To predict failure/ pass with demographic features

1. Algorithm used:
2. Logistic regression
3. Gaussian Naïve Bayes
4. SVM-linear kernel
5. SVM-rbf kernel
6. Decision tree
7. Random forest
8. Results on the test data (25% of the total dataset)

including accuracy, precision, recall, f-score, kappa-score

(por)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm / metrics | Accuracy | Precision | Recall | f-score | Kappa score |
| Logistic regression | 0.73 | 0.52 | 0.31 | 0.39 | 0.23 |
| Gaussian Naïve Bayes | 0.6748 | 0.41 | 0.42 | 0.42 | 0.19 |
| SVM-linear kernel | 0.7239 | 0 | 0 | 0 | 0 |
| Decision tree | 0.7116 | 0.46 | 0.29 | 0.36 | 0.1832 |
| Random forest | 0.7178 | 0.48 | 0.27 | 0.34 | 0.1814 |

(mat)

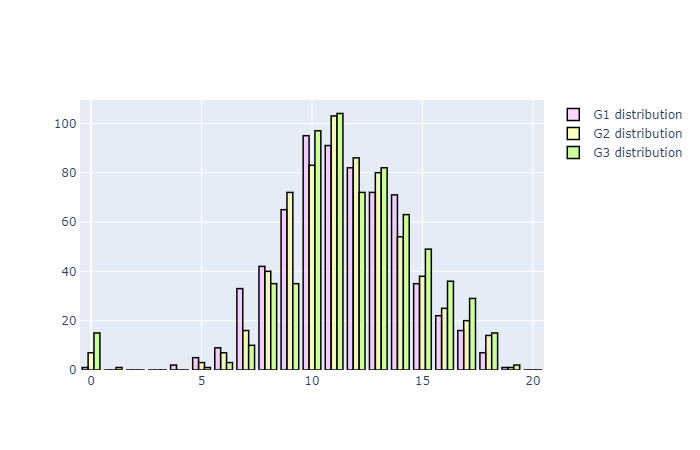
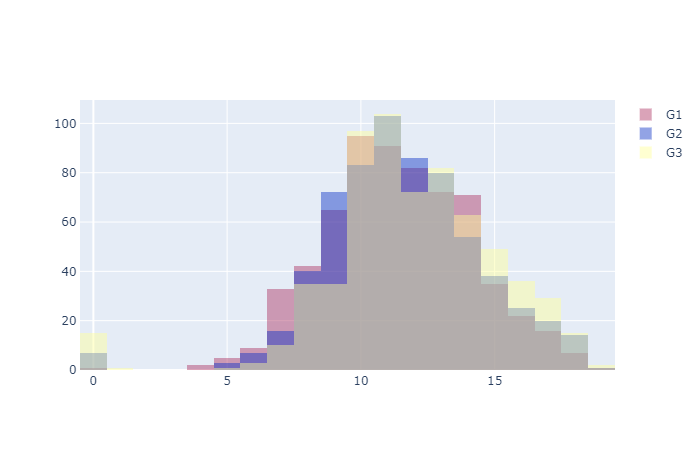
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm / metrics | Accuracy | Precision | Recall | f-score | Kappa score |
| Logistic regression | 0.62 | 0.37 | 0.62 | 0.46 | 0.2029 |
| Gaussian Naïve Bayes | 0.61 | 0.33 | 0.46 | 0.39 | 0.12 |
| SVM-linear kernel | 0.636 | 0.25 | 0.19 | 0.22 | -0.014 |
| Decision tree | 0.636 | 0.35 | 0.46 | 0.40 | 0.1457 |
| Random forest | 0.565 | 0.24 | 0.31 | 0.27 | -0.032 |

The variance of the model is quite high

the test results and the model performance varies a lot with regards to different test set

(logistic regression precision varies from 0.3 to 0.6)

1. Predicted failure possibility of the test set by each model (X\_test\_data.xsxl)
2. Feature importance ranked by each model: LR/Decision tree/ random forest (feature\_importance.xsxl)
3. Plot of G1, G2, G3 （por）



(mat)

