atlantbh

DATA SCIENCE INTERNSHIP – MONTH 1

Mentor: Belma Ibrahimović
Intern: Mehmed Kadrić

MOTIVATION

- Effects of machine learning algorithms are all around us
- Curiosity
- Digit recognition

TABLE OF CONTENTS

- TASK 1: Python for Data Science
- TASK 2: Classification in Data Science
- TASK 3: Anomaly Detection
- Conclusion

PYTHON FOR DATA SCIENCE

- Important for DS
- Easy to use language
- Most common libraries used in data science:
  - Numpy
  - Pandas
  - Scikit-learn
  - Matplotlib

„Started from the bottom, now I‘m here“

CLASSIFICATION IN DATA SCIENCE

- Use-case scenario:
    - Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew (32% survival rate)
    - Latest discovery have shown that there were additional 10 passengers
- Goal:
    - Create the system which would predict probability of their survival
- Dataset was given
- Supervised learning

TITANIC

Preprocessing

- 1303 examples, 4 features
- Missing values in Age column and PClass column
- Only one outlier in PClass column
- Dimensionality reduction

	Name	PClass	Age	Sex	Survived
1	Allen, Miss Elisabeth Walton	1st	29	female	1

	Title	PClass	LifeStage	Survived
1	Miss	1st	2	1

TITANIC

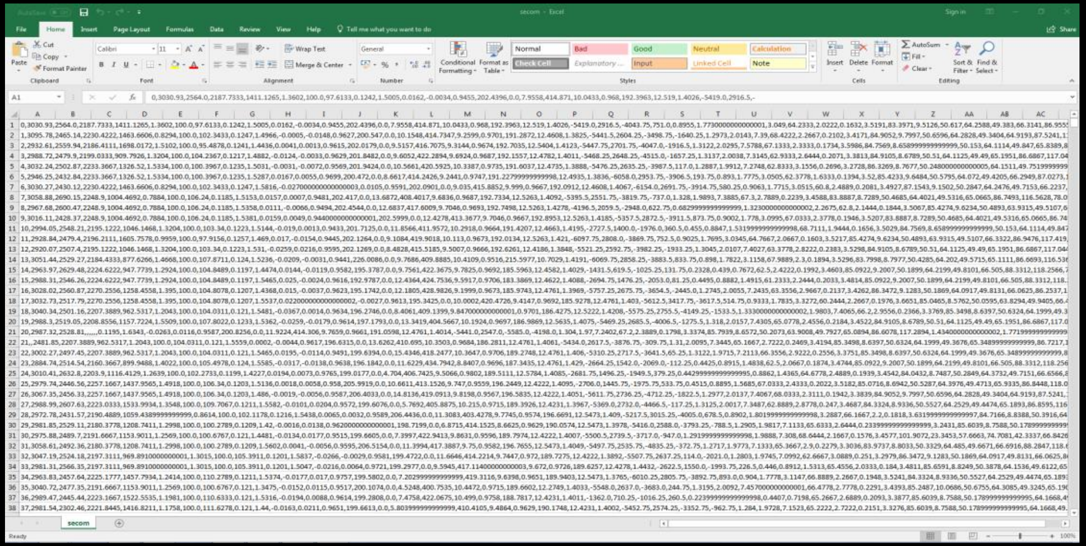
- There are many classification algorithms such as: Logistic Regression, Naive Bayes classifier, SVM, KNN, Decision Trees, etc.
- Support Vector Machine Classifier gave the best results

Name	Probability of survival	Actually survived
Sage, Miss Constance	0.0148	0
Sage, Miss Dorothy	0.0148	0
Wilkinson, Mrs Elizabeth Anne	**0.2134**	**1**
Thomas, Master Assad Alexander	0.0344	0
Zakarian, Mr Artun	0.0079	0
Zakarian, Mr Maprieder	0.0079	0
Anderson, Mr Harry	0.8047	1
Andrews, Miss Kornelia Theodosia	**0.2133**	**1**
Brown, Mrs John Murray (Caroline Lane Lamson)	0.8413	1
Carter, Mr William Ernest	0.7876	1

Accuracy
Precision
Recall
F1-score

Anomaly detection

- Develop system which can be used for smart production monitoring and fault detection in manufacturing processes
- Given dataset: 1567 instances, 591 features
- Unbalanced dataset: 104 fails
- Only numeric values
- Missing values
- Outliers

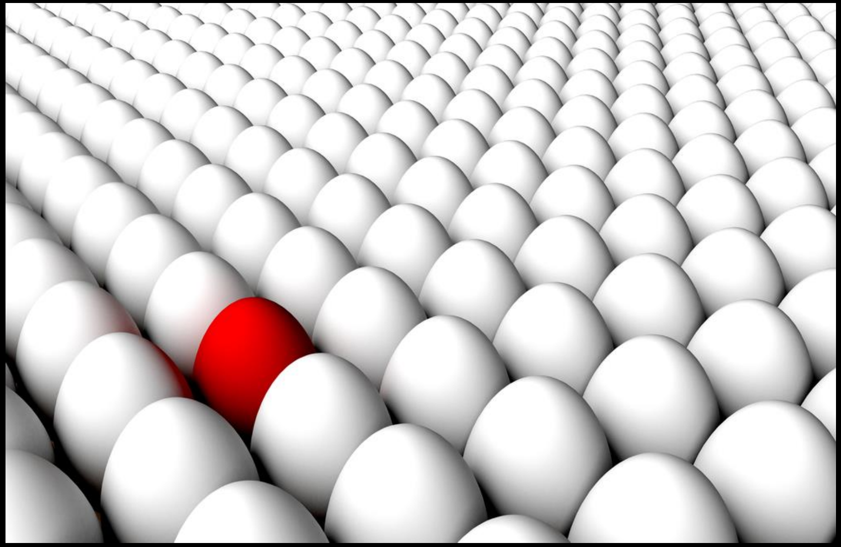


```
A1   fx  0,3030.93,2564.0,2187.7333,1411.1265,1.3602,100.0,97.6133,0.1242,1.5005,0.0162,-0.0034,0.9455,202.4396,0.0,7.9558,414.871,10.0433,0.968,192.3963,12.519,1.4026,-5419.0,2916.5,-

    A  B  C  D  E  F  G  H  I  J  K  L  M  N  O  P  Q  R  S  T  U  V  W  X  Y  Z  AA  AB  AC
1   0,3030.93,2564.0,2187.7333,1411.1265,1.3602,100.0,97.6133,0.1242,1.5005,0.0162,-0.0034,0.9455,202.4396,0.0,7.9558,414.871,10.0433,0.968,192.3963,12.519,1.4026,-5419.0,2916.5,-4043.75,751.0,0.8955,1.7730000000000001,3.049,64.2333,2.0222,0.1632,3.5191,83.3971,9.5126,50.617,64.2588,49.383,66.3141,86.9555
2   1,3095.78,2465.14,2230.4222,1463.6606,0.8294,100.0,102.3433,0.1247,1.4966,-0.0005,-0.0148,0.9627,200.547,0.0,10.1548,414.7347,9.2599,0.9701,191.2872,12.4608,1.3825,-5441.5,2604.25,-3498.75,-1640.25,1.2973,2.0143,7.39,68.4222,2.2667,0.2102,3.4171,84.9052,9.7997,50.6596,64.2828,49.3404,64.9193,87.5241,1.1
3   2,2932.61,2559.94,2186.4111,1698.0172,1.5102,100.0,95.4878,0.1241,1.4436,0.0041,0.0013,0.9615,202.0179,0.0,9.5157,416.7075,9.3144,0.9674,192.7035,12.5404,1.4123,-5447.75,2701.75,-4047.0,-1916.5,1.3122,2.0295,7.5788,67.1333,2.3333,0.1734,3.5986,84.7569,8.6589999999999999,50.153,64.1114,49.847,65.8389,8
4   3,2988.72,2479.9,2199.0333,909.7926,1.3204,100.0,104.2367,0.1217,1.4882,-0.0124,-0.0033,0.9629,201.8482,0.0,9.6052,422.2894,9.6924,0.9687,192.1557,12.4782,1.4011,-5468.25,2648.25,-4515.0,-1657.25,1.3137,2.0038,7.3145,62.9333,2.6444,0.2071,3.3813,84.9105,8.6789,50.51,64.1125,49.49,65.1951,86.6867,117.04
5   4,3032.24,2502.87,2233.3667,1326.52,1.5334,100.0,100.3967,0.1235,1.5031,-0.0031,-0.0072,0.9569,201.9424,0.0,10.5661,420.5925,10.3387,0.9735,191.6037,12.4735,1.3888,-5476.25,2635.25,-3987.5,117.0,1.2887,1.9912,7.2748,62.8333,3.1556,0.2696,3.2728,86.3269,8.7677,50.248000000000005,64.1511,49.7519999999
6   5,2946.25,2432.84,2233.3667,1326.52,1.5334,100.0,100.3967,0.1235,1.5287,0.0167,0.0055,0.9699,200.472,0.0,8.6617,414.2426,9.2441,0.9747,191.22799999999998,12.4935,1.3836,-6058.0,2953.75,-3906.5,193.75,0.893,1.7775,3.0505,62.3778,1.6333,0.1394,3.52,85.4233,9.6484,50.5795,64.072,49.4205,66.2949,87.0273,1
7   6,3030.27,2430.12,2230.4222,1463.6606,0.8294,100.0,102.3433,0.1247,1.5816,-0.027000000000000003,0.0105,0.9591,202.0901,0.0,9.035,415.8852,9.999,0.9667,192.0912,12.4608,1.4067,-6154.0,2691.75,-3914.75,580.25,0.9063,1.7715,3.0515,60.8,2.4889,0.2081,3.4927,87.1543,9.1502,50.2847,64.2476,49.7153,66.2237,
8   7,3058.88,2690.15,2248.9,1004.4692,0.7884,100.0,106.24,0.1185,1.5153,0.0157,0.0007,0.9481,202.417,0.0,13.6872,408.4017,9.6836,0.9687,192.7334,12.5263,1.4092,-5395.5,2551.75,-3819.75,-737.0,1.328,1.9893,7.3885,67.3,2.7889,0.2239,3.4588,83.8887,8.7289,50.4685,64.4021,49.5316,65.0665,86.7493,116.5628,78.0
9   8,2967.68,2600.47,2248.9,1004.4692,0.7884,100.0,106.24,0.1185,1.5358,0.0111,-0.0066,0.9494,202.4544,0.0,12.6837,417.6009,9.7046,0.9693,192.7498,12.5263,1.4278,-4196.5,2059.5,-2948.0,622.75,0.6829999999999999,1.3230000000000002,2.2675,62.8,2.1444,0.1844,3.5067,85.4274,9.6234,50.4893,63.9315,49.5107,6
10  9,3016.11,2428.37,2248.9,1004.4692,0.7884,100.0,106.24,0.1185,1.5381,0.0159,0.0049,0.9440000000000001,202.5999,0.0,12.4278,413.3677,9.7046,0.9667,192.8953,12.5263,1.4185,-5357.5,2872.5,-3911.5,873.75,0.9002,1.778,3.0995,67.0333,2.3778,0.1946,3.5207,83.8887,8.7289,50.4685,64.4021,49.5316,65.0665,86.74
11  10,2994.05,2548.21,2195.1222,1046.1468,1.3204,100.0,103.34,0.1223,1.5144,-0.019,0.0013,0.9433,201.7125,0.0,11.8566,411.9572,10.2918,0.9664,191.4207,12.4663,1.4195,-2727.5,1400.0,-1976.0,360.5,0.455,0.8847,1.5319999999999998,68.7111,1.9444,0.1656,3.5029,84.7569,8.6589999999999999,50.153,64.1114,49.847
12  11,2928.84,2479.4,2196.2111,1605.7578,0.9959,100.0,97.9156,0.1257,1.469,0.017,-0.0154,0.9445,202.1264,0.0,9.1084,419.9018,10.113,0.9673,192.0134,12.5263,1.421,-6097.75,2808.0,-3869.75,752.5,0.9025,1.7695,3.0345,64.7667,2.0667,0.1603,3.5217,85.4274,9.6234,50.4893,63.9315,49.5107,66.3322,86.9476,117.419,
13  12,2920.07,2507.4,2195.1222,1046.1468,1.3204,100.0,103.34,0.1223,1.531,-0.0259,0.0216,0.9595,202.1269,0.0,8.4828,415.5185,9.5007,0.9666,192.6261,12.4186,1.3848,-5521.25,2592.75,-3982.25,-1933.25,1.3045,2.0107,7.4027,63.3778,2.8222,0.2383,3.5298,84.9105,8.6789,50.51,64.1125,49.49,65.1951,86.6867,117.044
14  13,3051.44,2529.27,2184.4333,877.6266,1.4668,100.0,107.8711,0.124,1.5236,-0.0209,-0.0031,0.9441,226.0086,0.0,9.7686,409.8885,10.4109,0.9516,215.5977,10.7029,1.4191,-6069.75,2858.25,-3883.5,833.75,0.898,1.7822,3.1158,67.9889,2.3,0.1894,3.5296,83.7998,8.7977,50.4285,64.202,49.5715,65.1111,86.6693,116.536
15  14,2963.97,2629.48,2224.6222,947.7739,1.2924,100.0,104.8489,0.1197,1.4474,0.0144,-0.0119,0.9582,195.3787,0.0,9.7561,422.3675,9.7825,0.9692,185.5963,12.4582,1.4029,-1431.5,619.5,-1025.25,131.75,0.2328,0.439,0.7672,62.5,2.4222,0.1992,3.4603,85.0922,9.2007,50.1899,64.2199,49.8101,66.505,88.3312,118.2566,7
16  15,2988.31,2546.26,2224.6222,947.7739,1.2924,100.0,104.8489,0.1197,1.5465,0.025,-0.0024,0.9616,192.9787,0.0,12.4364,424.7536,9.5917,0.9706,183.3869,12.4622,1.4088,-2694.75,1476.25,-2053.0,81.25,0.4495,0.8882,1.4915,61.2333,2.2444,0.2033,3.4814,85.0922,9.2007,50.1899,64.2199,49.8101,66.505,88.3312,118.
17  16,3028.02,2560.87,2270.2556,1258.4558,1.395,100.0,104.8078,0.1207,1.4368,0.015,-0.0037,0.9623,195.1742,0.0,12.1805,428.9826,9.1999,0.9673,185.9743,12.4761,1.3969,-5757.25,2675.75,-3654.5,-2445.0,1.2745,2.0055,7.2435,63.3556,2.9667,0.2137,3.4262,86.3472,9.1283,50.1869,64.0917,49.8131,66.0625,86.2537,1
18  17,3032.73,2517.79,2270.2556,1258.4558,1.395,100.0,104.8078,0.1207,1.5537,0.022000000000000002,-0.0027,0.9613,195.3425,0.0,10.0002,420.4726,9.4147,0.9692,185.9278,12.4761,1.403,-5612.5,3417.75,-3617.5,514.75,0.9333,1.7835,3.3272,60.2444,2.2667,0.1976,3.6651,85.0465,8.5762,50.0595,63.8294,49.9405,66.4
19  18,3040.34,2501.16,2207.3889,962.5317,1.2043,100.0,104.0311,0.121,1.5481,-0.0367,0.0014,0.9634,196.2746,0.0,8.4061,409.1399,9.847000000000001,0.9701,186.4275,12.5222,1.4208,-5575.25,2755.5,-4149.25,-1533.5,1.3330000000000002,1.9803,7.4065,66.2,2.9556,0.2366,3.3769,85.3498,8.6397,50.6324,64.1999,49.3
20  19,2988.3,2519.05,2208.8556,1157.7224,1.5509,100.0,107.8022,0.1233,1.5362,-0.0259,-0.0179,0.9614,197.1793,0.0,13.3419,404.5667,10.1924,0.9697,186.9869,12.5635,1.4075,-5469.25,2685.5,-4006.5,-1275.5,1.318,2.0157,7.4305,65.0778,2.4556,0.2184,3.4522,84.9105,8.6789,50.51,64.1125,49.49,65.1951,86.6867,117.0
21  20,2987.32,2528.81,,,,,,0.1195,1.6343,-0.0263,0.0116,0.9587,200.8256,0.0,11.9224,414.306,9.7659,0.9661,191.0598,12.4761,1.4014,-5441.0,2547.0,-5585.0,-4198.0,1.304,1.97,7.2402,67.2,2.3889,0.1798,3.3374,85.7939,8.6572,50.2073,63.9008,49.7927,65.0894,86.6078,117.2894,1.4340000000000002,1.7719999999999
22  21,,2481.85,2207.3889,962.5317,1.2043,100.0,104.0311,0.121,1.5559,0.0002,-0.0044,0.9617,196.6315,0.0,13.6262,410.695,10.3503,0.9684,186.2811,12.4761,1.4061,-5434.0,2617.5,-3876.75,-309.75,1.31,2.0095,7.3445,65.1667,2.7222,0.2469,3.4194,85.3498,8.6397,50.6324,64.1999,49.3676,65.34899999999999,86.7217,1
23  22,3002.27,2497.45,2207.3889,962.5317,1.2043,100.0,104.0311,0.121,1.5465,0.0195,-0.0114,0.9491,199.6394,0.0,15.4346,418.2477,10.3647,0.9706,189.2748,12.4761,1.406,-5310.25,2717.5,-3641.5,65.25,1.3122,1.9715,7.2113,66.3556,2.9222,0.2556,3.3751,85.3498,8.6397,50.6324,64.1999,49.3676,65.34899999999999,8
24  23,2884.74,2514.54,2160.3667,899.9488,1.4022,100.0,105.4978,0.124,1.5585,-0.0317,-0.0138,0.9638,196.1842,0.0,11.6229,434.7942,8.8407,0.9696,187.3435,12.4761,1.429,-2664.25,1542.0,-2069.0,-112.25,0.4425,0.8915,1.4838,62.5,2.0667,0.1874,3.4744,85.0922,9.2007,50.1899,64.2199,49.8101,66.505,88.3312,118.256
25  24,3010.41,2632.8,2203.9,1116.4129,1.2639,100.0,102.2733,0.1199,1.4227,0.0194,0.0073,0.9765,199.0177,0.0,4.704,406.7425,9.5066,0.9802,189.5111,12.5784,1.4085,-2681.75,1496.25,-1949.5,379.25,0.44299999999999995,0.8862,1.4365,64.6778,2.4889,0.1939,3.4542,84.0432,8.7487,50.2849,64.3732,49.7151,66.6566,8
26  25,2979.74,2446.56,2257.1667,1437.9565,1.4918,100.0,106.34,0.1203,1.5136,0.0018,0.0058,0.958,205.9919,0.0,10.6611,413.1526,9.747,0.9559,196.2449,12.4222,1.4095,-2706.0,1445.75,-1975.75,533.75,0.4515,0.8895,1.5685,67.0333,2.4333,0.2022,3.5182,85.0716,8.6942,50.5287,64.3976,49.4713,65.9335,86.8448,118.0
27  26,3067.35,2456.33,2257.1667,1437.9565,1.4918,100.0,106.34,0.1203,1.486,-0.0019,-0.0056,0.9587,206.4033,0.0,14.8136,419.0913,9.8198,0.9567,196.5835,12.4222,1.4051,-5611.75,2736.25,-4712.25,-1822.5,1.2977,2.0137,7.4067,68.0333,2.3111,0.1942,3.3839,84.9052,9.7997,50.6596,64.2828,49.3404,64.9193,87.5241,1
28  27,2988.99,2607.63,2223.0333,1533.9934,1.3548,100.0,109.7067,0.1211,1.5582,-0.0101,0.0204,0.9572,199.6076,0.0,5.7692,405.8875,10.215,0.9715,189.3926,12.4231,1.3967,-5369.0,2732.0,-4466.5,-117.25,1.3125,2.0017,7.3487,62.8889,2.8778,0.247,3.4687,84.3324,8.9336,50.5527,64.2529,49.4474,65.1893,86.8595,116
29  28,2972.78,2431.57,2190.4889,1059.4389999999999,0.8614,100.0,102.1178,0.1216,1.5438,0.0065,0.0032,0.9589,206.4436,0.0,11.3083,403.4278,9.7745,0.9574,196.6691,12.5473,1.409,-5217.5,3015.25,-4005.0,678.5,0.8902,1.8019999999999998,3.2887,66.1667,2.2,0.1818,3.6319999999999997,84.7166,8.8388,50.3916,64
30  29,2981.85,2529.11,2180.3778,1208.7411,1.2998,100.0,100.2789,0.1209,1.42,-0.0016,0.0138,0.9620000000000001,198.7199,0.0,6.8715,414.1525,8.6625,0.9629,190.0574,12.5473,1.3978,-5416.0,2588.0,-3793.25,-788.5,1.2905,1.9817,7.1133,65.6333,2.6444,0.23399999999999999,3.2431,85.6039,8.7588,50.17899999999
31  30,2975.88,2489.7,2191.6667,1153.9011,1.2569,100.0,100.6767,0.121,1.4481,-0.0134,0.0177,0.9515,199.6605,0.0,7.3997,422.9413,9.8631,0.9596,189.7974,12.4222,1.4007,-5500.5,2739.5,-3717.0,-947.0,1.2919999999999998,1.9888,7.308,68.6444,2.1667,0.1576,3.4577,101.9072,23.3453,57.6663,74.7081,42.3337,66.8426
32  31,3058.61,2492.36,2180.3778,1208.7411,1.2998,100.0,100.2789,0.1209,1.5602,0.0041,-0.0056,0.9595,206.5154,0.0,11.3994,417.3887,9.75,0.9582,196.7655,12.5473,1.4049,-5497.75,2535.75,-4835.25,-372.75,1.2717,1.9773,7.1333,65.3667,2.9,0.2279,3.3036,83.9737,8.8033,50.3329,64.485,49.6671,66.6916,88.2847,118.6
33  32,3047.19,2524.18,2197.3111,969.8910000000001,1.3015,100.0,105.3911,0.1201,1.5837,-0.0266,-0.0029,0.9581,199.4722,0.0,11.6646,414.2214,9.7447,0.972,189.7275,12.4222,1.3892,-5507.75,2637.25,114.0,-2021.0,1.2803,1.9745,7.0992,62.6667,3.0889,0.251,3.2979,86.3472,9.1283,50.1869,64.0917,49.8131,66.0625,86
34  33,2981.31,2566.35,2197.3111,969.8910000000001,1.3015,100.0,105.3911,0.1201,1.5047,-0.0216,0.0064,0.9721,199.2977,0.0,9.5945,417.11400000000003,9.672,0.9726,189.6257,12.4278,1.4432,-2622.5,1550.0,-1993.75,226.5,0.446,0.8912,1.5313,65.4556,2.0333,0.184,3.4811,85.6591,8.8249,50.3878,64.1536,49.6122,65
35  34,2963.83,2457.64,2225.1777,1457.7934,1.2414,100.0,110.2789,0.1211,1.5374,-0.0177,0.017,0.9757,199.5802,0.0,7.202999999999999,419.3116,9.6398,0.9651,189.9403,12.5473,1.3765,-6010.25,2805.75,-3892.75,893.0,0.904,1.7778,3.1147,66.8889,2.2667,0.1948,3.5241,84.3324,8.9336,50.5527,64.2529,49.4474,65.1893
36  35,3040.72,2477.35,2191.6667,1153.9011,1.2569,100.0,100.6767,0.121,1.3475,-0.0152,0.0115,0.9517,200.1074,0.0,4.5248,400.7535,10.4472,0.9715,189.6602,12.2749,1.4033,-5548.0,2637.0,-3683.0,244.75,1.3195,2.0092,7.457000000000001,66.4778,2.7778,0.2291,3.4393,85.2487,10.0686,50.6755,64.3085,49.3245,65.190
37  36,2989.47,2445.44,2223.1667,1522.5535,1.1981,100.0,110.6333,0.121,1.5316,-0.0194,0.0088,0.9614,199.2808,0.0,7.4758,422.0675,10.499,0.9758,188.7817,12.4231,1.4011,-1362.0,710.25,-1016.25,260.5,0.22399999999999998,0.4407,0.7198,65.2667,2.6889,0.2093,3.3877,85.6039,8.7588,50.17899999999995,64.1668,4
38  37,2981.54,2302.46,2221.8445,1416.8211,1.1758,100.0,111.6278,0.121,1.44,-0.0163,0.0211,0.9651,199.6613,0.0,5.803999999999999,410.4105,9.4864,0.9629,190.1748,12.4231,1.4002,-5452.75,2574.25,-3352.75,-962.75,1.284,1.9728,7.1523,65.2222,2.7222,0.2151,3.3276,85.6039,8.7588,50.17899999999995,64.1668,49.
```

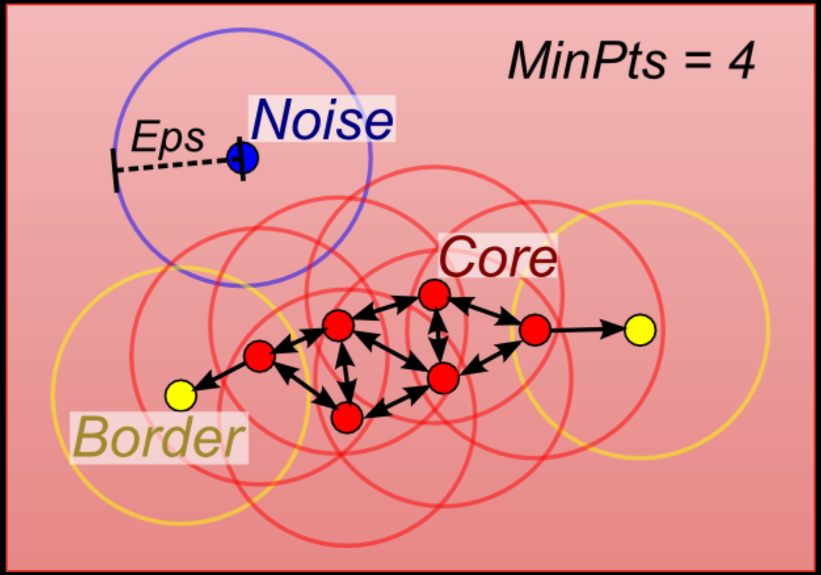
Anomaly detection

- Anomalies are patterns in data that do not conform to a well defined notion of normal behavior
- Anomaly detection techniques:
  - Classification based
  - Nearest Neighbor based
  - Clustering based
  - Statistical
  - Information theoretic
  - Spectral



Anomaly detection

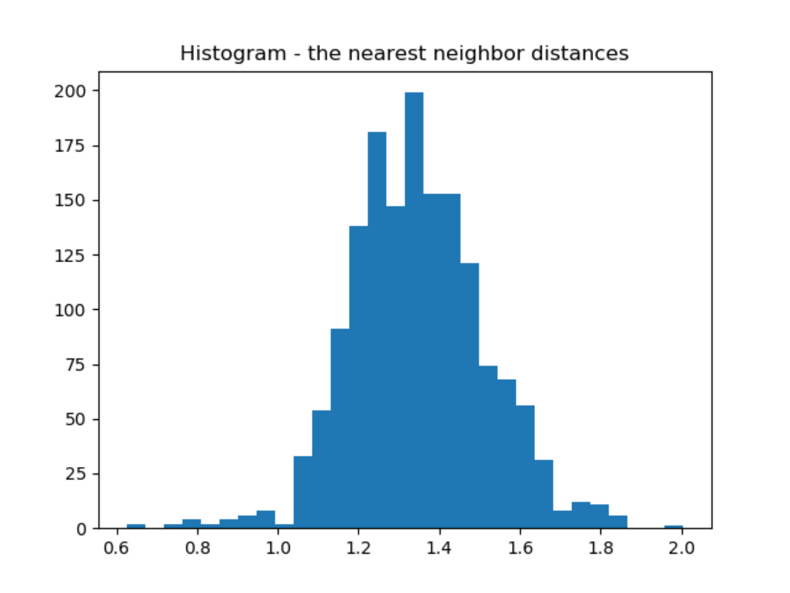
- DBSCAN (Density-based spatial clustering of applications with noise) algorithm
    - Core point
    - Border point
    - Noise point

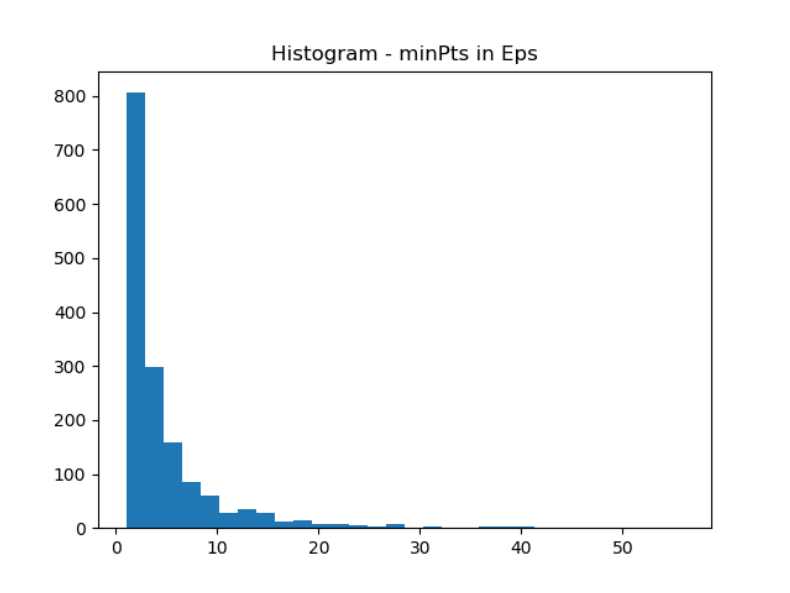


https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Anomaly detection

- Parameter tuning:
  - In DBSCAN learning algorithm, thare are two parameters: eps and minPts





Eps = 1.45

minPts = 6

Anomaly detection

- Unbalanced dataset – accuracy is not a good measure
- It‘s important to detect the fault in a timely manner

	Precision	**Recall**	**F1-score**	**Support**
Outlier	0.07	0.51	0.13	104
Not outlier	0.94	0.54	0.69	1463
Avg/total	0.88	0.54	0.65	1567

CONCLUSION

- Learn while working
- Math is fun
- Advantage of creative mindset
- There is no general approach for solving machine learning problems

THANK YOU!