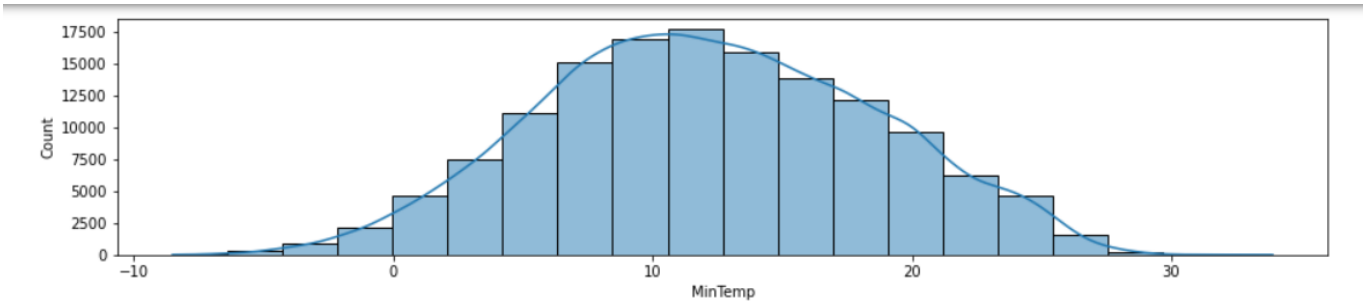
We have checked missing values and deleted features which had more than 15% of data values missing: 'Evaporation', 'Sunshine', 'Cloud9am', 'Cloud3pm'

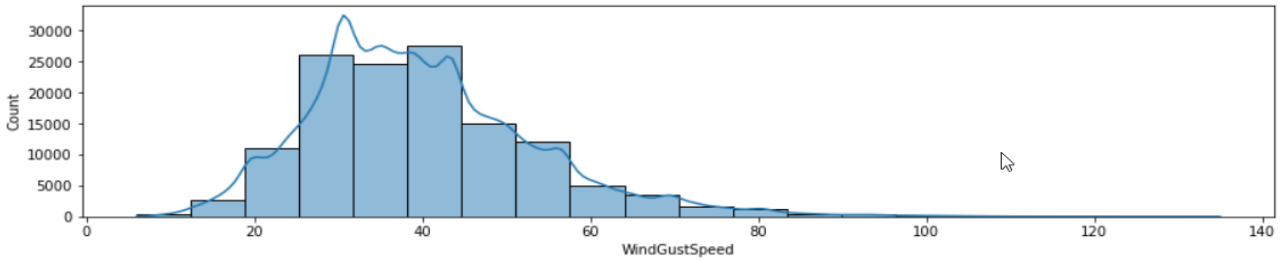
We have also removed rows where target variables are missing

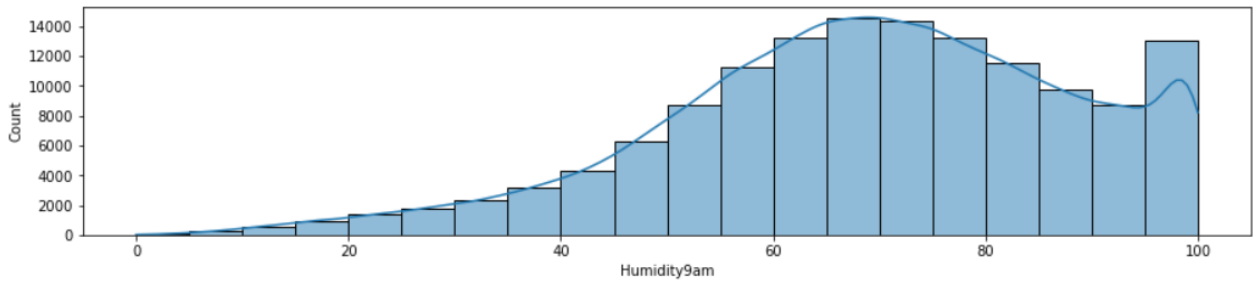
The dataset now contains 140787 samples

Data visualization

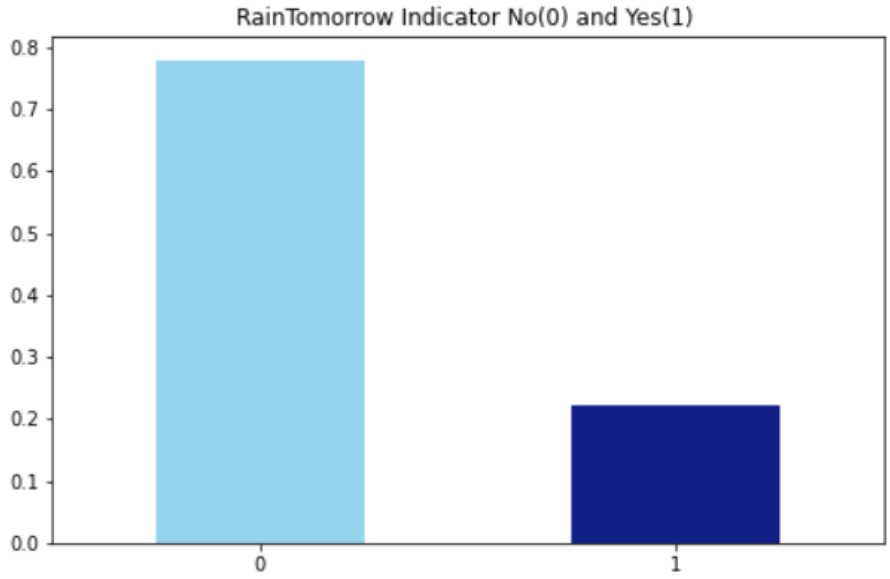
Numerical features distributions:







Target variable (Rain Tomorrow)



Data preparation

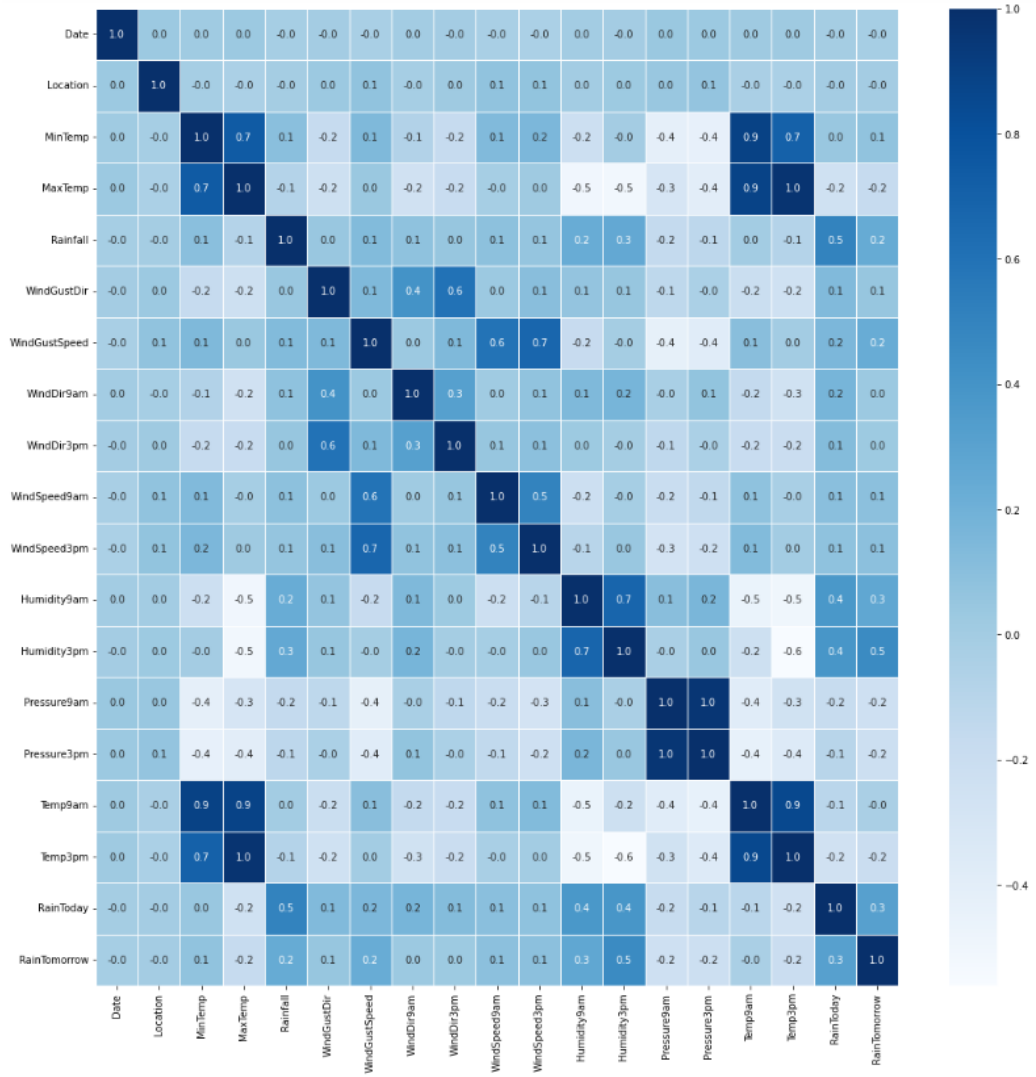
We have changed the categorical values for the feature Rain Today and for the target Rain Tomorrow into 0 and 1, where 0 means no rain and 1 means it rained.

We have used Label Encoder to transform all our categorical data into numerical data.

We have filled in missing values from numerical features with the mean of that feature.

For categorical values we have dropped all samples that have missing values.

We have computed the heatmap:



We have chosen the upper triangle for our heatmap and dropped all features that have a correlation bigger than 0.7, so it won't influence our model. These features are: 'MaxTemp', 'Pressure3pm', 'Temp9am', 'Temp3pm'.

Now our data looks like below:

	Date	Location	MinTemp	Rainfall	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3p
0	377	1	13.4	0.6	13	44.0	13	14	20.0	24.0	71.0	22
1	378	1	7.4	0.0	14	44.0	6	15	4.0	22.0	44.0	25
2	379	1	12.9	0.0	15	46.0	13	15	19.0	26.0	38.0	30
3	380	1	9.2	0.0	4	24.0	9	0	11.0	9.0	45.0	16
4	381	1	17.5	1.0	13	41.0	1	7	7.0	20.0	82.0	33
...	...	...	...	...	...	...	...	...	...	...	...	
145454	3411	39	3.5	0.0	0	31.0	2	0	15.0	13.0	59.0	27
145455	3412	39	2.8	0.0	0	31.0	9	1	13.0	11.0	51.0	24
145456	3413	39	3.6	0.0	6	22.0	9	3	13.0	9.0	56.0	21
145457	3414	39	5.4	0.0	3	37.0	9	14	9.0	9.0	53.0	24
145458	3415	39	7.8	0.0	9	28.0	10	3	13.0	7.0	51.0	24

123710 rows × 15 columns

We have then used Standard Scaler to standardize features by removing the mean and scaling to unit variance (dividing all the values by the standard deviation). This is especially useful for dimensionality reduction.

Models

For the modelling part we have used the following 4 models:

- Logistic Regression
- Random Forest Classifier
- Decision Trees
- Ada Boost Classifier

For dimensionality reduction in combination with the upper models we have used:

- **PCA**
- **SVD**
- Random Forest Built-in **Feature Importance**

We have compared the models to see which performs best.

1. Logistic Regression

No dimensionality reduction

	precision	recall	f1-score	support
0	0.87	0.95	0.90	19284
1	0.72	0.48	0.58	5458
accuracy			0.84	24742
macro avg	0.79	0.71	0.74	24742
weighted avg	0.83	0.84	0.83	24742

```
Accuracy score:  0.844474981812303
Done in:  0.15537500381469727
```

Logistic Regression with SVD

	Number of components	Cumulative Explained Variance Ration	Accuracy	Train Time
12	13.000000	0.981759	0.841533	0.695953
11	12.000000	0.962722	0.829892	0.621327
10	11.000000	0.935111	0.829440	0.602556
8	9.000000	0.869036	0.827168	0.583026
7	8.000000	0.832205	0.826950	0.570825
9	10.000000	0.903422	0.826853	0.589371
6	7.000000	0.779583	0.822617	0.532996
3	4.000000	0.577951	0.817557	0.513586
4	5.000000	0.650866	0.817549	0.551089
2	3.000000	0.492273	0.817323	0.483068
5	6.000000	0.717220	0.817129	0.553323
1	2.000000	0.365357	0.800792	0.462124
0	1.000000	0.195983	0.785749	0.416258

Logistic Regression with PCA

	Number of components	Cumulative Explained Variance Ration	Accuracy	Train Time
12	13.000000	0.981759	0.841533	0.393997
13	14.000000	1.000000	0.841242	0.366103
11	12.000000	0.962722	0.829892	0.330699
10	11.000000	0.935111	0.829440	0.566657
8	9.000000	0.869036	0.827168	0.532114
7	8.000000	0.832205	0.826950	0.547677
9	10.000000	0.903422	0.826853	0.494351
6	7.000000	0.779583	0.822617	0.549508
3	4.000000	0.577951	0.817557	0.475420
4	5.000000	0.650866	0.817549	0.536294
2	3.000000	0.492273	0.817323	0.455748
5	6.000000	0.717220	0.817129	0.517912
1	2.000000	0.365357	0.800792	0.412511
0	1.000000	0.195983	0.785749	0.467665

Logistic Regression with Feature Selection

	Number of features	Cumulative Importance	Accuracy	Train Time
13	14.000000	1.000000	0.841242	0.423524
10	11.000000	0.880883	0.840433	0.323873
9	10.000000	0.838000	0.840377	0.364703
12	13.000000	0.965767	0.840352	0.373825
11	12.000000	0.923662	0.840247	0.358822
8	9.000000	0.791423	0.839851	0.288735
6	7.000000	0.688010	0.838534	0.276170
7	8.000000	0.742362	0.838000	0.333853
3	4.000000	0.475209	0.837539	0.260195
4	5.000000	0.551801	0.837248	0.280182
5	6.000000	0.622793	0.837200	0.315233
2	3.000000	0.396312	0.831978	0.248940
1	2.000000	0.315922	0.831428	0.220188
0	1.000000	0.227757	0.826700	0.233186

With logistic regression performed best the model without any dimensionality reduction. However, the differences are quite small between the results with/ without dimensionality reduction. Also, the train time for the model without dim red was better.

2. Random Forest Classifier

No dimensionality reduction

	precision	recall	f1-score	support
0	0.87	0.95	0.91	19284
1	0.76	0.50	0.60	5458
accuracy			0.85	24742
macro avg	0.82	0.73	0.76	24742
weighted avg	0.85	0.85	0.84	24742

```
Accuracy score:  0.8549430118826288
Done in:  10.874882936477661
```

Random Forest with PCA

	Number of components	Cumulative Explained Variance Ration	Accuracy	Train Time(sec)	Train Time(min)
13	14.000000	1.000000	0.840239	121.026489	2.017108
12	13.000000	0.981759	0.839593	119.179883	1.986331
11	12.000000	0.962722	0.828478	131.189410	2.186490
10	11.000000	0.935111	0.828082	118.002326	1.966705
9	10.000000	0.903422	0.826441	118.124508	1.968742
8	9.000000	0.869036	0.824889	120.821040	2.013684
7	8.000000	0.832205	0.823531	83.968261	1.399471
6	7.000000	0.779583	0.818018	82.475725	1.374595
4	5.000000	0.650866	0.808116	81.184834	1.353081
5	6.000000	0.717220	0.806434	82.041873	1.367365
3	4.000000	0.577951	0.805893	80.240445	1.337341
2	3.000000	0.492273	0.804890	47.196544	0.786609
1	2.000000	0.365357	0.776396	53.581976	0.893033
0	1.000000	0.195983	0.689281	77.871586	1.297860

Random Forest with SVD

	Number of components	Cumulative Explained Variance Ration	Accuracy	Train Time(sec)	Train_time(min)
12	13.000000	0.981759	0.839593	119.367111	1.989452
11	12.000000	0.962722	0.828478	118.751925	1.979199
10	11.000000	0.935111	0.828082	117.356465	1.955941
9	10.000000	0.903422	0.826441	113.836936	1.897282
8	9.000000	0.869036	0.824889	112.334381	1.872240
7	8.000000	0.832205	0.823531	77.850080	1.297501
6	7.000000	0.779583	0.818018	78.785198	1.313087
4	5.000000	0.650866	0.808116	77.136508	1.285608
5	6.000000	0.717220	0.806434	78.217545	1.303626
3	4.000000	0.577951	0.805893	75.588615	1.259810
2	3.000000	0.492273	0.804810	44.413764	0.740229
1	2.000000	0.365357	0.776138	51.226410	0.853773
0	1.000000	0.195983	0.688683	75.770887	1.262848

Random Forest with Feature Selection

	Number of features	Cumulative Importance	Accuracy	Train Time	Train_time(min)
6	7.000000	0.688010	0.840918	40.665926	0.677765
13	14.000000	1.000000	0.839819	49.386018	0.823100
12	13.000000	0.965767	0.839277	52.385454	0.873091
11	12.000000	0.923662	0.839059	53.791553	0.896526
10	11.000000	0.880883	0.837547	55.311434	0.921857
5	6.000000	0.622793	0.836666	42.137853	0.702298
9	10.000000	0.838000	0.835559	54.169214	0.902820
8	9.000000	0.791423	0.834726	54.177573	0.902960
7	8.000000	0.742362	0.833546	41.051974	0.684200
4	5.000000	0.551801	0.833174	37.198207	0.619970
0	1.000000	0.227757	0.826425	7.023191	0.117053
3	4.000000	0.475209	0.824606	33.676324	0.561272
2	3.000000	0.396312	0.806515	21.942604	0.365710
1	2.000000	0.315922	0.805295	19.087661	0.318128

All dimension reduction models performed way worse than the model with no dimension reduction, especially when speaking about train time.

3. Decision Tree

No dimensionality reduction

```
              precision    recall  f1-score   support

           0       0.87      0.85      0.86     19284
           1       0.51      0.54      0.52      5458

    accuracy                           0.78     24742
   macro avg       0.69      0.70      0.69     24742
weighted avg       0.79      0.78      0.78     24742


Accuracy score:  0.7819093040174602
Done in:  0.7023036479949951
```

Decision Tree with PCA

	Number of components	Cumulative Explained Variance Ration	Accuracy	Train Time(sec)	Train Time(min)
12	13.000000	0.981759	0.758273	8.434982	0.140583
13	14.000000	1.000000	0.757902	9.275909	0.154598
11	12.000000	0.962722	0.748210	7.998102	0.133302
10	11.000000	0.935111	0.747029	7.434014	0.123900
9	10.000000	0.903422	0.744620	6.660630	0.111011
8	9.000000	0.869036	0.743481	6.015391	0.100257
7	8.000000	0.832205	0.741565	5.235675	0.087261
6	7.000000	0.779583	0.737628	4.485478	0.074758
2	3.000000	0.492273	0.736343	2.185447	0.036424
3	4.000000	0.577951	0.732859	2.760129	0.046002
4	5.000000	0.650866	0.730434	3.318372	0.055306
5	6.000000	0.717220	0.727985	3.879949	0.064666
1	2.000000	0.365357	0.709684	1.850000	0.030833
0	1.000000	0.195983	0.689281	1.629412	0.027157

Decision Tree with SVD

	Number of components	Cumulative Explained Variance Ration	Accuracy	Train Time(sec)	Train_time(min)
12	13.000000	0.981759	0.758273	8.830007	0.147167
11	12.000000	0.962722	0.748210	8.331408	0.138857
10	11.000000	0.935111	0.747029	7.476247	0.124604
9	10.000000	0.903422	0.744620	6.651896	0.110865
8	9.000000	0.869036	0.743481	6.030850	0.100514
7	8.000000	0.832205	0.741565	5.367823	0.089464
6	7.000000	0.779583	0.737628	4.604477	0.076741
2	3.000000	0.492273	0.735987	2.196380	0.036606
3	4.000000	0.577951	0.732859	2.777183	0.046286
4	5.000000	0.650866	0.730434	3.412137	0.056869
5	6.000000	0.717220	0.727985	3.932911	0.065549
1	2.000000	0.365357	0.711915	1.931123	0.032185
0	1.000000	0.195983	0.689451	1.590617	0.026510

Decision Tree with Feature Selection

	Number of features	Cumulative Importance	Accuracy	Train Time	Train_time(min)
0	1.000000	0.227757	0.826425	0.116814	0.001947
1	2.000000	0.315922	0.804470	0.449639	0.007494
6	7.000000	0.688010	0.764716	2.081114	0.034685
5	6.000000	0.622793	0.760246	1.866547	0.031109
3	4.000000	0.475209	0.759930	1.007744	0.016796
4	5.000000	0.551801	0.759292	1.401623	0.023360
2	3.000000	0.396312	0.756301	0.782015	0.013034
13	14.000000	1.000000	0.754644	3.144254	0.052404
12	13.000000	0.965767	0.753747	3.095620	0.051594
11	12.000000	0.923662	0.750683	2.888059	0.048134
10	11.000000	0.880883	0.749584	2.718935	0.045316
9	10.000000	0.838000	0.745946	2.607660	0.043461
8	9.000000	0.791423	0.744402	2.459892	0.040998
7	8.000000	0.742362	0.741888	2.349135	0.039152

Decision Tree with feature selection has the best accuracy and the best time using only **one** component.

4. Ada Boost Classifier

No dimensionality reduction

	precision	recall	f1-score	support
0	0.87	0.94	0.90	19284
1	0.71	0.49	0.58	5458
accuracy			0.84	24742
macro avg	0.79	0.72	0.74	24742
weighted avg	0.83	0.84	0.83	24742

```
Accuracy score:  0.844474981812303
Done in:  2.739335775375366
```

Ada Boost with PCA

	Number of components	Cumulative Explained Variance Ration	Accuracy	Train Time(sec)	Train Time(min)
12	13.000000	0.981759	0.835963	35.870055	0.597834
13	14.000000	1.000000	0.835963	38.667035	0.644451
11	12.000000	0.962722	0.824622	32.546723	0.542445
10	11.000000	0.935111	0.824178	30.630276	0.510505
8	9.000000	0.869036	0.822997	25.551632	0.425861
7	8.000000	0.832205	0.822431	22.840964	0.380683
9	10.000000	0.903422	0.822318	27.960992	0.466017
6	7.000000	0.779583	0.817290	20.427556	0.340459
2	3.000000	0.492273	0.813459	10.599501	0.176658
3	4.000000	0.577951	0.812990	13.130004	0.218833
4	5.000000	0.650866	0.812950	15.576473	0.259608
5	6.000000	0.717220	0.812254	17.654550	0.294242
1	2.000000	0.365357	0.798723	8.344784	0.139080
0	1.000000	0.195983	0.786695	5.869580	0.097826

Ada Boost with SVD

	Number of components	Cumulative Explained Variance Ration	Accuracy	Train Time(sec)	Train_time(min)
12	13.000000	0.981759	0.835963	36.589637	0.609827
11	12.000000	0.962722	0.824622	33.997792	0.566630
10	11.000000	0.935111	0.824178	31.650194	0.527503
8	9.000000	0.869036	0.822997	25.570459	0.426174
7	8.000000	0.832205	0.822431	24.295465	0.404924
9	10.000000	0.903422	0.822318	28.581761	0.476363
6	7.000000	0.779583	0.817290	21.279226	0.354654
2	3.000000	0.492273	0.813233	10.991038	0.183184
3	4.000000	0.577951	0.812990	13.589546	0.226492
4	5.000000	0.650866	0.812950	16.161525	0.269359
5	6.000000	0.717220	0.812254	18.693815	0.311564
1	2.000000	0.365357	0.798796	8.530503	0.142175
0	1.000000	0.195983	0.786695	5.967052	0.099451

Ada Boost with Feature Selection

	Number of features	Cumulative Importance	Accuracy	Train Time	Train_time(min)
6	7.000000	0.688010	0.840757	9.543211	0.159054
9	10.000000	0.838000	0.840409	11.345314	0.189089
8	9.000000	0.791423	0.840158	11.047948	0.184132
10	11.000000	0.880883	0.839746	11.945113	0.199085
5	6.000000	0.622793	0.838760	8.925587	0.148760
4	5.000000	0.551801	0.838736	7.569662	0.126161
12	13.000000	0.965767	0.838606	13.152711	0.219212
13	14.000000	1.000000	0.838606	13.517733	0.225296
11	12.000000	0.923662	0.838550	12.616699	0.210278
7	8.000000	0.742362	0.838162	10.088520	0.168142
3	4.000000	0.475209	0.837717	6.589810	0.109830
2	3.000000	0.396312	0.832479	5.945192	0.099087
1	2.000000	0.315922	0.830943	5.096945	0.084949
0	1.000000	0.227757	0.827702	4.094575	0.068243

Again our best model is Ada Boost with no dimension reduction.

Conclusions:

The best accuracy had Random Forest with no dimensionality reduction: 85,4% in 10 sec.

Feature selection using Decisional Tree had an accuracy of 82,6% in 0,11 sec, with only one feature.