SC1_Proj

Alessio

13 January 2020

Introduction

One of the most common types of cancer diagnosed in women is breast cancer. There are multiple tests that people are subjected to, but one of the most indicative ones is fine needle aspiration which involes extracting a sample of cells to be examined under a microscope. Multiple numerical metrics are computed from the obtained images. The aim is to use the extracted metrics to make accurate diagnoses.

The dataset consists of 569 images which have been processed as described and a total of 30 variables have been computed for each observation.

The aim of this report is to implement a number of classification algorithms, use them to obtain predictions, and compare their performances.

TODO: describe

```r
data <- read_csv("../data/data.csv")
glimpse(data) %>%
  kable()
```

```
## Observations: 569
## Variables: 33
## $ id                      <dbl> 842302, 842517, 84300903, 84348301, 8435840...
## $ diagnosis               <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M"...
## $ radius_mean             <dbl> 17.990, 20.570, 19.690, 11.420, 20.290, 12....
## $ texture_mean            <dbl> 10.38, 17.77, 21.25, 20.38, 14.34, 15.70, 1...
## $ perimeter_mean          <dbl> 122.80, 132.90, 130.00, 77.58, 135.10, 82.5...
## $ area_mean               <dbl> 1001.0, 1326.0, 1203.0, 386.1, 1297.0, 477....
## $ smoothness_mean         <dbl> 0.11840, 0.08474, 0.10960, 0.14250, 0.10030...
## $ compactness_mean        <dbl> 0.27760, 0.07864, 0.15990, 0.28390, 0.13280...
## $ concavity_mean          <dbl> 0.30010, 0.08690, 0.19740, 0.24140, 0.19800...
## $ `concave points_mean`   <dbl> 0.14710, 0.07017, 0.12790, 0.10520, 0.10430...
## $ symmetry_mean           <dbl> 0.2419, 0.1812, 0.2069, 0.2597, 0.1809, 0.2...
## $ fractal_dimension_mean  <dbl> 0.07871, 0.05667, 0.05999, 0.09744, 0.05883...
## $ radius_se               <dbl> 1.0950, 0.5435, 0.7456, 0.4956, 0.7572, 0.3...
## $ texture_se              <dbl> 0.9053, 0.7339, 0.7869, 1.1560, 0.7813, 0.8...
## $ perimeter_se            <dbl> 8.589, 3.398, 4.585, 3.445, 5.438, 2.217, 3...
## $ area_se                 <dbl> 153.40, 74.08, 94.03, 27.23, 94.44, 27.19, ...
## $ smoothness_se           <dbl> 0.006399, 0.005225, 0.006150, 0.009110, 0.0...
## $ compactness_se          <dbl> 0.049040, 0.013080, 0.040060, 0.074580, 0.0...
## $ concavity_se            <dbl> 0.05373, 0.01860, 0.03832, 0.05661, 0.05688...
## $ `concave points_se`     <dbl> 0.015870, 0.013400, 0.020580, 0.018670, 0.0...
## $ symmetry_se             <dbl> 0.03003, 0.01389, 0.02250, 0.05963, 0.01756...
## $ fractal_dimension_se    <dbl> 0.006193, 0.003532, 0.004571, 0.009208, 0.0...
## $ radius_worst            <dbl> 25.38, 24.99, 23.57, 14.91, 22.54, 15.47, 2...
```

```
## $ texture_worst           <dbl> 17.33, 23.41, 25.53, 26.50, 16.67, 23.75, 2...
## $ perimeter_worst         <dbl> 184.60, 158.80, 152.50, 98.87, 152.20, 103....
## $ area_worst              <dbl> 2019.0, 1956.0, 1709.0, 567.7, 1575.0, 741....
## $ smoothness_worst        <dbl> 0.1622, 0.1238, 0.1444, 0.2098, 0.1374, 0.1...
## $ compactness_worst       <dbl> 0.6656, 0.1866, 0.4245, 0.8663, 0.2050, 0.5...
## $ concavity_worst         <dbl> 0.71190, 0.24160, 0.45040, 0.68690, 0.40000...
## $ `concave points_worst`  <dbl> 0.26540, 0.18600, 0.24300, 0.25750, 0.16250...
## $ symmetry_worst          <dbl> 0.4601, 0.2750, 0.3613, 0.6638, 0.2364, 0.3...
## $ fractal_dimension_worst <dbl> 0.11890, 0.08902, 0.08758, 0.17300, 0.07678...
## $ X33                     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
```

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness
842302	M	17.990	10.38	122.80	1001.0	0.11840	
842517	M	20.570	17.77	132.90	1326.0	0.08474	
84300903	M	19.690	21.25	130.00	1203.0	0.10960	
84348301	M	11.420	20.38	77.58	386.1	0.14250	
84358402	M	20.290	14.34	135.10	1297.0	0.10030	
843786	M	12.450	15.70	82.57	477.1	0.12780	
844359	M	18.250	19.98	119.60	1040.0	0.09463	
84458202	M	13.710	20.83	90.20	577.9	0.11890	
844981	M	13.000	21.82	87.50	519.8	0.12730	
84501001	M	12.460	24.04	83.97	475.9	0.11860	
845636	M	16.020	23.24	102.70	797.8	0.08206	
84610002	M	15.780	17.89	103.60	781.0	0.09710	
846226	M	19.170	24.80	132.40	1123.0	0.09740	
846381	M	15.850	23.95	103.70	782.7	0.08401	
84667401	M	13.730	22.61	93.60	578.3	0.11310	
84799002	M	14.540	27.54	96.73	658.8	0.11390	
848406	M	14.680	20.13	94.74	684.5	0.09867	
84862001	M	16.130	20.68	108.10	798.8	0.11700	
849014	M	19.810	22.15	130.00	1260.0	0.09831	
8510426	B	13.540	14.36	87.46	566.3	0.09779	
8510653	B	13.080	15.71	85.63	520.0	0.10750	
8510824	B	9.504	12.44	60.34	273.9	0.10240	
8511133	M	15.340	14.26	102.50	704.4	0.10730	
851509	M	21.160	23.04	137.20	1404.0	0.09428	
852552	M	16.650	21.38	110.00	904.6	0.11210	
852631	M	17.140	16.40	116.00	912.7	0.11860	
852763	M	14.580	21.53	97.41	644.8	0.10540	
852781	M	18.610	20.25	122.10	1094.0	0.09440	
852973	M	15.300	25.27	102.40	732.4	0.10820	
853201	M	17.570	15.05	115.00	955.1	0.09847	
853401	M	18.630	25.11	124.80	1088.0	0.10640	
853612	M	11.840	18.70	77.93	440.6	0.11090	
85382601	M	17.020	23.98	112.80	899.3	0.11970	
854002	M	19.270	26.47	127.90	1162.0	0.09401	
854039	M	16.130	17.88	107.00	807.2	0.10400	
854253	M	16.740	21.59	110.10	869.5	0.09610	
854268	M	14.250	21.72	93.63	633.0	0.09823	
854941	B	13.030	18.42	82.61	523.8	0.08983	
855133	M	14.990	25.20	95.54	698.8	0.09387	
855138	M	13.480	20.82	88.40	559.2	0.10160	
855167	M	13.440	21.58	86.18	563.0	0.08162	
855563	M	10.950	21.35	71.90	371.1	0.12270	

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness
855625	M	19.070	24.81	128.30	1104.0	0.09081	
856106	M	13.280	20.28	87.32	545.2	0.10410	
85638502	M	13.170	21.81	85.42	531.5	0.09714	
857010	M	18.650	17.60	123.70	1076.0	0.10990	
85713702	B	8.196	16.84	51.71	201.9	0.08600	
85715	M	13.170	18.66	85.98	534.6	0.11580	
857155	B	12.050	14.63	78.04	449.3	0.10310	
857156	B	13.490	22.30	86.91	561.0	0.08752	
857343	B	11.760	21.60	74.72	427.9	0.08637	
857373	B	13.640	16.34	87.21	571.8	0.07685	
857374	B	11.940	18.24	75.71	437.6	0.08261	
857392	M	18.220	18.70	120.30	1033.0	0.11480	
857438	M	15.100	22.02	97.26	712.8	0.09056	
85759902	B	11.520	18.75	73.34	409.0	0.09524	
857637	M	19.210	18.57	125.50	1152.0	0.10530	
857793	M	14.710	21.59	95.55	656.9	0.11370	
857810	B	13.050	19.31	82.61	527.2	0.08060	
858477	B	8.618	11.79	54.34	224.5	0.09752	
858970	B	10.170	14.88	64.55	311.9	0.11340	
858981	B	8.598	20.98	54.66	221.8	0.12430	
858986	M	14.250	22.15	96.42	645.7	0.10490	
859196	B	9.173	13.86	59.20	260.9	0.07721	
85922302	M	12.680	23.84	82.69	499.0	0.11220	
859283	M	14.780	23.94	97.40	668.3	0.11720	
859464	B	9.465	21.01	60.11	269.4	0.10440	
859465	B	11.310	19.04	71.80	394.1	0.08139	
859471	B	9.029	17.33	58.79	250.5	0.10660	
859487	B	12.780	16.49	81.37	502.5	0.09831	
859575	M	18.940	21.31	123.60	1130.0	0.09009	
859711	B	8.888	14.64	58.79	244.0	0.09783	
859717	M	17.200	24.52	114.20	929.4	0.10710	
859983	M	13.800	15.79	90.43	584.1	0.10070	
8610175	B	12.310	16.52	79.19	470.9	0.09172	
8610404	M	16.070	19.65	104.10	817.7	0.09168	
8610629	B	13.530	10.94	87.91	559.2	0.12910	
8610637	M	18.050	16.15	120.20	1006.0	0.10650	
8610862	M	20.180	23.97	143.70	1245.0	0.12860	
8610908	B	12.860	18.00	83.19	506.3	0.09934	
861103	B	11.450	20.97	73.81	401.5	0.11020	
8611161	B	13.340	15.86	86.49	520.0	0.10780	
8611555	M	25.220	24.91	171.50	1878.0	0.10630	
8611792	M	19.100	26.29	129.10	1132.0	0.12150	
8612080	B	12.000	15.65	76.95	443.3	0.09723	
8612399	M	18.460	18.52	121.10	1075.0	0.09874	
86135501	M	14.480	21.46	94.25	648.2	0.09444	
86135502	M	19.020	24.59	122.00	1076.0	0.09029	
861597	B	12.360	21.80	79.78	466.1	0.08772	
861598	B	14.640	15.24	95.77	651.9	0.11320	
861648	B	14.620	24.02	94.57	662.7	0.08974	
861799	M	15.370	22.76	100.20	728.2	0.09200	
861853	B	13.270	14.76	84.74	551.7	0.07355	
862009	B	13.450	18.30	86.60	555.1	0.10220	

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness
862028	M	15.060	19.83	100.30	705.6	0.10390	
86208	M	20.260	23.03	132.40	1264.0	0.09078	
86211	B	12.180	17.84	77.79	451.1	0.10450	
862261	B	9.787	19.94	62.11	294.5	0.10240	
862485	B	11.600	12.84	74.34	412.6	0.08983	
862548	M	14.420	19.77	94.48	642.5	0.09752	
862717	M	13.610	24.98	88.05	582.7	0.09488	
862722	B	6.981	13.43	43.79	143.5	0.11700	
862965	B	12.180	20.52	77.22	458.7	0.08013	
862980	B	9.876	19.40	63.95	298.3	0.10050	
862989	B	10.490	19.29	67.41	336.1	0.09989	
863030	M	13.110	15.56	87.21	530.2	0.13980	
863031	B	11.640	18.33	75.17	412.5	0.11420	
863270	B	12.360	18.54	79.01	466.7	0.08477	
86355	M	22.270	19.67	152.80	1509.0	0.13260	
864018	B	11.340	21.26	72.48	396.5	0.08759	
864033	B	9.777	16.99	62.50	290.2	0.10370	
86408	B	12.630	20.76	82.15	480.4	0.09933	
86409	B	14.260	19.65	97.83	629.9	0.07837	
864292	B	10.510	20.19	68.64	334.2	0.11220	
864496	B	8.726	15.83	55.84	230.9	0.11500	
864685	B	11.930	21.53	76.53	438.6	0.09768	
864726	B	8.950	15.76	58.74	245.2	0.09462	
864729	M	14.870	16.67	98.64	682.5	0.11620	
864877	M	15.780	22.91	105.70	782.6	0.11550	
865128	M	17.950	20.01	114.20	982.0	0.08402	
865137	B	11.410	10.82	73.34	403.3	0.09373	
86517	M	18.660	17.12	121.40	1077.0	0.10540	
865423	M	24.250	20.20	166.20	1761.0	0.14470	
865432	B	14.500	10.89	94.28	640.7	0.11010	
865468	B	13.370	16.39	86.10	553.5	0.07115	
86561	B	13.850	17.21	88.44	588.7	0.08785	
866083	M	13.610	24.69	87.76	572.6	0.09258	
866203	M	19.000	18.91	123.40	1138.0	0.08217	
866458	B	15.100	16.39	99.58	674.5	0.11500	
866674	M	19.790	25.12	130.40	1192.0	0.10150	
866714	B	12.190	13.29	79.08	455.8	0.10660	
8670	M	15.460	19.48	101.70	748.9	0.10920	
86730502	M	16.160	21.54	106.20	809.8	0.10080	
867387	B	15.710	13.93	102.00	761.7	0.09462	
867739	M	18.450	21.91	120.20	1075.0	0.09430	
868202	M	12.770	22.47	81.72	506.3	0.09055	
868223	B	11.710	16.67	74.72	423.6	0.10510	
868682	B	11.430	15.39	73.06	399.8	0.09639	
868826	M	14.950	17.57	96.85	678.1	0.11670	
868871	B	11.280	13.39	73.00	384.8	0.11640	
868999	B	9.738	11.97	61.24	288.5	0.09250	
869104	M	16.110	18.05	105.10	813.0	0.09721	
869218	B	11.430	17.31	73.66	398.0	0.10920	
869224	B	12.900	15.92	83.74	512.2	0.08677	
869254	B	10.750	14.97	68.26	355.3	0.07793	
869476	B	11.900	14.65	78.11	432.8	0.11520	

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness
869691	M	11.800	16.58	78.99	432.0	0.10910	
86973701	B	14.950	18.77	97.84	689.5	0.08138	
86973702	B	14.440	15.18	93.97	640.1	0.09970	
869931	B	13.740	17.91	88.12	585.0	0.07944	
871001501	B	13.000	20.78	83.51	519.4	0.11350	
871001502	B	8.219	20.70	53.27	203.9	0.09405	
8710441	B	9.731	15.34	63.78	300.2	0.10720	
87106	B	11.150	13.08	70.87	381.9	0.09754	
8711002	B	13.150	15.34	85.31	538.9	0.09384	
8711003	B	12.250	17.94	78.27	460.3	0.08654	
8711202	M	17.680	20.74	117.40	963.7	0.11150	
8711216	B	16.840	19.46	108.40	880.2	0.07445	
871122	B	12.060	12.74	76.84	448.6	0.09311	
871149	B	10.900	12.96	68.69	366.8	0.07515	
8711561	B	11.750	20.18	76.10	419.8	0.10890	
8711803	M	19.190	15.94	126.30	1157.0	0.08694	
871201	M	19.590	18.15	130.70	1214.0	0.11200	
8712064	B	12.340	22.22	79.85	464.5	0.10120	
8712289	M	23.270	22.04	152.10	1686.0	0.08439	
8712291	B	14.970	19.76	95.50	690.2	0.08421	
87127	B	10.800	9.71	68.77	357.6	0.09594	
8712729	M	16.780	18.80	109.30	886.3	0.08865	
8712766	M	17.470	24.68	116.10	984.6	0.10490	
8712853	B	14.970	16.95	96.22	685.9	0.09855	
87139402	B	12.320	12.39	78.85	464.1	0.10280	
87163	M	13.430	19.63	85.84	565.4	0.09048	
87164	M	15.460	11.89	102.50	736.9	0.12570	
871641	B	11.080	14.71	70.21	372.7	0.10060	
871642	B	10.660	15.15	67.49	349.6	0.08792	
872113	B	8.671	14.45	54.42	227.2	0.09138	
872608	B	9.904	18.06	64.60	302.4	0.09699	
87281702	M	16.460	20.11	109.30	832.9	0.09831	
873357	B	13.010	22.22	82.01	526.4	0.06251	
873586	B	12.810	13.06	81.29	508.8	0.08739	
873592	M	27.220	21.87	182.10	2250.0	0.10940	
873593	M	21.090	26.57	142.70	1311.0	0.11410	
873701	M	15.700	20.31	101.20	766.6	0.09597	
873843	B	11.410	14.92	73.53	402.0	0.09059	
873885	M	15.280	22.41	98.92	710.6	0.09057	
874158	B	10.080	15.11	63.76	317.5	0.09267	
874217	M	18.310	18.58	118.60	1041.0	0.08588	
874373	B	11.710	17.19	74.68	420.3	0.09774	
874662	B	11.810	17.39	75.27	428.9	0.10070	
874839	B	12.300	15.90	78.83	463.7	0.08080	
874858	M	14.220	23.12	94.37	609.9	0.10750	
875093	B	12.770	21.41	82.02	507.4	0.08749	
875099	B	9.720	18.22	60.73	288.1	0.06950	
875263	M	12.340	26.86	81.15	477.4	0.10340	
87556202	M	14.860	23.21	100.40	671.4	0.10440	
875878	B	12.910	16.33	82.53	516.4	0.07941	
875938	M	13.770	22.29	90.63	588.9	0.12000	
877159	M	18.080	21.84	117.40	1024.0	0.07371	

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness
877486	M	19.180	22.49	127.50	1148.0	0.08523	
877500	M	14.450	20.22	94.49	642.7	0.09872	
877501	B	12.230	19.56	78.54	461.0	0.09586	
877989	M	17.540	19.32	115.10	951.6	0.08968	
878796	M	23.290	26.67	158.90	1685.0	0.11410	
87880	M	13.810	23.75	91.56	597.8	0.13230	
87930	B	12.470	18.60	81.09	481.9	0.09965	
879523	M	15.120	16.68	98.78	716.6	0.08876	
879804	B	9.876	17.27	62.92	295.4	0.10890	
879830	M	17.010	20.26	109.70	904.3	0.08772	
8810158	B	13.110	22.54	87.02	529.4	0.10020	
8810436	B	15.270	12.91	98.17	725.5	0.08182	
881046502	M	20.580	22.14	134.70	1290.0	0.09090	
8810528	B	11.840	18.94	75.51	428.0	0.08871	
8810703	M	28.110	18.47	188.50	2499.0	0.11420	
881094802	M	17.420	25.56	114.50	948.0	0.10060	
8810955	M	14.190	23.81	92.87	610.7	0.09463	
8810987	M	13.860	16.93	90.96	578.9	0.10260	
8811523	B	11.890	18.35	77.32	432.2	0.09363	
8811779	B	10.200	17.48	65.05	321.2	0.08054	
8811842	M	19.800	21.56	129.70	1230.0	0.09383	
88119002	M	19.530	32.47	128.00	1223.0	0.08420	
8812816	B	13.650	13.16	87.88	568.9	0.09646	
8812818	B	13.560	13.90	88.59	561.3	0.10510	
8812844	B	10.180	17.53	65.12	313.1	0.10610	
8812877	M	15.750	20.25	102.60	761.3	0.10250	
8813129	B	13.270	17.02	84.55	546.4	0.08445	
88143502	B	14.340	13.47	92.51	641.2	0.09906	
88147101	B	10.440	15.46	66.62	329.6	0.10530	
88147102	B	15.000	15.51	97.45	684.5	0.08371	
88147202	B	12.620	23.97	81.35	496.4	0.07903	
881861	M	12.830	22.33	85.26	503.2	0.10880	
881972	M	17.050	19.08	113.40	895.0	0.11410	
88199202	B	11.320	27.08	71.76	395.7	0.06883	
88203002	B	11.220	33.81	70.79	386.8	0.07780	
88206102	M	20.510	27.81	134.40	1319.0	0.09159	
882488	B	9.567	15.91	60.21	279.6	0.08464	
88249602	B	14.030	21.25	89.79	603.4	0.09070	
88299702	M	23.210	26.97	153.50	1670.0	0.09509	
883263	M	20.480	21.46	132.50	1306.0	0.08355	
883270	B	14.220	27.85	92.55	623.9	0.08223	
88330202	M	17.460	39.28	113.40	920.6	0.09812	
88350402	B	13.640	15.60	87.38	575.3	0.09423	
883539	B	12.420	15.04	78.61	476.5	0.07926	
883852	B	11.300	18.19	73.93	389.4	0.09592	
88411702	B	13.750	23.77	88.54	590.0	0.08043	
884180	M	19.400	23.50	129.10	1155.0	0.10270	
884437	B	10.480	19.86	66.72	337.7	0.10700	
884448	B	13.200	17.43	84.13	541.6	0.07215	
884626	B	12.890	14.11	84.95	512.2	0.08760	
88466802	B	10.650	25.22	68.01	347.0	0.09657	
884689	B	11.520	14.93	73.87	406.3	0.10130	

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness
884948	M	20.940	23.56	138.90	1364.0	0.10070	
88518501	B	11.500	18.45	73.28	407.4	0.09345	
885429	M	19.730	19.82	130.70	1206.0	0.10620	
8860702	M	17.300	17.08	113.00	928.2	0.10080	
886226	M	19.450	19.33	126.50	1169.0	0.10350	
886452	M	13.960	17.05	91.43	602.4	0.10960	
88649001	M	19.550	28.77	133.60	1207.0	0.09260	
886776	M	15.320	17.27	103.20	713.3	0.13350	
887181	M	15.660	23.20	110.20	773.5	0.11090	
88725602	M	15.530	33.56	103.70	744.9	0.10630	
887549	M	20.310	27.06	132.90	1288.0	0.10000	
888264	M	17.350	23.06	111.00	933.1	0.08662	
888570	M	17.290	22.13	114.40	947.8	0.08999	
889403	M	15.610	19.38	100.00	758.6	0.07840	
889719	M	17.190	22.07	111.60	928.3	0.09726	
88995002	M	20.730	31.12	135.70	1419.0	0.09469	
8910251	B	10.600	18.95	69.28	346.4	0.09688	
8910499	B	13.590	21.84	87.16	561.0	0.07956	
8910506	B	12.870	16.21	82.38	512.2	0.09425	
8910720	B	10.710	20.39	69.50	344.9	0.10820	
8910721	B	14.290	16.82	90.30	632.6	0.06429	
8910748	B	11.290	13.04	72.23	388.0	0.09834	
8910988	M	21.750	20.99	147.30	1491.0	0.09401	
8910996	B	9.742	15.67	61.50	289.9	0.09037	
8911163	M	17.930	24.48	115.20	998.9	0.08855	
8911164	B	11.890	17.36	76.20	435.6	0.12250	
8911230	B	11.330	14.16	71.79	396.6	0.09379	
8911670	M	18.810	19.98	120.90	1102.0	0.08923	
8911800	B	13.590	17.84	86.24	572.3	0.07948	
8911834	B	13.850	15.18	88.99	587.4	0.09516	
8912049	M	19.160	26.60	126.20	1138.0	0.10200	
8912055	B	11.740	14.02	74.24	427.3	0.07813	
89122	M	19.400	18.18	127.20	1145.0	0.10370	
8912280	M	16.240	18.77	108.80	805.1	0.10660	
8912284	B	12.890	15.70	84.08	516.6	0.07818	
8912521	B	12.580	18.40	79.83	489.0	0.08393	
8912909	B	11.940	20.76	77.87	441.0	0.08605	
8913	B	12.890	13.12	81.89	515.9	0.06955	
8913049	B	11.260	19.96	73.72	394.1	0.08020	
89143601	B	11.370	18.89	72.17	396.0	0.08713	
89143602	B	14.410	19.73	96.03	651.0	0.08757	
8915	B	14.960	19.10	97.03	687.3	0.08992	
891670	B	12.950	16.02	83.14	513.7	0.10050	
891703	B	11.850	17.46	75.54	432.7	0.08372	
891716	B	12.720	13.78	81.78	492.1	0.09667	
891923	B	13.770	13.27	88.06	582.7	0.09198	
891936	B	10.910	12.35	69.14	363.7	0.08518	
892189	M	11.760	18.14	75.00	431.1	0.09968	
892214	B	14.260	18.17	91.22	633.1	0.06576	
892399	B	10.510	23.09	66.85	334.2	0.10150	
892438	M	19.530	18.90	129.50	1217.0	0.11500	
892604	B	12.460	19.89	80.43	471.3	0.08451	

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness
89263202	M	20.090	23.86	134.70	1247.0	0.10800	
892657	B	10.490	18.61	66.86	334.3	0.10680	
89296	B	11.460	18.16	73.59	403.1	0.08853	
893061	B	11.600	24.49	74.23	417.2	0.07474	
89344	B	13.200	15.82	84.07	537.3	0.08511	
89346	B	9.000	14.40	56.36	246.3	0.07005	
893526	B	13.500	12.71	85.69	566.2	0.07376	
893548	B	13.050	13.84	82.71	530.6	0.08352	
893783	B	11.700	19.11	74.33	418.7	0.08814	
89382601	B	14.610	15.69	92.68	664.9	0.07618	
89382602	B	12.760	13.37	82.29	504.1	0.08794	
893988	B	11.540	10.72	73.73	409.1	0.08597	
894047	B	8.597	18.60	54.09	221.2	0.10740	
894089	B	12.490	16.85	79.19	481.6	0.08511	
894090	B	12.180	14.08	77.25	461.4	0.07734	
894326	M	18.220	18.87	118.70	1027.0	0.09746	
894329	B	9.042	18.90	60.07	244.5	0.09968	
894335	B	12.430	17.00	78.60	477.3	0.07557	
894604	B	10.250	16.18	66.52	324.2	0.10610	
894618	M	20.160	19.66	131.10	1274.0	0.08020	
894855	B	12.860	13.32	82.82	504.8	0.11340	
895100	M	20.340	21.51	135.90	1264.0	0.11700	
89511501	B	12.200	15.21	78.01	457.9	0.08673	
89511502	B	12.670	17.30	81.25	489.9	0.10280	
89524	B	14.110	12.88	90.03	616.5	0.09309	
895299	B	12.030	17.93	76.09	446.0	0.07683	
8953902	M	16.270	20.71	106.90	813.7	0.11690	
895633	M	16.260	21.88	107.50	826.8	0.11650	
896839	M	16.030	15.51	105.80	793.2	0.09491	
896864	B	12.980	19.35	84.52	514.0	0.09579	
897132	B	11.220	19.86	71.94	387.3	0.10540	
897137	B	11.250	14.78	71.38	390.0	0.08306	
897374	B	12.300	19.02	77.88	464.4	0.08313	
89742801	M	17.060	21.00	111.80	918.6	0.11190	
897604	B	12.990	14.23	84.08	514.3	0.09462	
897630	M	18.770	21.43	122.90	1092.0	0.09116	
897880	B	10.050	17.53	64.41	310.8	0.10070	
89812	M	23.510	24.27	155.10	1747.0	0.10690	
89813	B	14.420	16.54	94.15	641.2	0.09751	
898143	B	9.606	16.84	61.64	280.5	0.08481	
89827	B	11.060	14.96	71.49	373.9	0.10330	
898431	M	19.680	21.68	129.90	1194.0	0.09797	
89864002	B	11.710	15.45	75.03	420.3	0.11500	
898677	B	10.260	14.71	66.20	321.6	0.09882	
898678	B	12.060	18.90	76.66	445.3	0.08386	
89869	B	14.760	14.74	94.87	668.7	0.08875	
898690	B	11.470	16.03	73.02	402.7	0.09076	
899147	B	11.950	14.96	77.23	426.7	0.11580	
899187	B	11.660	17.07	73.70	421.0	0.07561	
899667	M	15.750	19.22	107.10	758.6	0.12430	
899987	M	25.730	17.46	174.20	2010.0	0.11490	
9010018	M	15.080	25.74	98.00	716.6	0.10240	

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness
901011	B	11.140	14.07	71.24	384.6	0.07274	
9010258	B	12.560	19.07	81.92	485.8	0.08760	
9010259	B	13.050	18.59	85.09	512.0	0.10820	
901028	B	13.870	16.21	88.52	593.7	0.08743	
9010333	B	8.878	15.49	56.74	241.0	0.08293	
901034301	B	9.436	18.32	59.82	278.6	0.10090	
901034302	B	12.540	18.07	79.42	491.9	0.07436	
901041	B	13.300	21.57	85.24	546.1	0.08582	
9010598	B	12.760	18.84	81.87	496.6	0.09676	
9010872	B	16.500	18.29	106.60	838.1	0.09686	
9010877	B	13.400	16.95	85.48	552.4	0.07937	
901088	M	20.440	21.78	133.80	1293.0	0.09150	
9011494	M	20.200	26.83	133.70	1234.0	0.09905	
9011495	B	12.210	18.02	78.31	458.4	0.09231	
9011971	M	21.710	17.25	140.90	1546.0	0.09384	
9012000	M	22.010	21.90	147.20	1482.0	0.10630	
9012315	M	16.350	23.29	109.00	840.4	0.09742	
9012568	B	15.190	13.21	97.65	711.8	0.07963	
9012795	M	21.370	15.10	141.30	1386.0	0.10010	
901288	M	20.640	17.35	134.80	1335.0	0.09446	
9013005	B	13.690	16.07	87.84	579.1	0.08302	
901303	B	16.170	16.07	106.30	788.5	0.09880	
901315	B	10.570	20.22	70.15	338.3	0.09073	
9013579	B	13.460	28.21	85.89	562.1	0.07517	
9013594	B	13.660	15.15	88.27	580.6	0.08268	
9013838	M	11.080	18.83	73.30	361.6	0.12160	
901549	B	11.270	12.96	73.16	386.3	0.12370	
901836	B	11.040	14.93	70.67	372.7	0.07987	
90250	B	12.050	22.72	78.75	447.8	0.06935	
90251	B	12.390	17.48	80.64	462.9	0.10420	
902727	B	13.280	13.72	85.79	541.8	0.08363	
90291	M	14.600	23.29	93.97	664.7	0.08682	
902975	B	12.210	14.09	78.78	462.0	0.08108	
902976	B	13.880	16.16	88.37	596.6	0.07026	
903011	B	11.270	15.50	73.38	392.0	0.08365	
90312	M	19.550	23.21	128.90	1174.0	0.10100	
90317302	B	10.260	12.22	65.75	321.6	0.09996	
903483	B	8.734	16.84	55.27	234.3	0.10390	
903507	M	15.490	19.97	102.40	744.7	0.11600	
903516	M	21.610	22.28	144.40	1407.0	0.11670	
903554	B	12.100	17.72	78.07	446.2	0.10290	
903811	B	14.060	17.18	89.75	609.1	0.08045	
90401601	B	13.510	18.89	88.10	558.1	0.10590	
90401602	B	12.800	17.46	83.05	508.3	0.08044	
904302	B	11.060	14.83	70.31	378.2	0.07741	
904357	B	11.800	17.26	75.26	431.9	0.09087	
90439701	M	17.910	21.02	124.40	994.0	0.12300	
904647	B	11.930	10.91	76.14	442.7	0.08872	
904689	B	12.960	18.29	84.18	525.2	0.07351	
9047	B	12.940	16.17	83.18	507.6	0.09879	
904969	B	12.340	14.95	78.29	469.1	0.08682	
904971	B	10.940	18.59	70.39	370.0	0.10040	

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness
905189	B	16.140	14.86	104.30	800.0	0.09495	
905190	B	12.850	21.37	82.63	514.5	0.07551	
90524101	M	17.990	20.66	117.80	991.7	0.10360	
905501	B	12.270	17.92	78.41	466.1	0.08685	
905502	B	11.360	17.57	72.49	399.8	0.08858	
905520	B	11.040	16.83	70.92	373.2	0.10770	
905539	B	9.397	21.68	59.75	268.8	0.07969	
905557	B	14.990	22.11	97.53	693.7	0.08515	
905680	M	15.130	29.81	96.71	719.5	0.08320	
905686	B	11.890	21.17	76.39	433.8	0.09773	
905978	B	9.405	21.70	59.60	271.2	0.10440	
90602302	M	15.500	21.08	102.90	803.1	0.11200	
906024	B	12.700	12.17	80.88	495.0	0.08785	
906290	B	11.160	21.41	70.95	380.3	0.10180	
906539	B	11.570	19.04	74.20	409.7	0.08546	
906564	B	14.690	13.98	98.22	656.1	0.10310	
906616	B	11.610	16.02	75.46	408.2	0.10880	
906878	B	13.660	19.13	89.46	575.3	0.09057	
907145	B	9.742	19.12	61.93	289.7	0.10750	
907367	B	10.030	21.28	63.19	307.3	0.08117	
907409	B	10.480	14.98	67.49	333.6	0.09816	
90745	B	10.800	21.98	68.79	359.9	0.08801	
90769601	B	11.130	16.62	70.47	381.1	0.08151	
90769602	B	12.720	17.67	80.98	501.3	0.07896	
907914	M	14.900	22.53	102.10	685.0	0.09947	
907915	B	12.400	17.68	81.47	467.8	0.10540	
908194	M	20.180	19.54	133.80	1250.0	0.11330	
908445	M	18.820	21.97	123.70	1110.0	0.10180	
908469	B	14.860	16.94	94.89	673.7	0.08924	
908489	M	13.980	19.62	91.12	599.5	0.10600	
908916	B	12.870	19.54	82.67	509.2	0.09136	
909220	B	14.040	15.98	89.78	611.2	0.08458	
909231	B	13.850	19.60	88.68	592.6	0.08684	
909410	B	14.020	15.66	89.59	606.5	0.07966	
909411	B	10.970	17.20	71.73	371.5	0.08915	
909445	M	17.270	25.42	112.40	928.8	0.08331	
90944601	B	13.780	15.79	88.37	585.9	0.08817	
909777	B	10.570	18.32	66.82	340.9	0.08142	
9110127	M	18.030	16.85	117.50	990.0	0.08947	
9110720	B	11.990	24.89	77.61	441.3	0.10300	
9110732	M	17.750	28.03	117.30	981.6	0.09997	
9110944	B	14.800	17.66	95.88	674.8	0.09179	
911150	B	14.530	19.34	94.25	659.7	0.08388	
911157302	M	21.100	20.52	138.10	1384.0	0.09684	
9111596	B	11.870	21.54	76.83	432.0	0.06613	
9111805	M	19.590	25.00	127.70	1191.0	0.10320	
9111843	B	12.000	28.23	76.77	442.5	0.08437	
911201	B	14.530	13.98	93.86	644.2	0.10990	
911202	B	12.620	17.15	80.62	492.9	0.08583	
9112085	B	13.380	30.72	86.34	557.2	0.09245	
9112366	B	11.630	29.29	74.87	415.1	0.09357	
9112367	B	13.210	25.25	84.10	537.9	0.08791	

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactnes
9112594	B	13.000	25.13	82.61	520.2	0.08369	
9112712	B	9.755	28.20	61.68	290.9	0.07984	
911296201	M	17.080	27.15	111.20	930.9	0.09898	
911296202	M	27.420	26.27	186.90	2501.0	0.10840	
9113156	B	14.400	26.99	92.25	646.1	0.06995	
911320501	B	11.600	18.36	73.88	412.7	0.08508	
911320502	B	13.170	18.22	84.28	537.3	0.07466	
9113239	B	13.240	20.13	86.87	542.9	0.08284	
9113455	B	13.140	20.74	85.98	536.9	0.08675	
9113514	B	9.668	18.10	61.06	286.3	0.08311	
9113538	M	17.600	23.33	119.00	980.5	0.09289	
911366	B	11.620	18.18	76.38	408.8	0.11750	
9113778	B	9.667	18.49	61.49	289.1	0.08946	
9113816	B	12.040	28.14	76.85	449.9	0.08752	
911384	B	14.920	14.93	96.45	686.9	0.08098	
9113846	B	12.270	29.97	77.42	465.4	0.07699	
911391	B	10.880	15.62	70.41	358.9	0.10070	
911408	B	12.830	15.73	82.89	506.9	0.09040	
911654	B	14.200	20.53	92.41	618.4	0.08931	
911673	B	13.900	16.62	88.97	599.4	0.06828	
911685	B	11.490	14.59	73.99	404.9	0.10460	
911916	M	16.250	19.51	109.80	815.8	0.10260	
912193	B	12.160	18.03	78.29	455.3	0.09087	
91227	B	13.900	19.24	88.73	602.9	0.07991	
912519	B	13.470	14.06	87.32	546.3	0.10710	
912558	B	13.700	17.64	87.76	571.1	0.09950	
912600	B	15.730	11.28	102.80	747.2	0.10430	
913063	B	12.450	16.41	82.85	476.7	0.09514	
913102	B	14.640	16.85	94.21	666.0	0.08641	
913505	M	19.440	18.82	128.10	1167.0	0.10890	
913512	B	11.680	16.17	75.49	420.5	0.11280	
913535	M	16.690	20.20	107.10	857.6	0.07497	
91376701	B	12.250	22.44	78.18	466.5	0.08192	
91376702	B	17.850	13.23	114.60	992.1	0.07838	
914062	M	18.010	20.56	118.40	1007.0	0.10010	
914101	B	12.460	12.83	78.83	477.3	0.07372	
914102	B	13.160	20.54	84.06	538.7	0.07335	
914333	B	14.870	20.21	96.12	680.9	0.09587	
914366	B	12.650	18.17	82.69	485.6	0.10760	
914580	B	12.470	17.31	80.45	480.1	0.08928	
914769	M	18.490	17.52	121.30	1068.0	0.10120	
91485	M	20.590	21.24	137.80	1320.0	0.10850	
914862	B	15.040	16.74	98.73	689.4	0.09883	
91504	M	13.820	24.49	92.33	595.9	0.11620	
91505	B	12.540	16.32	81.25	476.3	0.11580	
915143	M	23.090	19.83	152.10	1682.0	0.09342	
915186	B	9.268	12.87	61.49	248.7	0.16340	
915276	B	9.676	13.14	64.12	272.5	0.12550	
91544001	B	12.220	20.04	79.47	453.1	0.10960	
91544002	B	11.060	17.12	71.25	366.5	0.11940	
915452	B	16.300	15.70	104.70	819.8	0.09427	
915460	M	15.460	23.95	103.80	731.3	0.11830	

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness
91550	B	11.740	14.69	76.31	426.0	0.08099	
915664	B	14.810	14.70	94.66	680.7	0.08472	
915691	M	13.400	20.52	88.64	556.7	0.11060	
915940	B	14.580	13.66	94.29	658.8	0.09832	
91594602	M	15.050	19.07	97.26	701.9	0.09215	
916221	B	11.340	18.61	72.76	391.2	0.10490	
916799	M	18.310	20.58	120.80	1052.0	0.10680	
916838	M	19.890	20.26	130.50	1214.0	0.10370	
917062	B	12.880	18.22	84.45	493.1	0.12180	
917080	B	12.750	16.70	82.51	493.8	0.11250	
917092	B	9.295	13.90	59.96	257.8	0.13710	
91762702	M	24.630	21.60	165.50	1841.0	0.10300	
91789	B	11.260	19.83	71.30	388.1	0.08511	
917896	B	13.710	18.68	88.73	571.0	0.09916	
917897	B	9.847	15.68	63.00	293.2	0.09492	
91805	B	8.571	13.10	54.53	221.3	0.10360	
91813701	B	13.460	18.75	87.44	551.1	0.10750	
91813702	B	12.340	12.27	78.94	468.5	0.09003	
918192	B	13.940	13.17	90.31	594.2	0.12480	
918465	B	12.070	13.44	77.83	445.2	0.11000	
91858	B	11.750	17.56	75.89	422.9	0.10730	
91903901	B	11.670	20.02	75.21	416.2	0.10160	
91903902	B	13.680	16.33	87.76	575.5	0.09277	
91930402	M	20.470	20.67	134.70	1299.0	0.09156	
919537	B	10.960	17.62	70.79	365.6	0.09687	
919555	M	20.550	20.86	137.80	1308.0	0.10460	
91979701	M	14.270	22.55	93.77	629.8	0.10380	
919812	B	11.690	24.44	76.37	406.4	0.12360	
921092	B	7.729	25.49	47.98	178.8	0.08098	
921362	B	7.691	25.44	48.34	170.4	0.08668	
921385	B	11.540	14.44	74.65	402.9	0.09984	
921386	B	14.470	24.99	95.81	656.4	0.08837	
921644	B	14.740	25.42	94.70	668.6	0.08275	
922296	B	13.210	28.06	84.88	538.4	0.08671	
922297	B	13.870	20.70	89.77	584.8	0.09578	
922576	B	13.620	23.23	87.19	573.2	0.09246	
922577	B	10.320	16.35	65.31	324.9	0.09434	
922840	B	10.260	16.58	65.85	320.8	0.08877	
923169	B	9.683	19.34	61.05	285.7	0.08491	
923465	B	10.820	24.21	68.89	361.6	0.08192	
923748	B	10.860	21.48	68.51	360.5	0.07431	
923780	B	11.130	22.44	71.49	378.4	0.09566	
924084	B	12.770	29.43	81.35	507.9	0.08276	
924342	B	9.333	21.94	59.01	264.0	0.09240	
924632	B	12.880	28.92	82.50	514.3	0.08123	
924934	B	10.290	27.61	65.67	321.4	0.09030	
924964	B	10.160	19.59	64.73	311.7	0.10030	
925236	B	9.423	27.88	59.26	271.3	0.08123	
925277	B	14.590	22.68	96.39	657.1	0.08473	
925291	B	11.510	23.93	74.52	403.5	0.09261	
925292	B	14.050	27.15	91.38	600.4	0.09929	
925311	B	11.200	29.37	70.67	386.0	0.07449	

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness
925622	M	15.220	30.62	103.40	716.9	0.10480	
926125	M	20.920	25.09	143.00	1347.0	0.10990	
926424	M	21.560	22.39	142.00	1479.0	0.11100	
926682	M	20.130	28.25	131.20	1261.0	0.09780	
926954	M	16.600	28.08	108.30	858.1	0.08455	
927241	M	20.600	29.33	140.10	1265.0	0.11780	
92751	B	7.760	24.54	47.92	181.0	0.05263	

```
colnames(data)[3:32] <- c('radius_m','texture_m', 'perim_m','area_m','smooth_m','compact_m','concav_m',
```

Dataset Preprocessing Visualisation and Exploration

```
colSums(is.na(data))
```

```
##           id    diagnosis     radius_m    texture_m      perim_m       area_m
##            0            0            0            0            0            0
##     smooth_m    compact_m     concav_m  concav_pt_m   symmetry_m   frac_dim_m
##            0            0            0            0            0            0
##    radius_se   texture_se     perim_se      area_se    smooth_se   compact_se
##            0            0            0            0            0            0
##    concav_se concav_pt_se  symmetry_se  frac_dim_se     radius_w    texture_w
##            0            0            0            0            0            0
##      perim_w       area_w     smooth_w    compact_w     concav_w  concav_pt_w
##            0            0            0            0            0            0
##   symmetry_w   frac_dim_w          X33
##            0            0          569
```

```
data %<>% mutate_at(vars(diagnosis), factor)
```

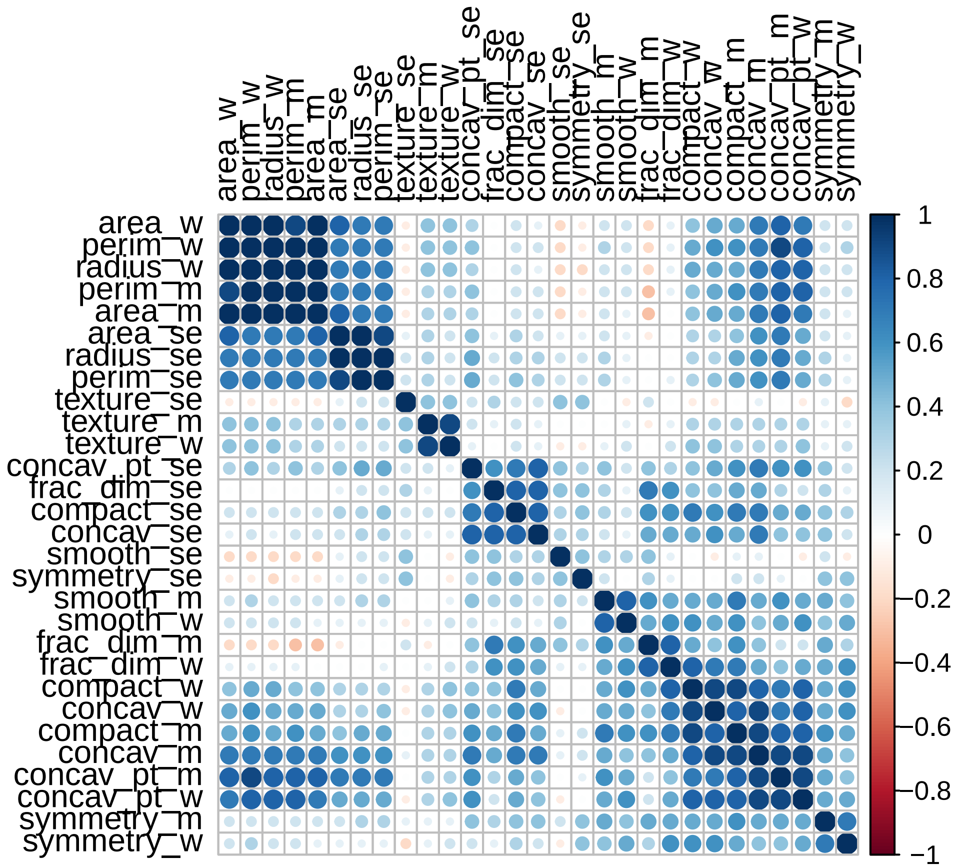
```
train <- data %>% sample_frac(0.8)
test <- anti_join(data,train, by='id')

# need ids for later
id_train <- train$id
id_test <- test$id
```

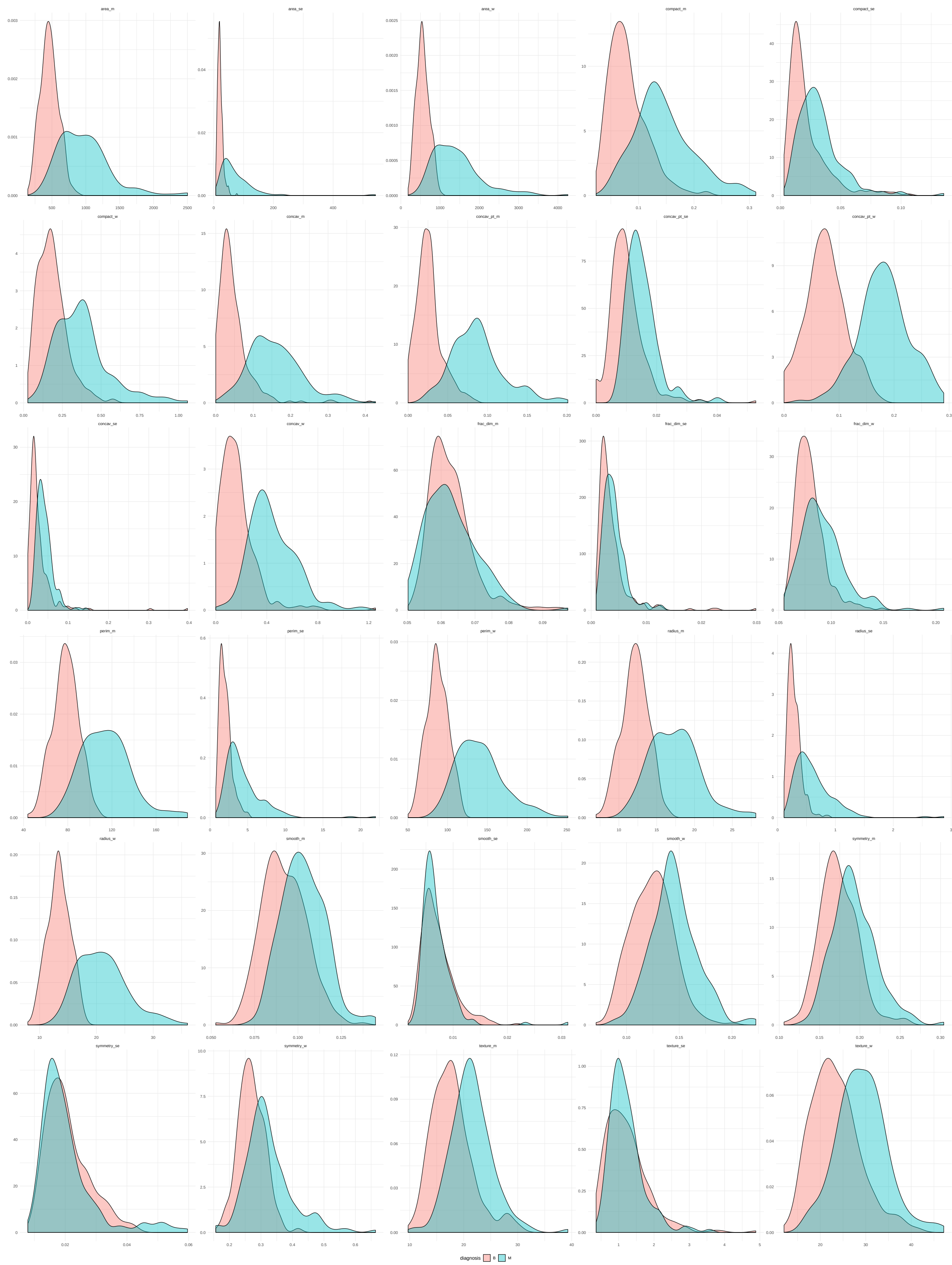
```
data %<>%
  dplyr::select(-c(id, X33))
train %<>%
  dplyr::select(-c(id, X33))
test %<>%
  dplyr::select(-c(id, X33))
```

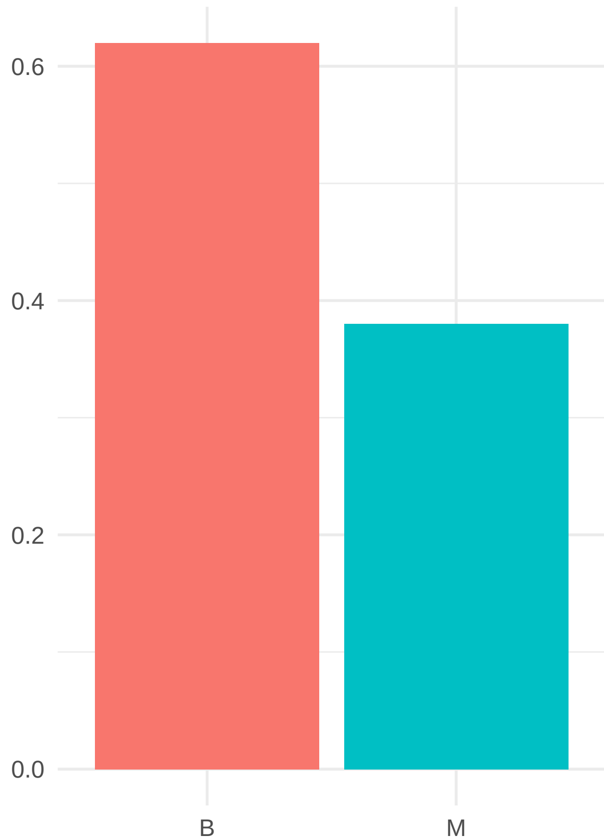
```
sum(is.na(data))
```

```
## [1] 0
```

```
training_data <- train[2:dim(train)[2]]
training_classes <- train[1]

test_data <- test[2:dim(test)[2]]
test_classes <- test[1]
```


area_w
perim_w
radius_w
perim_m
area_m
area_se
radius_se
perim_se
texture_se
texture_m
texture_w
concav_pt_se
frac_dim_se
compact_se
concav_se
smooth_se
symmetry_se
smooth_m
smooth_w
frac_dim_m
frac_dim_w
compact_w
concav_w
compact_m
concav_m
concav_pt_m
concav_pt_w
symmetry_m
symmetry_w
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1


area_m
area_se
area_w
compact_m
compact_se
compact_w
concav_m
concav_pt_m
concav_pt_se
concav_pt_w
concav_se
concav_w
frac_dim_m
frac_dim_se
frac_dim_w
perim_m
perim_se
perim_w
radius_m
radius_se
radius_w
smooth_m
smooth_se
smooth_w
symmetry_m
symmetry_se
symmetry_w
texture_m
texture_se
texture_w
diagnosis B M




```
confusion_plot <- function(actual,predicted){
  confusion_matrix <- as.data.frame(table(actual,predicted))
  g <-ggplot(confusion_matrix,aes(x=actual,y=predicted))+
    geom_tile(aes(fill=Freq))+
    geom_text(aes(label=sprintf("%1.0f", Freq)),color="white",fontface="bold")+
   labs(x="Actual class",y="Predicted class")+
    theme_minimal()
  return(g)
}
```

Dimensionality Reduction and Feature Selection

PCA

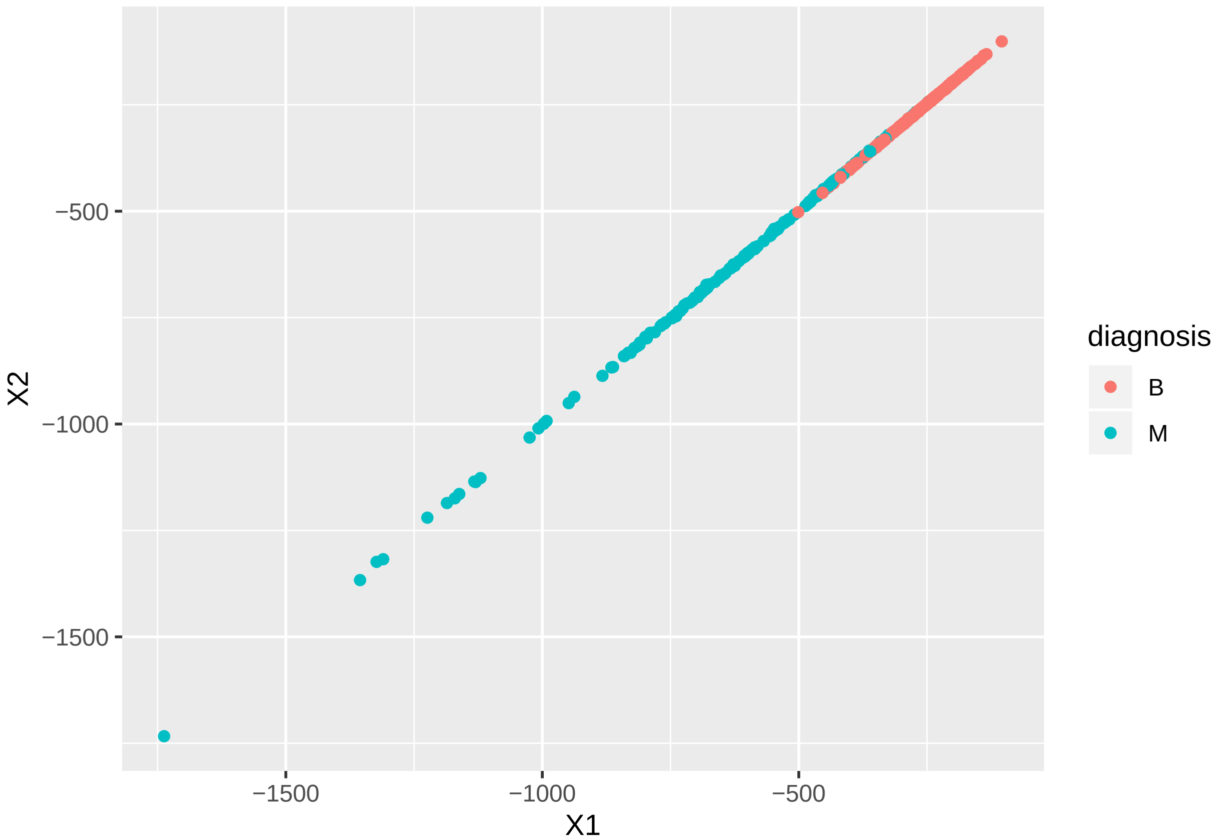
Code

```
normalise_z <-  function(X){
  mean_cols <- colMeans(X)
  sd_cols <- apply(X, 2, sd)
  mean_normalised_X <- t(apply(X, 1, function(x){x - mean_cols}))
  normalised_X <- t(apply(mean_normalised_X, 1, function(x){x / sd_cols}))
  return(normalised_X)
}


pca <- function(X, number_components_keep) {
  normalised_X <- normalise_z(X)
```

```
  corr_mat <- t(normalised_X) %*% normalised_X

  eigenvectors <- eigen(corr_mat, symmetric=TRUE)$vectors

  reduced_data <-  X %*% eigenvectors[,1:number_components_keep]
  relevant_eigs <-  eigenvectors[,1:number_components_keep]
  returnds <- list(reduced_data, relevant_eigs)
  names(returnds) <- c("reduced_data", "reduction_matrix")
  return(returnds)
}
```

Apply to dataset

```
pca_result <- pca(as.matrix(training_data), 2)
pca_reduced_training_data <- data.frame(cbind(pca_result$reduced_data, training_classes))

ggplot(data=pca_reduced_training_data, aes(x=X1, y=X2)) + geom_point(aes(colour=diagnosis))
```



tSNE

TODO: try different perplexity parameters

{r} #reduced_training_data <- tsne::tsne(training_data) # #reduced_train <- data.frame(cbind(reduced_training_data, training_classes)) #ggplot(data=reduced_training_data, aes(x=X1, y=X2)) + geom_point(aes(co

- Correlation Feature Selection
- LDA

Classification

To solve the problem of finding a SVM like classifier for non-separable data we must permit a certain number of points to violate the boundaries set however this number and the amount they violate the constraints by must be as small as possible. To formulate this we introduce a variable ϵ_i for each data point into the objective functions and the constraints leading to the optimisation problem:

$$
\min_{w,\epsilon_i} \frac{1}{2} w^T w + C \sum_{i=0}^{n} \epsilon_i
$$
$$
\text{such that } w \cdot x_i + b + \epsilon_i > 1 \text{ if } y_i = 1
$$
$$
\text{and } w \cdot x_i + b + \epsilon_i < -1 \text{ if } y_i = -1
$$

Note that we have swapped the sign of the b term in the equation for the hyperplane because I implemented it this way before realising they were different and am lazy.

As the above problem is convex (as it is quadratic) and Slater's condition holds then strong duality holds and we can take the Lagrangian of the optimisation problem and consider the result of the KKT conditions. By doing so we can reformulate the optimisation problem as the dual problem:

$$
\min_{\lambda} \frac{\overline{\lambda} X X^T \overline{\lambda^T}}{4} + \lambda^T \mathbf{1}
$$
$$
\text{such that } 0 \leq \lambda_i \leq C
$$
$$
\text{and } \sum_{i}^{n} \lambda_i y_i = 0
$$

where

$$
X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \text{ and } \overline{\lambda} = [\lambda_1 \cdot y_i, ..., \lambda_n \cdot y_n] \text{ and } \mathbf{1} = [1, ..., 1] \in \mathbb{R}^{\ltimes}
$$

As before we have to massage this optimisation problem into one that can be solved using `solve.QP`. In this formulation

$$
d = \mathbf{1}
$$

and

$$D = \begin{pmatrix} y_1 & 0 & ... & 0 \\ 0 & y_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & y_n \end{pmatrix} XX^T \begin{pmatrix} y_1 & 0 & ... & 0 \\ 0 & y_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & y_n \end{pmatrix}$$

A and b_0 require slightly more manipulation this time around with

$$A = \begin{pmatrix} y_1 & y_2 & ... & y_n \\ & & I & \\ & & -I & \end{pmatrix}$$

and

$$b_0 = \begin{pmatrix} 0 \\ \mathbf{0} \\ -\boldsymbol{C} \end{pmatrix}$$

where

$$\mathbf{0} = [0, ..., 0]^T \in \mathbb{R}^n$$

and

$$\boldsymbol{C} = [C, ..., C]^T \in \mathbb{R}^n$$

The code for this applied to the non-separable data can be found below.

```
C <- 1

X <- as.matrix(combined_class)[,1:2]
y <- as.matrix(combined_class)[,3]
Dmat2 <- diag(y) * X %*% t(X) %*% diag(y)
diag(Dmat2) <- diag(Dmat2) + 1e-11
dv2 <- rep(1, 30)

A2 <- rbind( y,diag(30))
A2 <- rbind(A2, -1*diag(30))

bv2 <- c(c(0), rep(0, 30), rep(-C, 30) )
model <- solve.QP(Dmat2, dv2, t(A2), bv2, meq = 1)
```

In order to recover w and b from λ we use the relationship

$$w = \sum_{i=0}^{n-1} \lambda_i x_i^T y_i$$

and

$$b = \text{mean}(\sum_{i=0}^{k} y_i - w \cdot x_i) \,.\, \forall\, i.0 < \lambda_i < C$$

Which can be made as functions in R as so:

```
calculate_b <- function(w, X, y, a, C) {
  ks <- sapply(a, function(x){return(x > 0 && x < C)})
  indices <- which(ks)
  sum_bs <- 0
  for(i in indices) {
    sum_bs <- sum_bs + (y[i] - w %*% X[i,])
  }
  return(sum_bs / length(indices))
```

```
}


recover_w <- function(a, y, X){
  colSums(diag(a) %*% diag(y) %*% X)
}
```

We can see the results of using the dual regression below

SVM

```
soft_margin_svm_plotter <- function(w, b) {
  plotter <- function(x) {
    return(1/w[2]   * -(b + (w[1]*as.numeric(x))))
  }
  return(plotter)
}
```

```
factor_to_label <- function(x)  {
  if(as.character(x) == "M")  {
    return(1)
  }
  else {
    return(-1)
  }
}
```

```
label_to_factor <- function(x) {
  if(x == 1) {
    return(as.factor("M"))
  }
  else{
    return(as.factor("B"))
  }
}
numeric_test_labels <- apply(test_classes, 1, factor_to_label)
numeric_training_labels <- apply(training_classes, 1, factor_to_label)
```

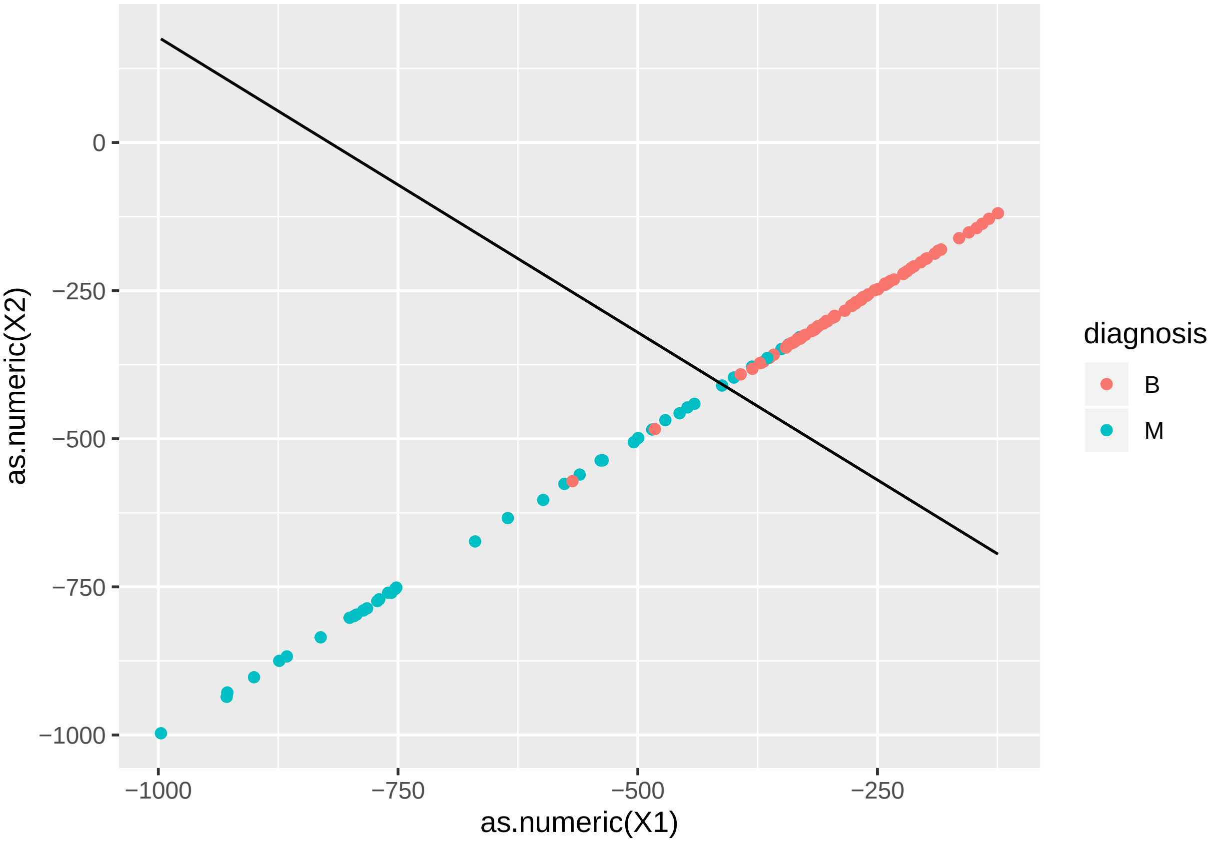
Use PCA then do SVM

```
model <- svm(X=pca_result$reduced_data,
             classes=numeric_training_labels,
             C=100000, margin_type='soft',
             kernel_function = linear_kernel,
             feature_map = linear_basis_function)


reduced_prediction_fn <- model$prediction_function

pca_reduced_prediction_fn <- function(x) {
  p <- x %*% pca_result$reduction_matrix
  reduced_prediction_fn(t(p))
}
```

```
predictions_svm <- apply(as.matrix(test_data),1, pca_reduced_prediction_fn)
accuracy_calc(numeric_test_labels, predictions_svm)
```

```
## [1] 92.98246
```

```
svm_plotter <- soft_margin_svm_plotter(model$params$w, model$params$b)
embedded_test_data <- data.frame(cbind(as.matrix(test_data) %*% pca_result$reduction_matrix), test_clas

ggplot(embedded_test_data, aes(x=as.numeric(X1), y=as.numeric(X2))) +
  geom_point(aes(colour=diagnosis)) +
  stat_function(fun=svm_plotter)
```



Naive Bayes

Mathematical setting

Let y be the class label that we want to assign to an observation $\boldsymbol{x} = (x_1, \cdots x_d)$, where $x_1, \cdots x_d$ are the features. The probability of an observation having label y is given by Bayes rule,

$$
\begin{aligned}
P(y|x_1, \cdots, x_d) &= \frac{P(x_1, \cdots, x_d|y_k)P(y)}{P(x_1, \cdots, x_d)} \\
&\propto P(x_1, \cdots, x_d|y_k)P(y).
\end{aligned}
$$

The prior class probability $P(y)$ can be easily obtained by the proportion of observation that are in the given class.

The main assumption is that every feature is conditionally independent given the class label y. The reason why this classifier is called *naive* is that very often this assumption is not actually realistic.

This assumption simplifies the posterior to

$$P(y|x_1, \cdots, x_d) \propto P(y) \prod_{i=1}^{d} P(x_i|y).$$

There are various types of Naive Bayes classifiers based on the type of features. In our case, since we have continuous variables we assume that all features are normally distributed. Therefore, the conditional probabilities can be calculated as

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Finally, to assign the class to an observation we use the Maximum A Posteriori decision rule. For every observation, we pick the class the has the highest probability

$$y = \underset{y}{\text{argmax}} P(y) \prod_{i=1}^{d} P(x_i|y).$$

Implementation

Here are some code snippets just to illustrate how these theoretical aspects are implemented. The full code can be found in the package.

The observations are stored as rows in X and the corresponding class labels are entires in the column matrix y.

First we calculate the prior class probabilities based on the number of observations in each class.

```
n <- dim(X)[1]
d <- dim(X)[2]
classes <- sort(unique(y)[, 1])
k <- length(classes)

prior <- rep(0, k)
for (i in 1:k) {
  prior[i] <- sum(y == classes[i]) / n
}
```

Then we create an array of the mean and sd of the data split by clasess and features.

```
summaries <- array(rep(1, d * k * 2), dim = c(k, d, 2))
for (i in 1:k) {
  X_k <- X[which(y == (i - 1)), ]
  summaries[i, , 1] <- apply(X_k, 2, mean)
  summaries[i, , 2] <- apply(X_k, 2, sd)
}
```

Finally, the predictions are obtained by taking the largest posterior class probability. Note that in order to avoid underflow, we take the maximum of the *log* posterior class probabilities.

```
probs <- matrix(rep(0, n * k), nrow = n)
for (obs in 1:n) {
  for (class in 1:k) {
    class_prob <- log(prior[class])
```

```r
    for (feat in 1:d) {
      mu <- summaries[class, feat, 1]
      sd <- summaries[class, feat, 2]
      cond <- dnorm(x_new[obs, feat], mu, sd, log = TRUE)
      class_prob <- class_prob + cond
      }
     probs[obs, class] <- class_prob
  }
}

pred <- apply(probs, 1, which.max)
```

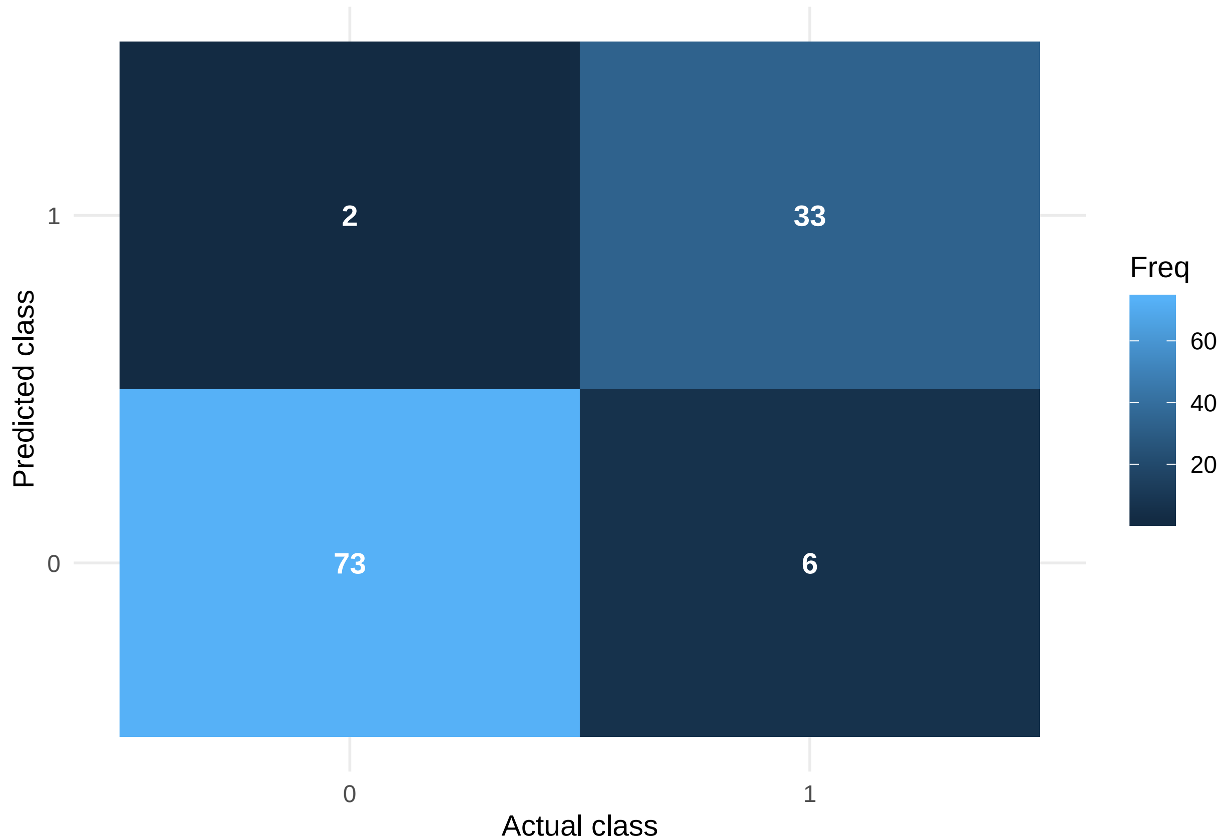
Fit model to dataset

```r
install_github("andreabecsek/NaiveBayes")
library(NaiveBayes)
```

```r
levels(training_classes$diagnosis) <- c(0,1)
training_classes %<>% as.matrix
mode(training_classes) <- 'numeric'

levels(test_classes$diagnosis) <-c(0,1)
test_classes %<>% as.matrix
mode(test_classes) <- 'numeric'
```

Fit the Naive Bayes model to the data, calculate predictions and check the accuracy using.

```r
model_naive <- naive_bayes(training_data,training_classes)

predictions_naive <- predict(model_naive,as.matrix(test_data))

confusion_plot(test_classes,predictions_naive)
```



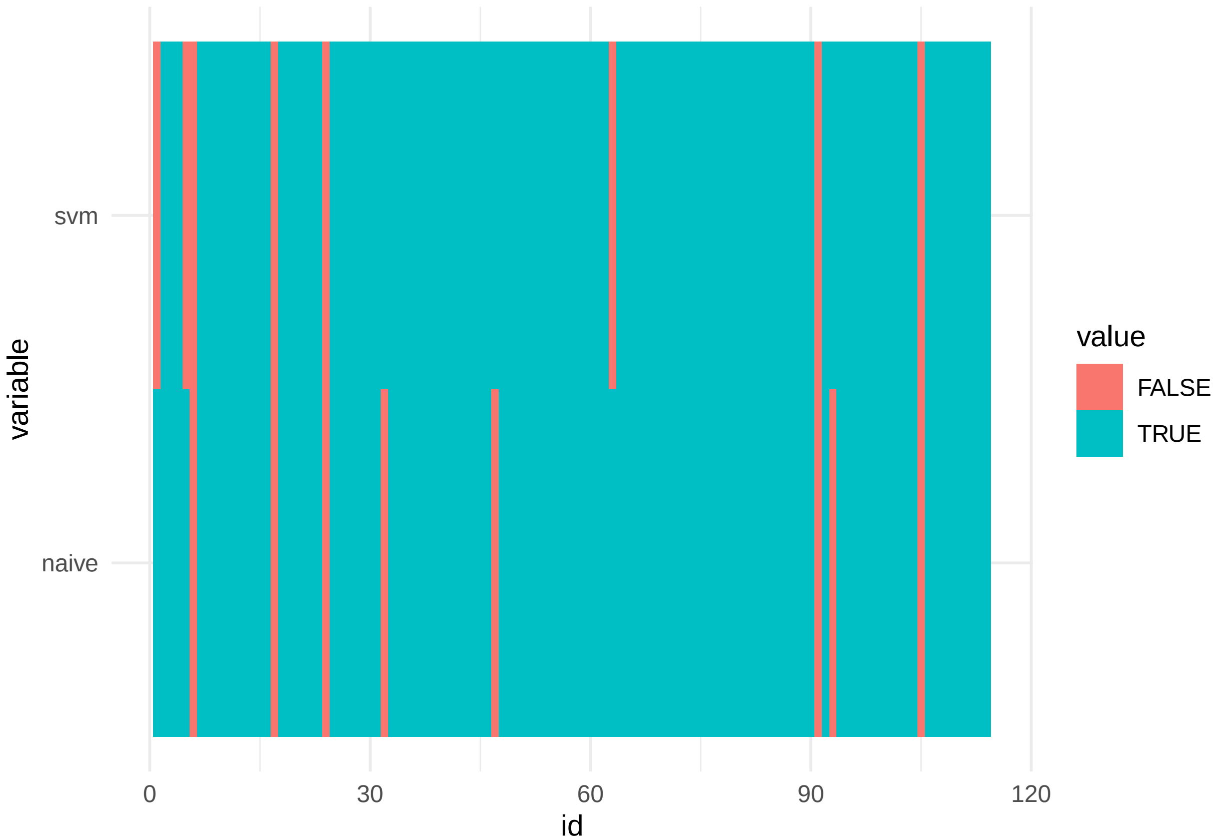
Conclusion

Merge all predictions

```
id <- seq(length(id_test))
all_predictions <- cbind(id,numeric_test_labels,predictions_naive,predictions_svm)
colnames(all_predictions) <- c('id','actual','naive','svm')
all_predictions[all_predictions==-1] <-0
all_predictions %<>% as.data.frame()
```

```
errors <- all_predictions %>%
  mutate(naive=naive==actual) %>%
  mutate(svm=svm==actual) %>%
  dplyr::select(-actual)
```

```
a <- errors %>%
  melt(id='id')

ggplot(a,aes(x=id,y=variable,fill=value))+
  geom_raster()+
  theme_minimal()
```



TODO: analyse results

- Naive Bayes
- Logistic Regression

Conclusion

- Evaluation of results
- Discuss outliers

TODO: create outlier plot