Gender Inequality Index Prediction by Machine Learning

- Course : Machine Learning

About GII

GII is a composite metric of gender inequality using three dimensions: reproductive health, empowerment and the labour market. A low GII value indicates low inequality between women and men, and vice-versa.

It shows the loss in potential human development due to inequality between female and male achievements in these dimensions. It ranges from 0, where women and men fare equally, to 1, where one gender fares as poorly as possible in all measured dimensions.

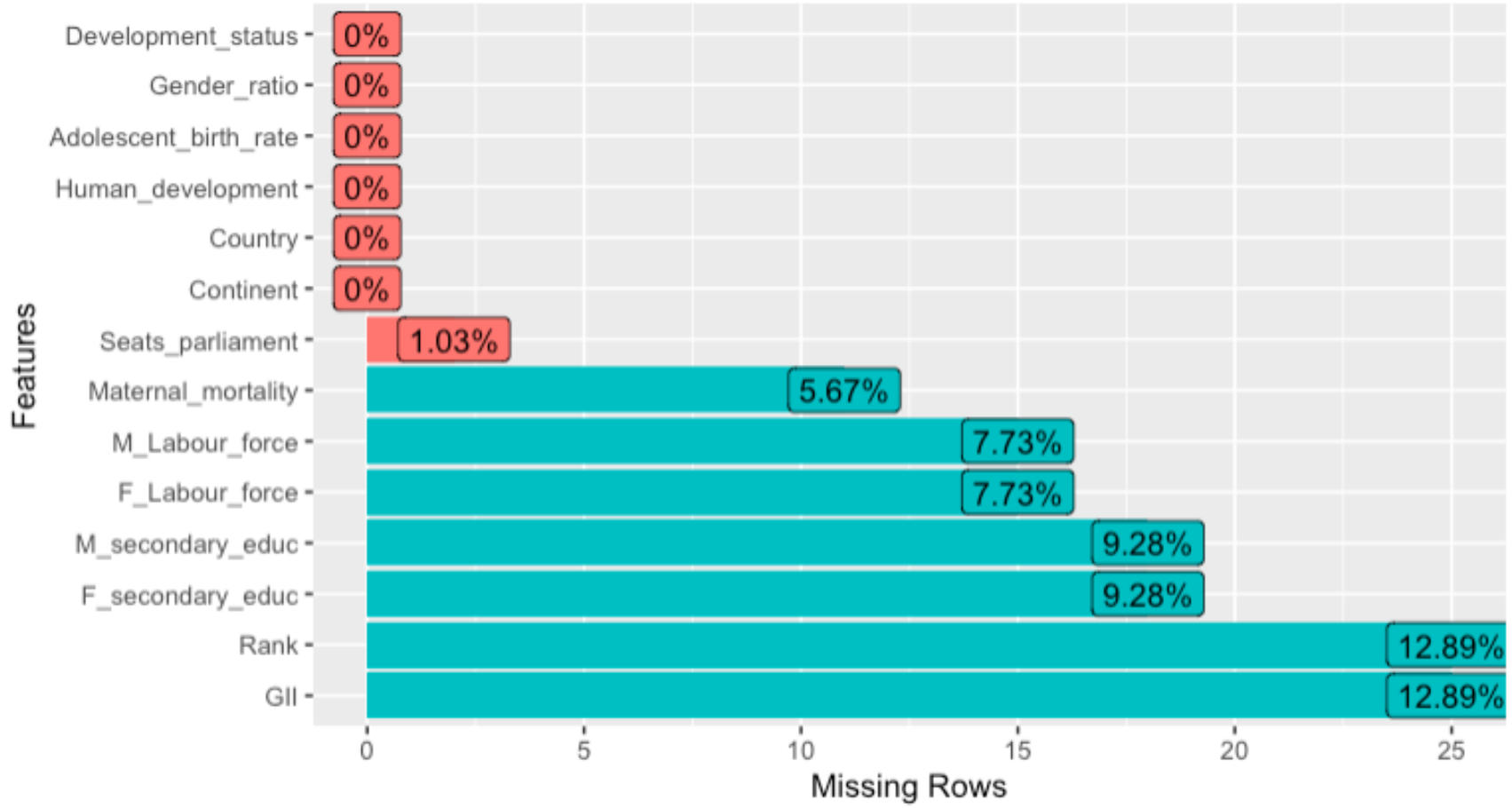
Research Question

- Developing a predictive model for the Gender Inequality Index (GII) by applying machine learning techniques, while considering multiple variables.
- Which machine learning models are most effective for predicting gender inequality, and how do their performance metrics compare?
- Which countries have the highest and lowest levels of gender inequality, and what factors contribute to these disparities? Is there a correlation between a country's level of human development and its level of gender inequality?

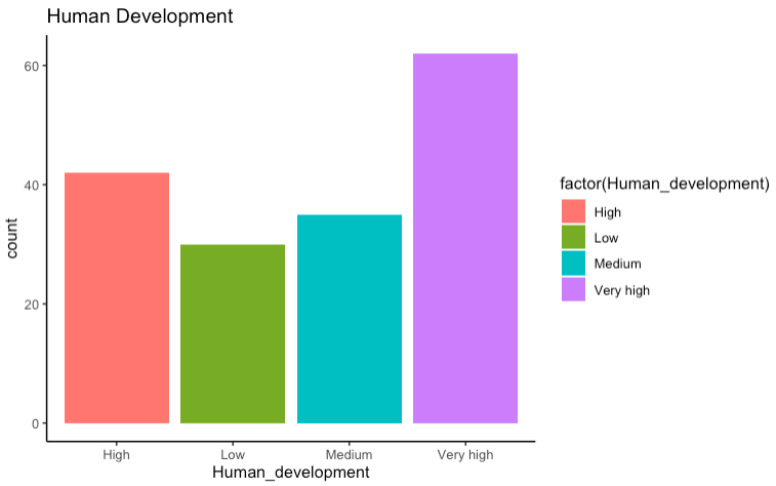
EDA

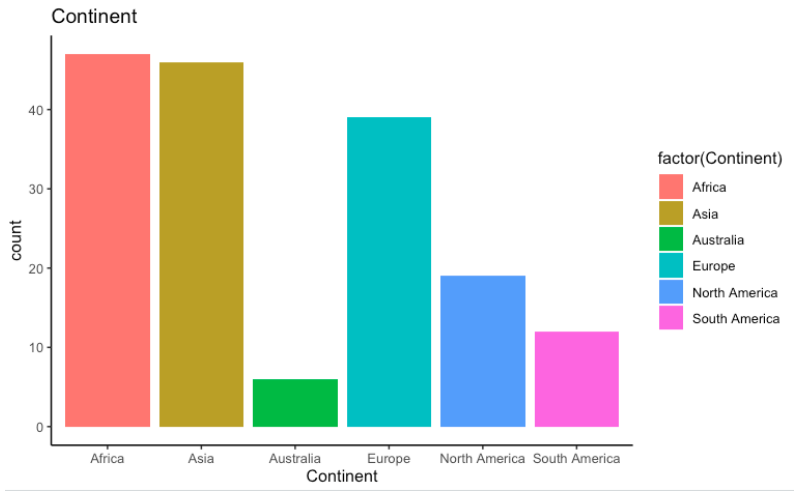
- Dataset : 194 rows and 14 columns
- Response variable
  - GII:Gender Inequality Index (response variable).
  - Low value = good equality, high value = high inequality
- 13 features
  - Human development category: Low- medium-high-Very High
  - Rank: Country Rank (highly correlated with GII)
  - Maternal mortality ratio (deaths per 100,000 live births)
  - Adolescent birth rate (births per 1,000 women ages 15–19)
  - Seats_parliament: Share of seats in parliament (% held by women)
  - F_secondary_educ: Females with at least some secondary education (% ages 25 and older)
  - M_secondary_educ: Males with at least some secondary education (% ages 25 and older)
  - F_Labour_force: Female - Labour force participation rate (% ages 15 and older)
  - M_Labour_force: Male - Labour force participation rate (% ages 15 and older)
  - Continent : Asia, Africa, North America, South America, Europe, and Australia
  - Country :190 Countries
  - Gender_ratio
  - Development_status : Developed or developing countries

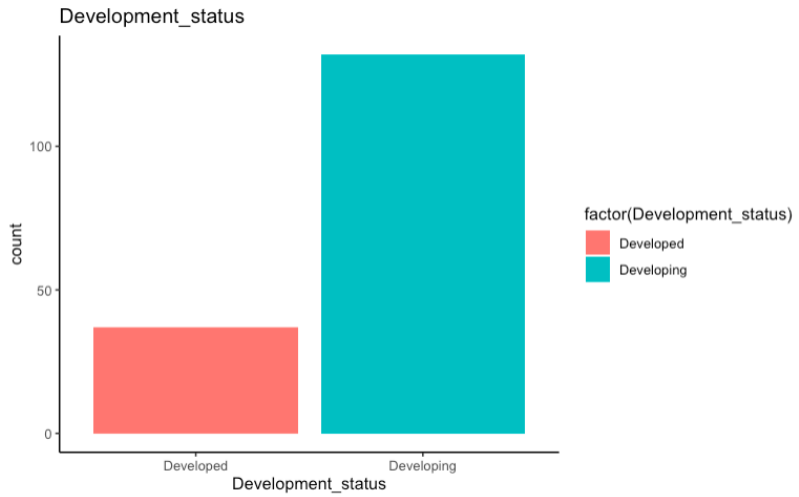
- Missing data



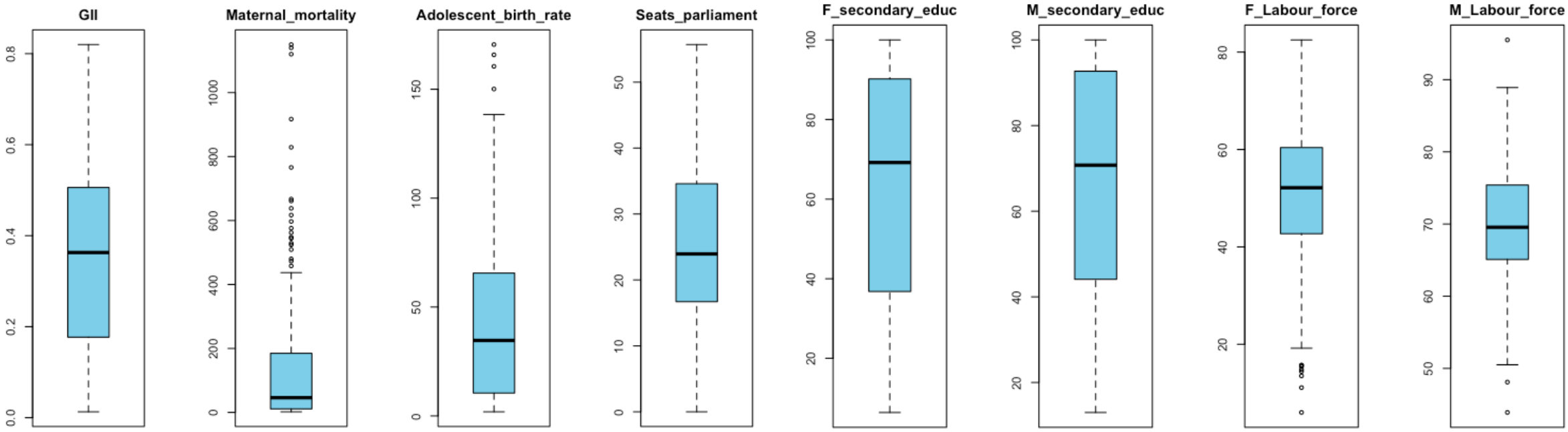
- Character Variables





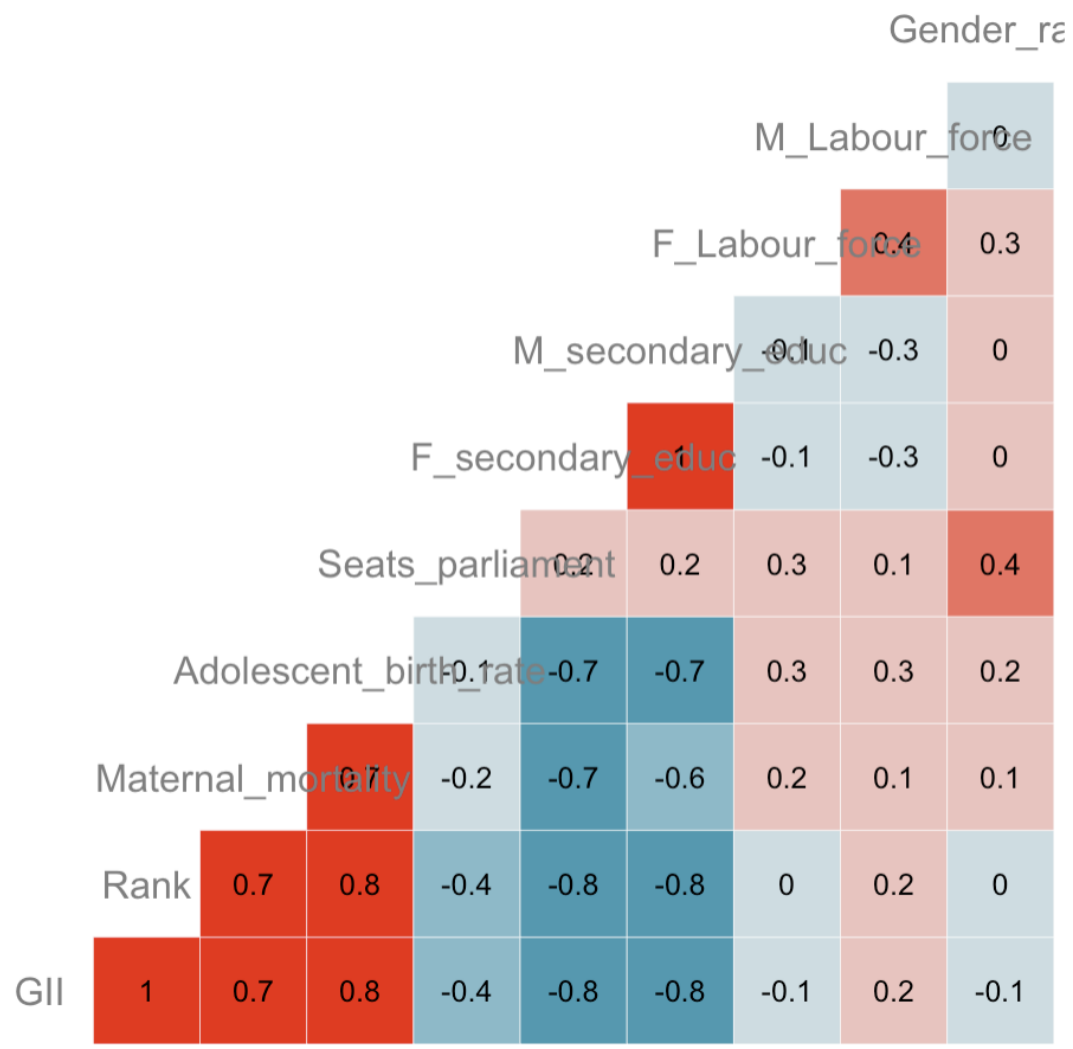


- Numerical Variables

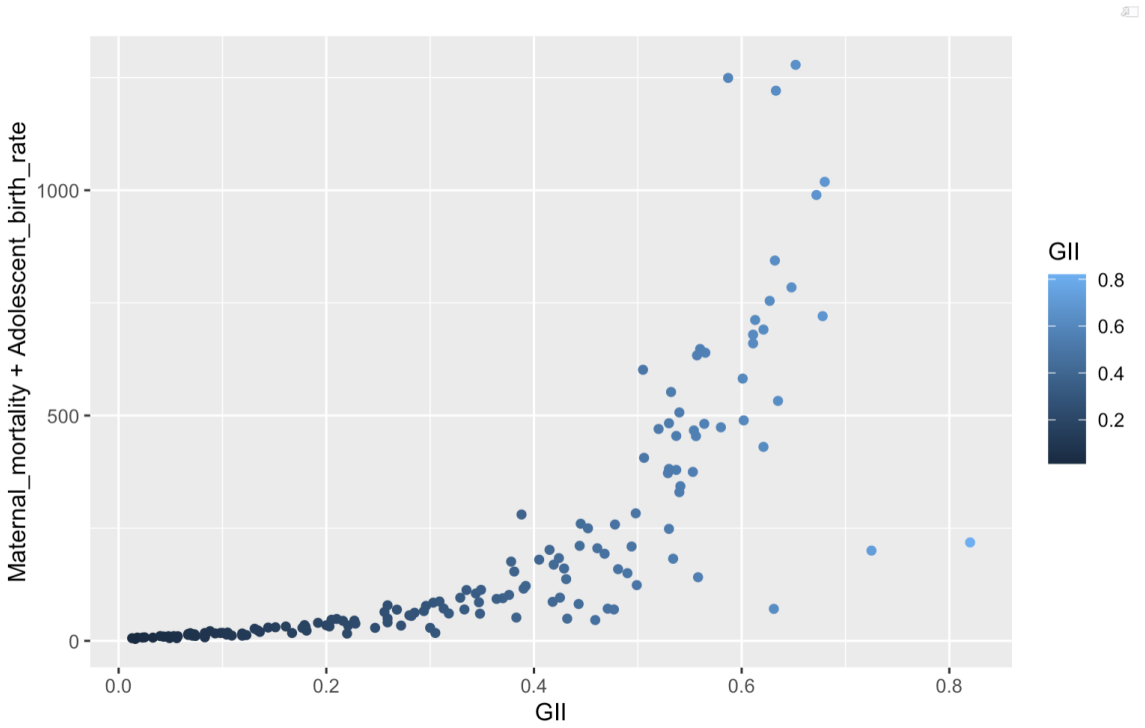


	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	mean <dbl>
1	GII	0	1	0.342787
2	Maternal_mortality	0	1	154.467456
3	Adolescent_birth_rate	0	1	44.488757
4	Seats_parliament	0	1	25.360947
5	F_secondary_educ	0	1	62.302959
6	M_secondary_educ	0	1	66.749704
7	F_Labour_force	0	1	50.378698
8	M_Labour_force	0	1	70.054438

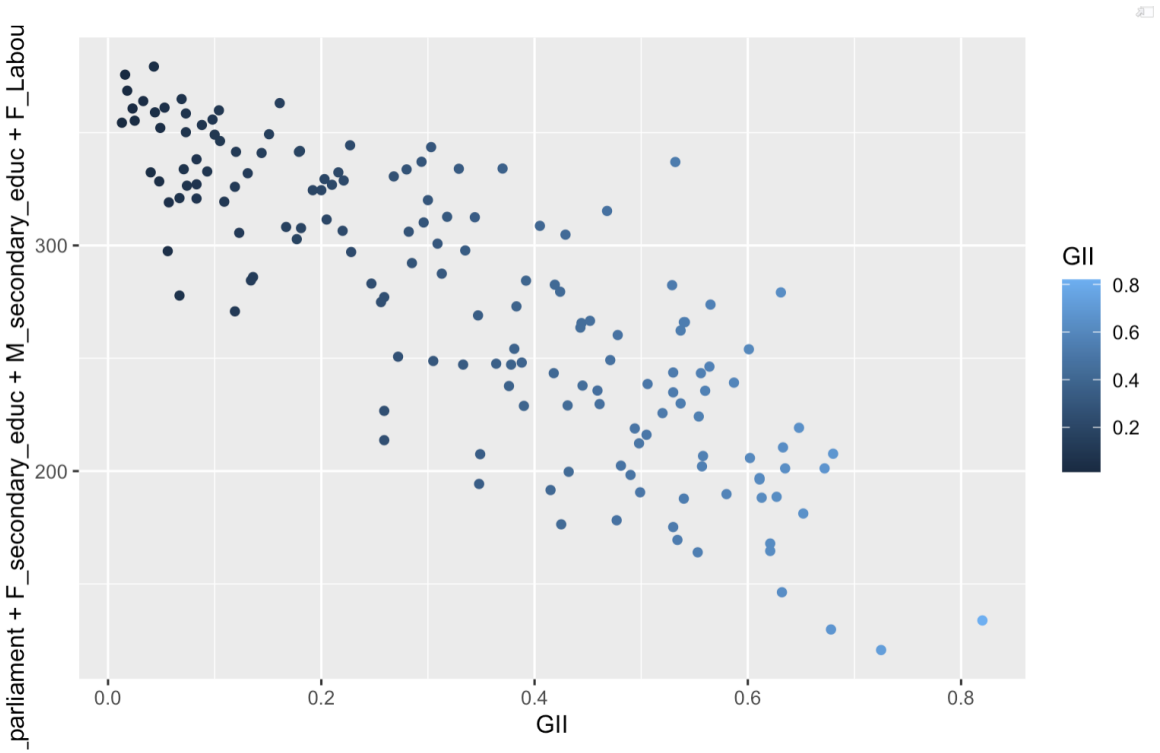
- Correlation
  - Strong relationship with GII
    :Rank(removed)
    Maternal_mortality,
    Adolescent_birth_rate,
    F_secondary_education,
    M_secondary_education,
    Development_status



- Scatterplot

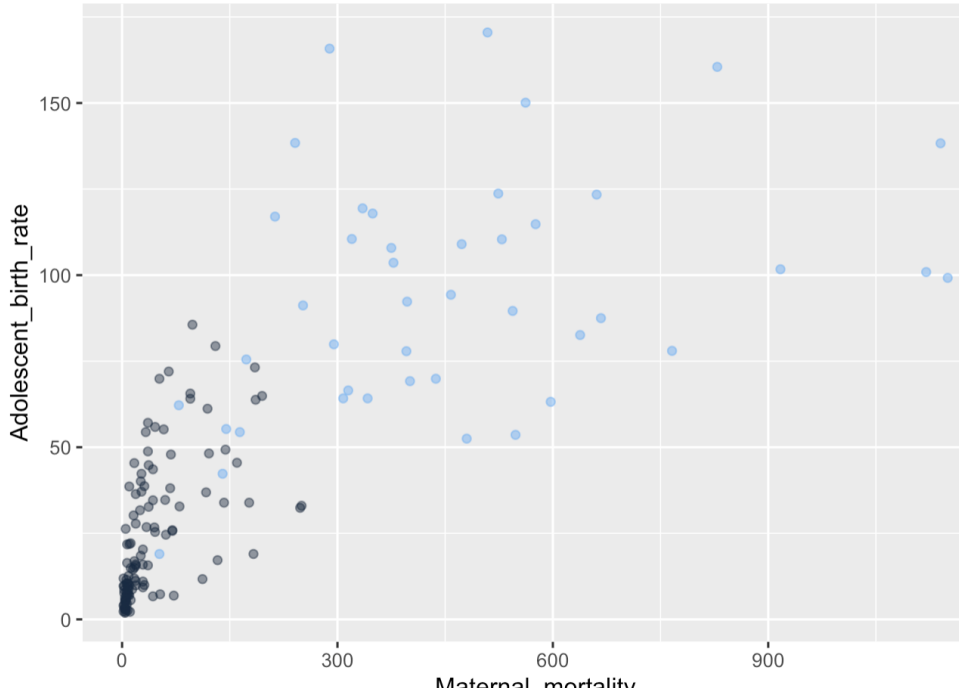


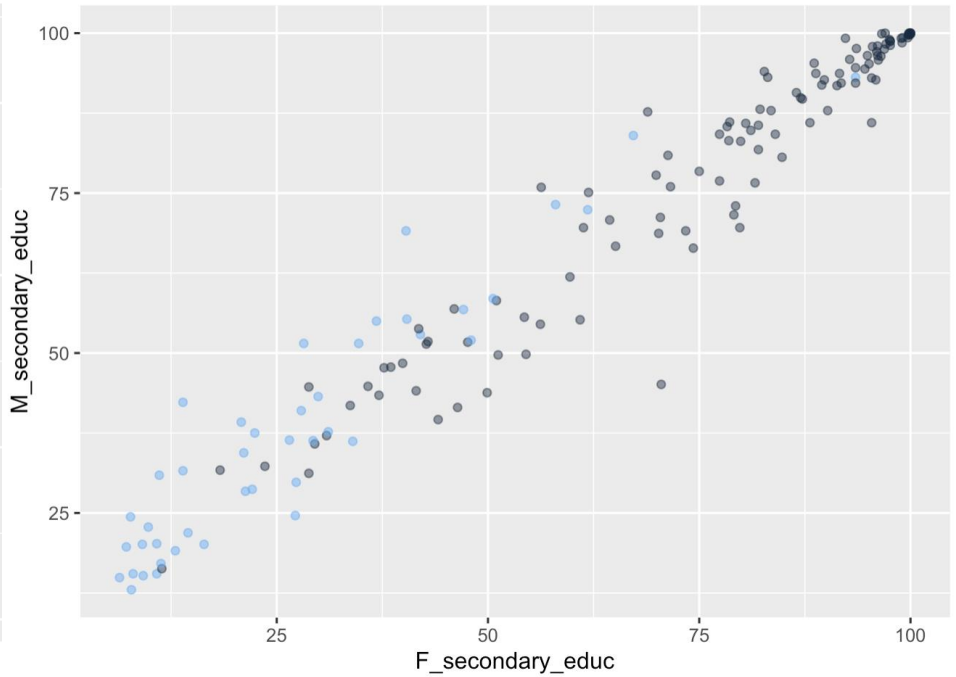
Proportional relationship with GII

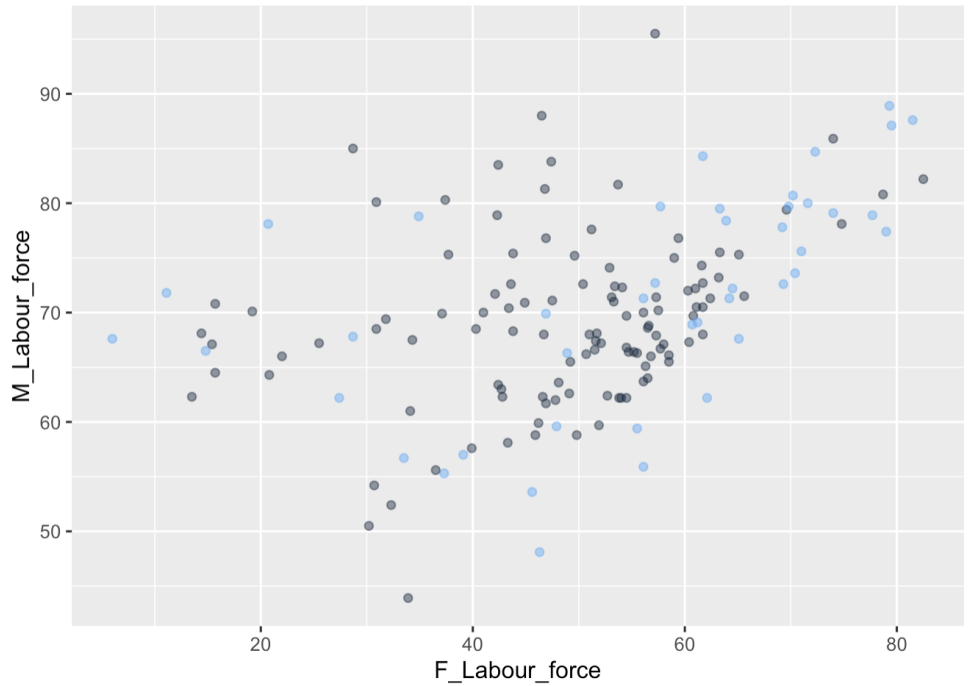


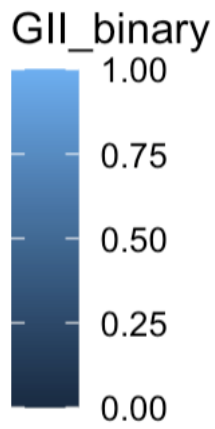
Inverse relationship with GII

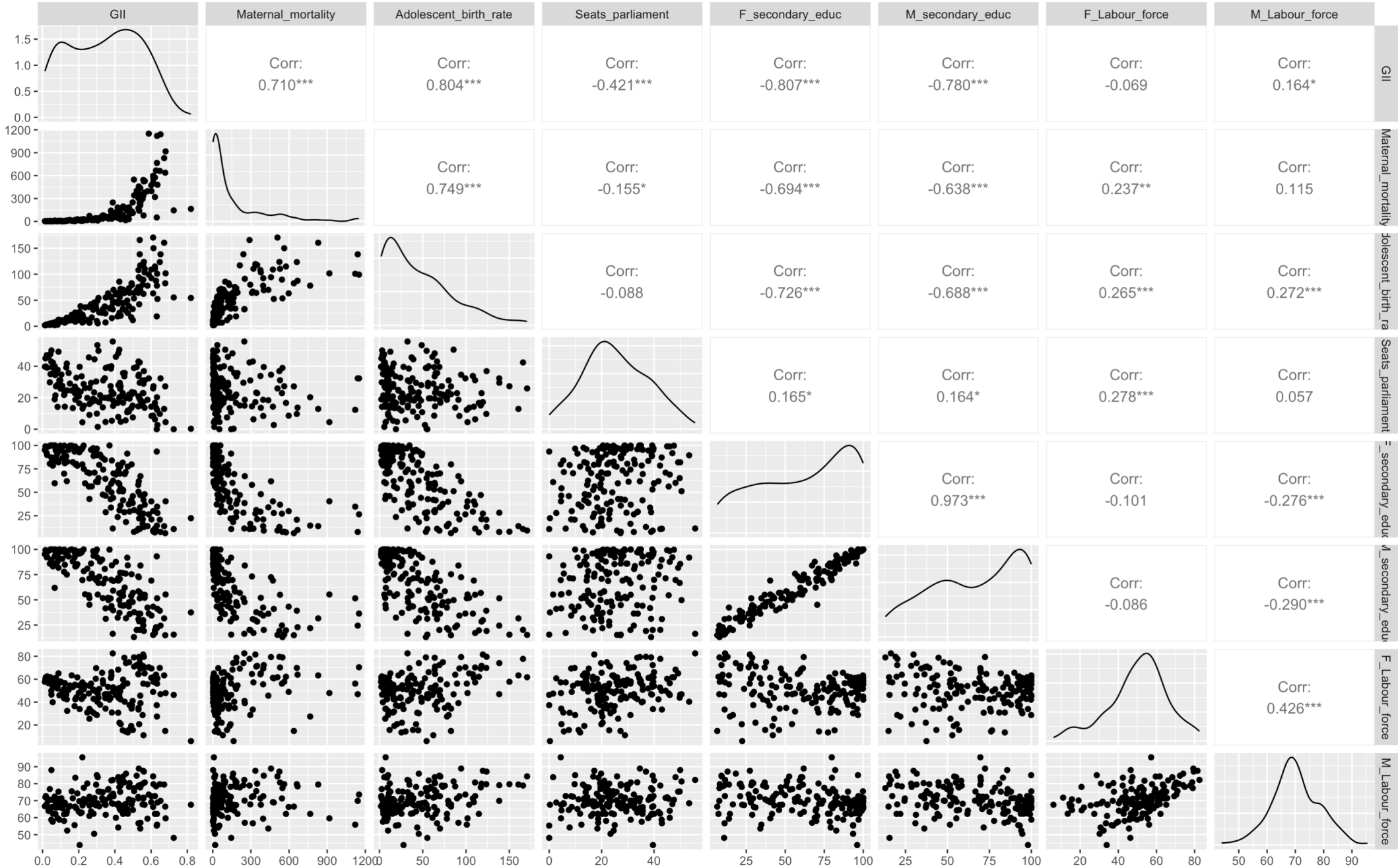
- Check if GII is separable or not



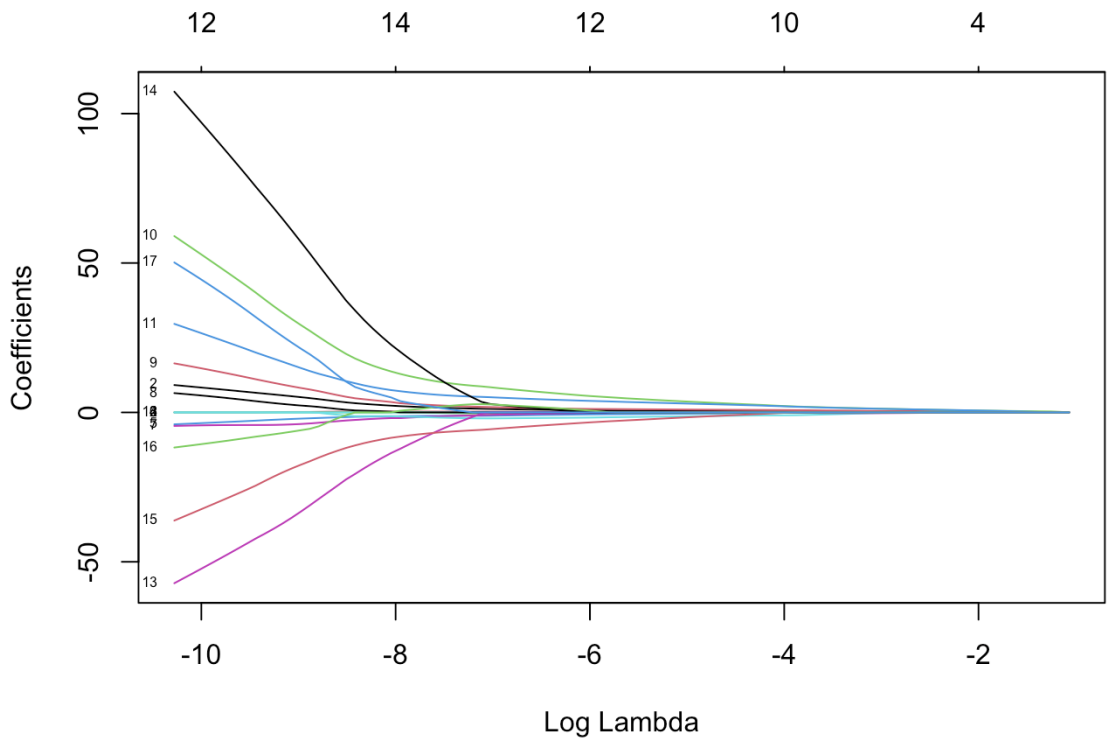







GII
Maternal_mortality
Adolescent_birth_rate
Seats_parliament
F_secondary_educ
M_secondary_educ
F_Labour_force
M_Labour_force
Corr: 0.710***
Corr: 0.804***
Corr: -0.421***
Corr: -0.807***
Corr: -0.780***
Corr: -0.069
Corr: 0.164*
Corr: 0.749***
Corr: -0.155*
Corr: -0.694***
Corr: -0.638***
Corr: 0.237**
Corr: 0.115
Corr: -0.088
Corr: -0.726***
Corr: -0.688***
Corr: 0.265***
Corr: 0.272***
Corr: 0.165*
Corr: 0.164*
Corr: 0.278***
Corr: 0.057
Corr: 0.973***
Corr: -0.101
Corr: -0.276***
Corr: -0.086
Corr: -0.290***
Corr: 0.426***

Lasso



	s0
(Intercept)	-3.87506232
(Intercept)	.
ContinentAustralia	0.48741126
ContinentAsia	.
ContinentNorth America	-0.30404803
ContinentSouth America	-0.43997634
ContinentAfrica	.
Human_developmentHigh	-0.04631958
Human_developmentMedium	.
Human_developmentLow	0.96509470
Maternal_mortality	3.74029846
Adolescent_birth_rate	3.01493243
Seats_parliament	-1.33173057
F_secondary_educ	.
M_secondary_educ	.
F_Labour_force	-1.76333018
M_Labour_force	.
Gender_ratio	.
Development_statusDeveloping	.

Logistic Model

- None of the variables are significant predictors.

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -3.241e+01  1.183e+04  -0.003    0.998
ContinentAustralia         1.768e+01  9.311e+03   0.002    0.998
ContinentAsia             -2.266e+00  1.201e+04   0.000    1.000
ContinentNorth America    -2.265e+01  1.946e+04  -0.001    0.999
ContinentSouth America    -2.307e+01  1.821e+04  -0.001    0.999
ContinentAfrica           -3.486e+00  1.201e+04   0.000    1.000
Human_developmentHigh      1.256e+01  7.301e+03   0.002    0.999
Human_developmentMedium    3.090e+01  1.053e+04   0.003    0.998
Human_developmentLow       5.134e+01  1.465e+04   0.004    0.997
Maternal_mortality         5.678e-03  8.003e-03   0.710    0.478
Adolescent_birth_rate      7.339e-02  4.833e-02   1.519    0.129
Seats_parliament          -1.024e-01  1.420e-01  -0.721    0.471
```

Linear Discriminant Analysis (LDA)

Some variables have larger coefficients than others in absolute value

- "ContinentAustralia"
- "ContinentAfrica"
- "Human_developmentLow"
- "Development_statusDeveloping"

```
Coefficients of linear discriminants:
                                    LD1
ContinentAustralia            2.157380700
ContinentAsia                -0.207879965
ContinentNorth America       -0.802144950
ContinentSouth America       -0.907669235
ContinentAfrica               0.535623752
Human_developmentHigh        -0.451715251
Human_developmentMedium      -0.242017690
Human_developmentLow          0.993394611
Maternal_mortality            0.001621479
Adolescent_birth_rate         0.034620175
Seats_parliament             -0.017326105
F_secondary_educ             -0.034663751
M_secondary_educ              0.032835468
F_Labour_force               -0.024236189
M_Labour_force               -0.004174159
Gender_ratio                  0.006680092
Development_statusDeveloping -0.544712266
```

LDA

predict_status	true_status 0	1
0	63	3
1	1	18

accuracy	sensitivity	specificity
0.9529412	0.984375	0.8571429

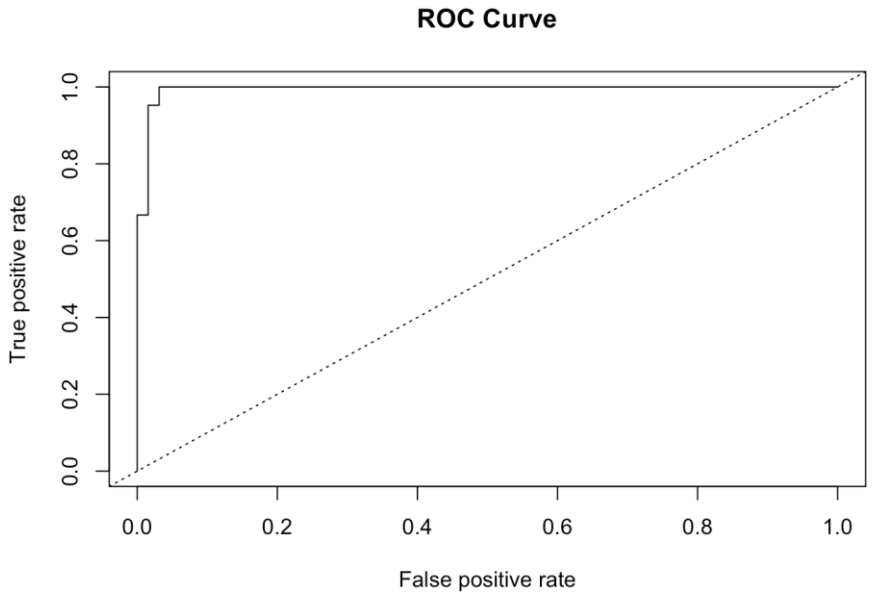
The accuracy score measure of how well the model performs overall.

The sensitivity score measure of how well the model performs in identifying positive cases.

The specificity score measure of how well the model performs in identifying negative cases.

LDA

- In this case, the ROC curve shows that the LDA model has an excellent performance, as the curve is very close to the top-left corner of the plot.
- This indicates that the model has a high true positive rate (TPR) and a low false positive rate (FPR) across all threshold settings.
- The AUC (area under the curve) is also very high at 0.994, which further confirms the high predictive power of the LDA model.

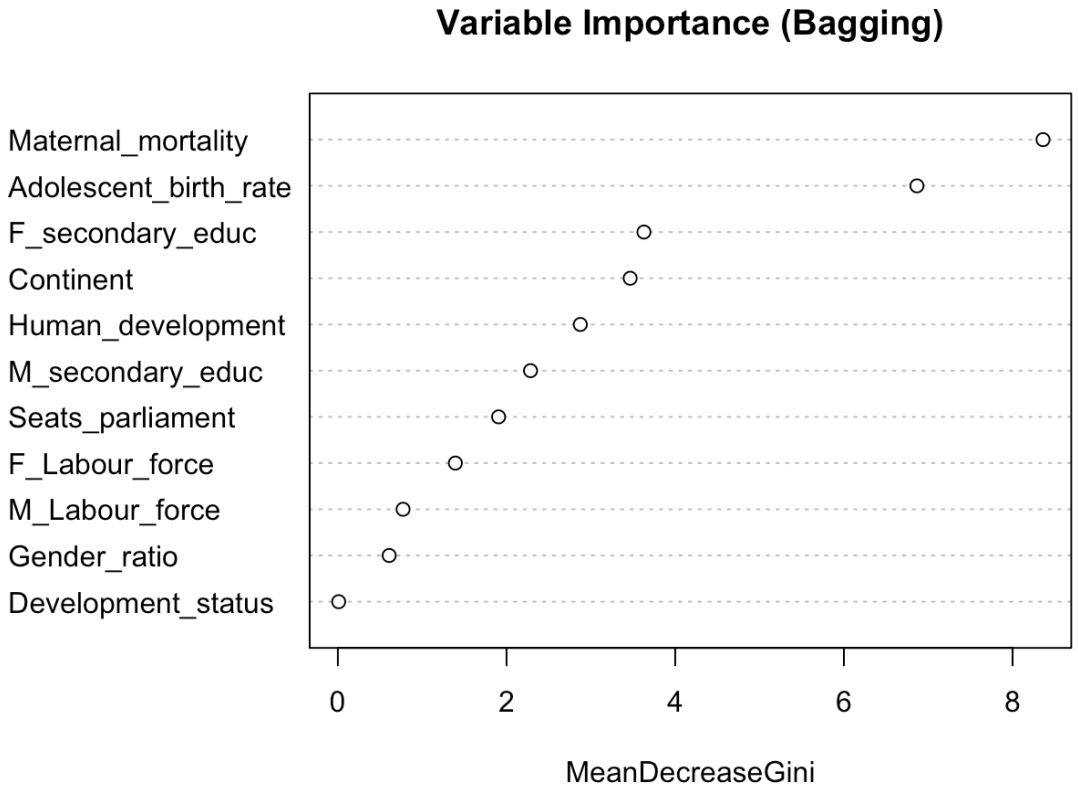



```
auc = as.numeric(performance(pred, "auc")@y.values)
auc
```

```
## [1] 0.9940476
```

Random Forest

- The higher the value of MeanDecreaseGini, the more important that variable is for the model.
- `Maternal_mortality` has the highest importance value



accuracy	sensitivity	specificity
0.9764706	0.984375	0.952381

Support Vector Machine

- We can see that the error rate is relatively low across all of the cost values tested, but the lowest error rate is achieved at the smallest value of cost.
- This suggests that a simpler model with a lower cost may be more appropriate for this data set.

```
- best parameters:
       cost
 0.04641589

- best performance: 0.04861111

- Detailed performance results:
           cost      error dispersion
1  1.000000e-03 0.26111111 0.13221626
2  3.593814e-03 0.26111111 0.13221626
3  1.291550e-02 0.07083333 0.06122674
4  4.641589e-02 0.04861111 0.06288462
5  1.668101e-01 0.04861111 0.06288462
6  5.994843e-01 0.05972222 0.08205031
7  2.154435e+00 0.09722222 0.11111111
8  7.742637e+00 0.07222222 0.10053867
9  2.782559e+01 0.07222222 0.10053867
10 1.000000e+02 0.10694444 0.08760894
```

```
Call:
svm(formula = GII_binary ~ ., data = train_set, kernel = "linear",
    cost = 0.04641589, scale = FALSE)
```

accuracy	sensitivity	specificity
0.9411765	0.984375	0.8095238

(Intercept)	Continent	Human_development
1.545714e+00	0.000000e+00	-4.641589e-02
Maternal_mortality	**Adolescent_birth_rate**	**Seats_parliament**
-1.660275e-02	1.660275e-02	4.641589e-02
F_secondary_educ	**M_secondary_educ**	**F_Labour_force**
0.000000e+00	-7.132585e-03	7.132585e-03
M_Labour_force	**Gender_ratio**	**Development_status**
0.000000e+00	-6.506973e-03	-7.953547e-02

Results

- LDA
- Random Forest
- SVM

accuracy	sensitivity	specificity
0.9529412	0.984375	0.8571429

accuracy	sensitivity	specificity
0.9764706	0.984375	0.952381

accuracy	sensitivity	specificity
0.9411765	0.984375	0.8095238

Results & Challenges

- LDA
  - Continent
  - Human Development
  - Development Status
- Random Forest
  - Maternal Mortality
  - Adolescent Birth Rate
  - Development Status
- SVM
  - Continent
  - Female Secondary Education
  - Male Labor Force

Thank you!