Why did you choose this model?

As a second ML model we decided to use a Random Forest Regressor model for analysing our dataset. A Random Forest Regressor Model has several advantages over other regression algorithms like an improved accuracy, handling non- linear relationships, little pre-processing required, robust outliers and shows a feature importance. Especially important and exciting for us is to analyse and explore the feature importance and their impact on the suicide rates. To know which factors are especially impacting the suicide rates can give our analysis a valuable depth and support people's health in the longterm.

How will you train the model?

For training the model we followed the following steps:

1. Prepare the data: Clean and preprocess the data and bring in suitable format
2. Select the right features: we are using the suicide rate as our y and the features of our dataset for the X
3. Split the data in training and testing data sets
4. Train the random forest model on the training data
5. Optimise the model by tuning the different hyperparameters like number of trees
6. Evaluate the model and its performance with R squared, MAE Mean Absolute Error, Mean Squared Error and RMSE.
7. Use the model to make predictions on new unseen data

What's the accuracy of the model?

```
In [18]: # Model Evaluation
         print('R^2:',metrics.r2_score(y_train, y_pred))
         print('Adjusted R^2:',1 - (1-metrics.r2_score(y_train, y_pred))*(len(y_train)-1)/(len(y_train)-X_train.shape[1]-1))
         print('MAE:',metrics.mean_absolute_error(y_train, y_pred))
         print('MSE:',metrics.mean_squared_error(y_train, y_pred))
         print('RMSE:',np.sqrt(metrics.mean_squared_error(y_train, y_pred)))

         R^2: 0.88522573150271
         Adjusted R^2: 0.8823152483985891
         MAE: 1.1354442054328884
         MSE: 8.841259491670437
         RMSE: 2.9734255483651237
```

1. **R^2 (Coefficient of Determination):** This metric measures the proportion of the variation in the target variable that is explained by the model. R squared measures how well the model fits the data and ranges from 0 to 1. A score of 0.88 means the the model fits quite good and is in 88% of the cases accurate.
2. **Adjusted R^2:** Similar to R^2, adjusted R^2 penalises models with a large number of features by reducing their R^2 scores. An adjusted R^2 of 0.88 indicates that the model explains 88% of the variance in the target variable, taking into account the number of features used.
3. **Mean Absolute Error (MAE):** This metric measures the average absolute difference between the actual and predicted values. A lower MAE indicates a better fit. Unlike mean squared error (MSE), which punishes large errors more heavily, MAE gives equal weight to all errors, regardless of their size. **In our case the MAE is 1.15**. And a low MAE shows that the model is making more accurate predictions and less errors.
4. **Mean Squared Error (MSE):** This metric measures the average squared difference between the actual and predicted values. A lower MSE indicates a better fit. **The MSE is 8.84.** The MSE punishes large errors more heavily.
5. **Root Mean Squared Error (RMSE):** This metric is the square root of the MSE and it provides an interpretable error metric in the same units as the target variable. A lower RMSE indicates a better fit. **The RMSE is 2.97.**

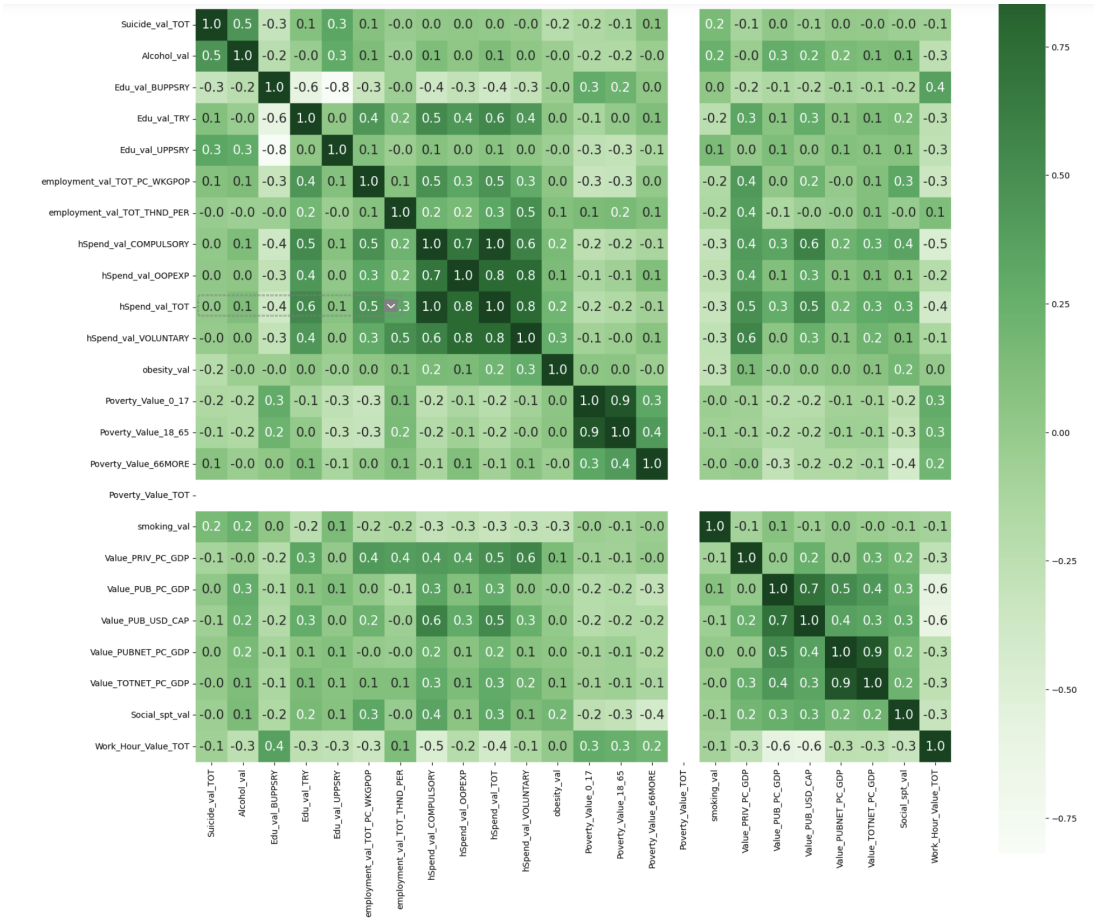
In general with a R^2 and adjusted R^2 of 0.88 our model indicates that the data fits the model quite well and a large proportion of the variation in the target variable is explained. The MAE, MSE, and RMSE are all relatively low, indicating that the model has a good fit.

How does the model work?

A random forest regressor is an ensemble machine learning model. For its prediction it uses multiple decision trees. The individual trees are trained on a random subset of the data and it predicts the target variable. At the end the predictions from all trees are then combined into the final prediction. The combination of many trees and individual random subsets of data makes it to a robust and accurate prediction model.

Heatmap

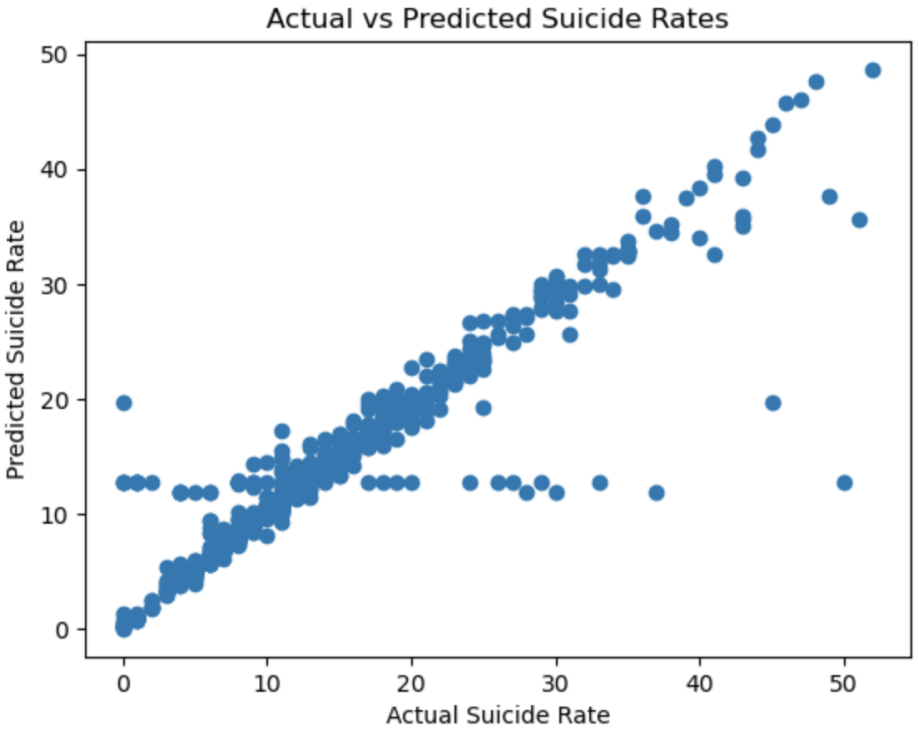
As an addition analysis we calculate a heatmap to see which features are correlating between each other and the suicide rate the most. The scale is from -1 to 1. From our understanding is -1 a very low correlation and 1 a high correlation.



The following features have a high correlation (0.7 or greater) to the suicide rate in the analysed countries. Health spending, Poverty rate up to the age of 65 and general economy in the country have the highest impact on the suicide rate.

hSpend_val_COMPULSORY
hSpend_val_OOPEXP
hSpend_val_TOT
hSpend_val_VOLUNTARY
Poverty_Value_0_17
Poverty_Value_18_65
Value_PUBNET_PC_GDP
Value_TOTNET_PC_GDP

Actual vs. Predicted Suicide Rates



The graph shows by using the trainings dataset, that the actual vs predicted suicide rate shows with the exception of some outliers a close to linear graph. That means that our model works pretty accurately in its prediction.