# Assignment 1. Austin Frownfelter

# Problem 2.

## (a)

The range for each attribute (in order, excluding class attribute)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Max | 17 | 199 | 122 | 99 | 846 | 67.1 | 2.42 | 81 |
| Min | 0 | 0 | 0 | 0 | 0 | 0 | 0.0780 | 21 |

## (b)

Means and Variances for each attribute (in order, excluding class attribute)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mean | 3.8451 | 120.8945 | 69.1055 | 20.5365 | 79.7995 | 31.9926 | 0.4719 | 33.2409 |
| Variance | 11 | 1021 | 374 | 254 | 13264 | 62 | 0 | 138 |

## (C)

Correlation with Target Class attribute

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| .2219 | .4666 | .0651 | .0748 | .1305 | .2927 | .1738 | .2384 |

The 2nd attribute has the highest correlation with the target class attribute. Given the target class is whether the patient shows signs of diabetes, it makes sense that the attribute measuring the glucose concentration would have the highest correlation. With my limited understanding of diabetes, blood-glucose levels are related to or affect diabetes. Some of the other attributes don’t really make sense to be correlated with diabetes, such as skin fold thickness, and they happen to have low correlation.

## (d)

The maximum correlation between any two attributes is 0.5443. The two attributes are 1 and 8, or the Number of times pregnant and the patient’s age.

## (E)

Having two attributes that are fully correlated may not necessarily affect target class prediction. If two attributes are very closely correlated (close to 100%), then they would be redundant in the classification since they are effectively equivalent.

## (F)

The most normal figure comes down to figure 6 and figure 3 (both shown below). In both cases there are outliers at the zero-bins, which seem to be data points that lack a measurement (since you cannot have zero blood pressure or BMI without being dead.) Overall, after omitting these outliers, Figure 3 (Diastolic Blood Pressure) appears to be the most normally distributed attribute.

|  |  |
| --- | --- |
| C:\Users\frown\AppData\Local\Microsoft\Windows\INetCache\Content.Word\partFfigure3.png | C:\Users\frown\AppData\Local\Microsoft\Windows\INetCache\Content.Word\partFfigure6.png |
| Figure 3 | Figure 6 |

## (G)

There appears to be a few scatter plots that could have some linear dependency. These linear relationships are not highly correlated, so the linear relationships would be weak. Attributes 4 and 6, or Triceps skin fold thickness are Body mass index, have the strongest linear relationship. If the zero measurements for Triceps skin fold thickness are removed (since they are likely not valid measurements), then there would be a general, loose linear correlation between these two attributes. Figure 1 and 7, or the Number of times pregnant and Body mass index, appear to not be correlated in any way.

|  |  |
| --- | --- |
| C:\Users\frown\AppData\Local\Microsoft\Windows\INetCache\Content.Word\partGfigure6v4.png | C:\Users\frown\AppData\Local\Microsoft\Windows\INetCache\Content.Word\partGfigure7v1.png |
| Figure 4v6 | Figure 1v7 |

# Problem 3.

## (A)

The normalized values for the first five entries of attribute 3 are listed below. The code will be uploaded on Courseweb.

0.1495

-0.1604

-0.2638

-0.1604

-1.5037

## (B)

The discretized values of the first five entries of attribute 3 are listed below.

5

5

5

5

3

# Problem 4.

## (A)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mean | 3.298 | 109.98 | 68.184 | 19.664 | 68.792 | 30.3042 | 0.4297 | 31.19 | 0 |
| StdDev | 3.0172 | 26.1412 | 18.0631 | 14.8899 | 98.8653 | 7.6899 | 0.2991 | 11.6677 | 0 |

Class-0 subset

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mean | 4.8657 | 141.2575 | 70.8246 | 22.1642 | 100.3358 | 35.1425 | 0.5505 | 37.0672 | 1 |
| StdDev | 3.7412 | 31.9396 | 21.4918 | 17.6797 | 138.6891 | 7.263 | 0.3724 | 10.9683 | 0 |

Class-1 subset

## (B)

Average length for the training dataset in one case of 20 runs was 504.75, which comes out to 65.72%.

## (C)

Average length for the training dataset is always 507, which is 66.02%, as close as possible for this data set. The data is different in every case.