# ML and PR Project Report – Part 1

What follows is a brief analysis of the statistics and distribution of the data relative to a fingerprint database, for which labels indicating whether a fingerprint is real or counterfeit are associated. The original dataset is trimmed to just 6 features for the sake of simplicity. **­­­**

**Features 0 and 1**

On the histogram, for both feature 0 and 1, the classes overlap considerably. We can infer visually that the mean is pretty much the same, while the variances, although not that much different from one another, show a greater value for the True class in the case of attribute-0, and a greater value for the False class in the case of attribute-1. These results are backed up by the actual calculation of the statistics:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| attribute-0 | Mean | Variance | attribute-1 | Mean | Variance |
| 0 – False | 0.00287744 | 0.56958105 | 0 – False | 0.01869316 | 1.42086571 |
| 1 – True | 0.00054454 | 1.43023345 | 1 – True | -0.00852437 | 0.57827792 |

Attribute 0 has a single mode for the False class and a couple for the True class, although adjacent. Attribute 1 has a single mode for the True class, and again a single mode for the False class, although the latter, the bucket that’s second in line has almost the same frequency as the mode, but again they are adjacent.

Immagine che contiene schermata, diagramma, testo, Diagramma

Descrizione generata automaticamenteImmagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

**Features 2 and 3**

For features 2 and 3, the classes do overlap, in the origin. The overlap covers a considerable amount of both attributes’ domain (more than 2/3rds of it) and some of their image (probably about 1/3rd of it). The means are clearly separated this time, and “mirrored” by the origin, meaning they are actually equal in magnitude, while the variances do not seem to differ much. Again, the statistics back-up the visual observations:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| attribute-2 | Mean | Variance | attribute-3 | Mean | Variance |
| 0 – False | -0.68094016 | 0.54997702 | 0 – False | 0.6708362 | 0.53604266 |
| 1 – True | 0.66523784 | 0.5489026 | 1 – True | -0.66419539 | 0.55334275 |

Both classes for both features show only one mode. If we consider the second most frequent bucket we may see 2 modes in most cases, but they are adjacent.

Immagine che contiene testo, diagramma, schermata, Diagramma

Descrizione generata automaticamente Immagine che contiene testo, schermata, diagramma, Carattere

Descrizione generata automaticamente

**Features 4 and 5**

For these last two features, classes do overlap, although in a different fashion compared to the previous ones: The true class is split into two “clusters”, centered near the edges of the False class, which is instead a single “cluster”.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| attribute-4 | Mean | Variance | attribute-5 | Mean | Variance |
| 0 – False | 0.02795697 | 0.6800736 | 0 – False | -0.0058274 | 0.70503844 |
| 1 – True | -0.04172518 | 1.31776792 | 1 – True | -0.02393848 | 1.28702609 |

As one would expect, we can then identify, for both attributes, double mode for the True class, while there’s only one for the False class.

**Immagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamente** Immagine che contiene testo, schermata, diagramma, Carattere

Descrizione generata automaticamente

Now, if we observe the scatter plots for these two features (4 and 5), we can understand the distribution of the data samples on this specific projection of features. What can be seen is that 4 clusters for each class appear, and their distribution show a specific type of correlation between the two features, which could probably be used for classification, since the clusters don’t overlap much. We can also observe that the extension of the domains for each class are pretty much the same for both features: the circular clusters are probably a consequence of this.

Immagine che contiene testo, schermata, Policromia, diagramma

Descrizione generata automaticamente

For completeness, I will add the scatter plot for the first two couples of attributes too.

Immagine che contiene testo, schermata, diagramma, Policromia

Descrizione generata automaticamenteImmagine che contiene testo, schermata, Policromia, diagramma

Descrizione generata automaticamente

**Conclusion**

To sum up, features 0 and 1 don’t give us much information about which class a sample may belong to, while the others may be useful.