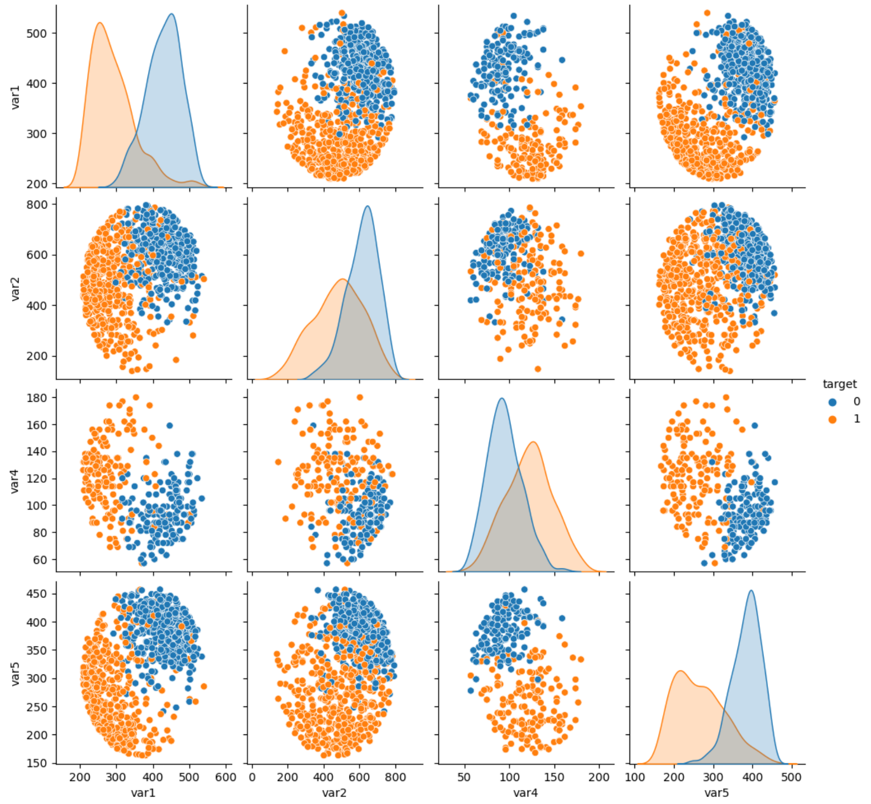
Name: Mosab Mohamed
Email: o.mohamed@innopolis.university

2. Theoretical Part

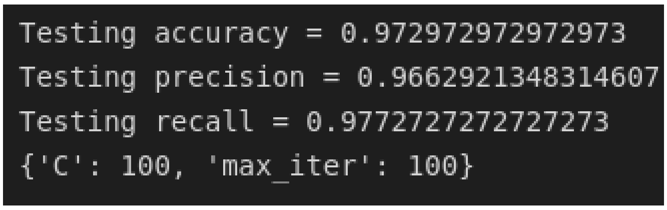
2.1) Regarding The Preprocessing

- **Which regression model was the most effective for the missing values, and why?**
- **Answer:**
    - The most effective model was Polynomial Regression with degree=3.
    - It's better than linear regression because the data is not linear.
    - And degree 3 is the best because less than that would result in underfitting and more than that would be overfitting.
    - 

- **What encoding technique did you use for encoding the categorical features, and why?**
- **Answer:**
    - I used one hot encoding for both var3 and var6
    - First I used it on var6 because it only consists of yes/no values which will be encoded to 2 columns and then reduced to one column because the 2 columns are redundant.
    - Then I used it on var3 for the sole reason that there are no ordinal relationships between the countries, and it might even mislead the model.

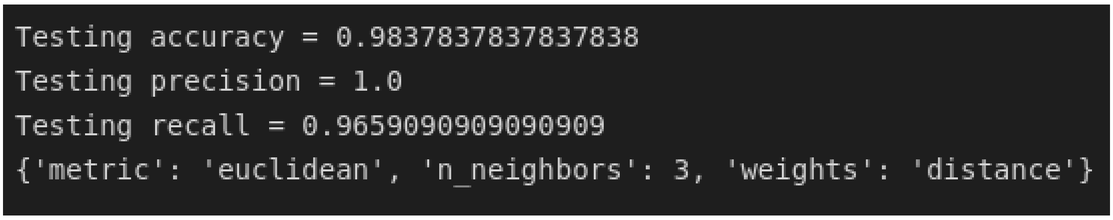
2.2) Regarding the training process

- **Which classification model performed best, and why?**
- **Answer:**
    - The different model performances with relatively close to each other, But the clear winner was KNN without PCA and with hyperparameters = {'metric': 'euclidean', 'n_neighbors': 3, 'weights': 'distance'}
    - Logistic:



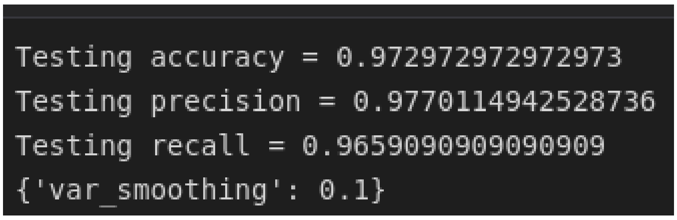
```
Testing accuracy = 0.972972972972973
Testing precision = 0.9662921348314607
Testing recall = 0.9772727272727273
{'C': 100, 'max_iter': 100}
```

    - KNN:



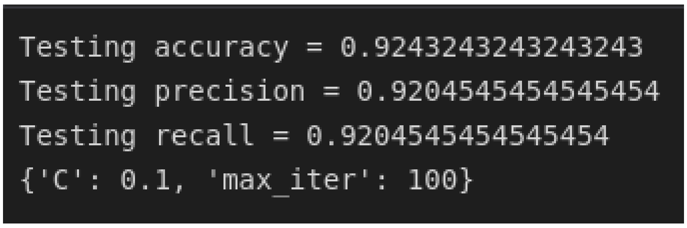
```
Testing accuracy = 0.9837837837837838
Testing precision = 1.0
Testing recall = 0.9659090909090909
{'metric': 'euclidean', 'n_neighbors': 3, 'weights': 'distance'}
```

    - Naive Bayes:



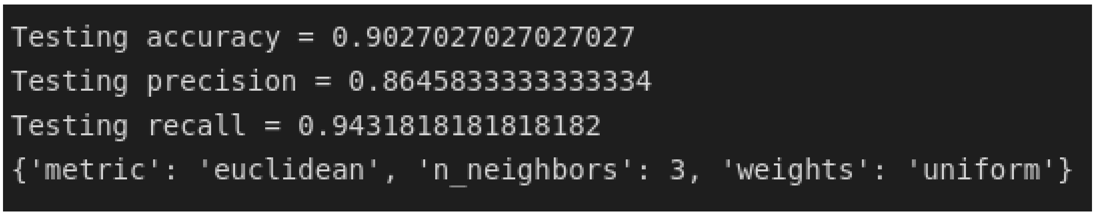
```
Testing accuracy = 0.972972972972973
Testing precision = 0.9770114942528736
Testing recall = 0.9659090909090909
{'var_smoothing': 0.1}
```

    - Logistic with PCA:



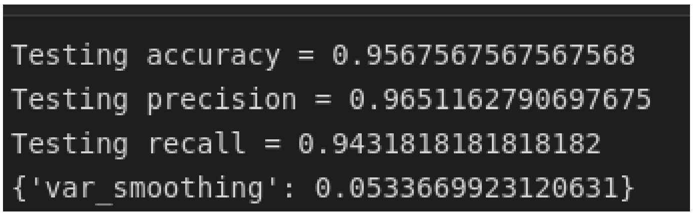
```
Testing accuracy = 0.9243243243243243
Testing precision = 0.9204545454545454
Testing recall = 0.9204545454545454
{'C': 0.1, 'max_iter': 100}
```

    - KNN with PCA



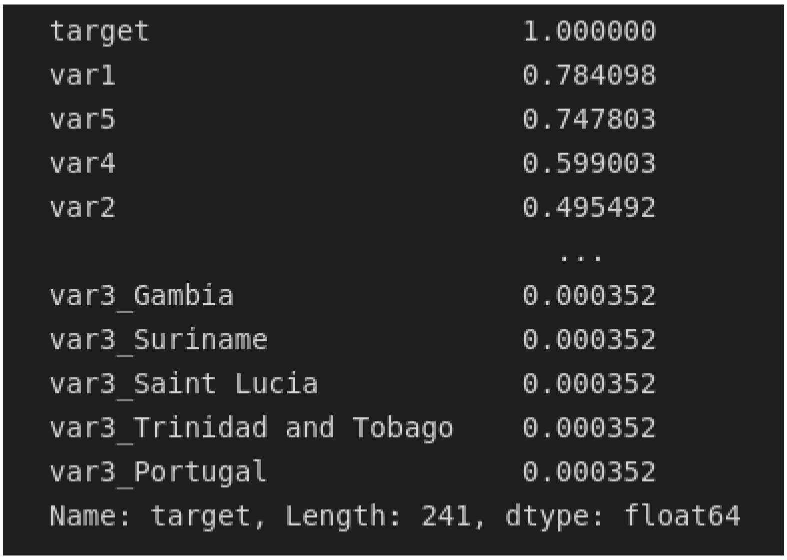
```
Testing accuracy = 0.9027027027027027
Testing precision = 0.8645833333333334
Testing recall = 0.9431818181818182
{'metric': 'euclidean', 'n_neighbors': 3, 'weights': 'uniform'}
```

    - Naive Bayes with PCA

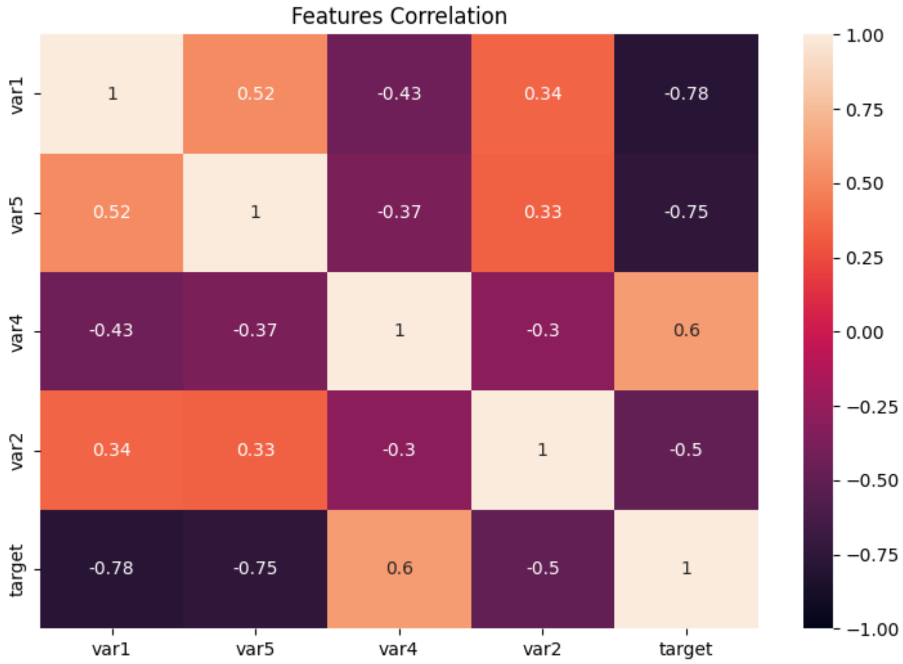


```
Testing accuracy = 0.9567567567567568
Testing precision = 0.9651162790697675
Testing recall = 0.9431818181818182
{'var_smoothing': 0.0533669923120631}
```

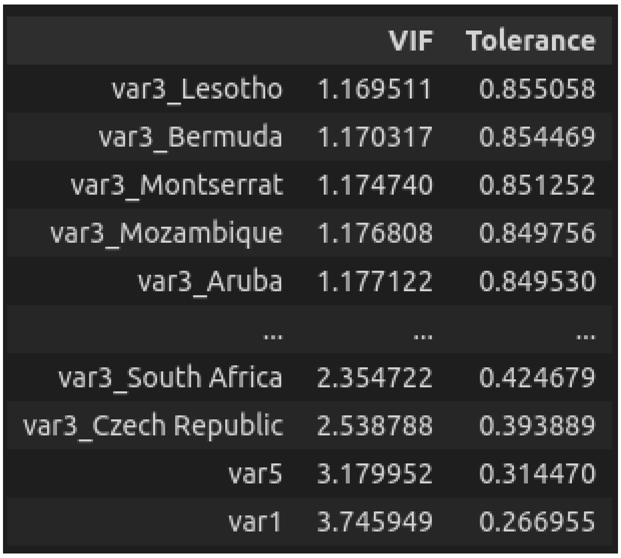
    - Because KNN handles outliers better than Naive Bayes and KNN gives non-linear solutions unlike logistic regression.

- **What were the most critical features with regards to the classification, and why?**
- **Answer:**
    - Var1, Var5, Var4, Var2
    - 

```
target                        1.000000
var1                          0.784098
var5                          0.747803
var4                          0.599003
var2                          0.495492
                                ...
var3_Gambia                   0.000352
var3_Suriname                 0.000352
var3_Saint Lucia              0.000352
var3_Trinidad and Tobago      0.000352
var3_Portugal                 0.000352
Name: target, Length: 241, dtype: float64
```

    - Because they had the highest correlation ranging above 0.4 while everything else was less than 0.01.
    - 

- **What features might be redundant or are not useful, and why?**
- **Answer:**
    - Var3 and Var6
    - They were not useful due to their low correlation with the target.

- **Did the dimensionality reduction by the PCA improve the model performance, and why?**
- **Answer:**
  - No
  - Because there was no multicollinearity in our dataset which resulted in us losing information because of the dimensionality reduction and not gaining anything in return because there was no multicollinearity
  - 

	VIF	Tolerance
var3_Lesotho	1.169511	0.855058
var3_Bermuda	1.170317	0.854469
var3_Montserrat	1.174740	0.851252
var3_Mozambique	1.176808	0.849756
var3_Aruba	1.177122	0.849530
...	...	...
var3_South Africa	2.354722	0.424679
var3_Czech Republic	2.538788	0.393889
var5	3.179952	0.314470
var1	3.745949	0.266955

  - Since we have no variable with VIF greater than 4 and no Tolerance smaller than 0.25, we can safely assume that there's no multicollinearity

- **Additional research:**
  - **What is a multi-label learning problem?**
  - **Answer:** Multi-label learning studies the problem where we find a model that maps the input to one of the multiple classes.

  - **Suggest an example in which you can transform the given problem into a multi-label problem ?**
  - **Answer:** For example, if we think of our problem as an ecommerce website and the dataset is the user's data and we predict something about those users, for example: their salary or income. With the target being something like this [Below Average, Average, Above Average].

  - **Will the models work as it is in that case, or would some changes be required?**
  - **Answer:** For Logistic Regression and Naive Bayes they will not work because they can't handle multiclass problems. KNN on the other hand will work fine but we will need to change the target features.
  - And to make Logistic Regression and Naive Bayes work on multi label problems we will need extensions or transform the problem somehow.