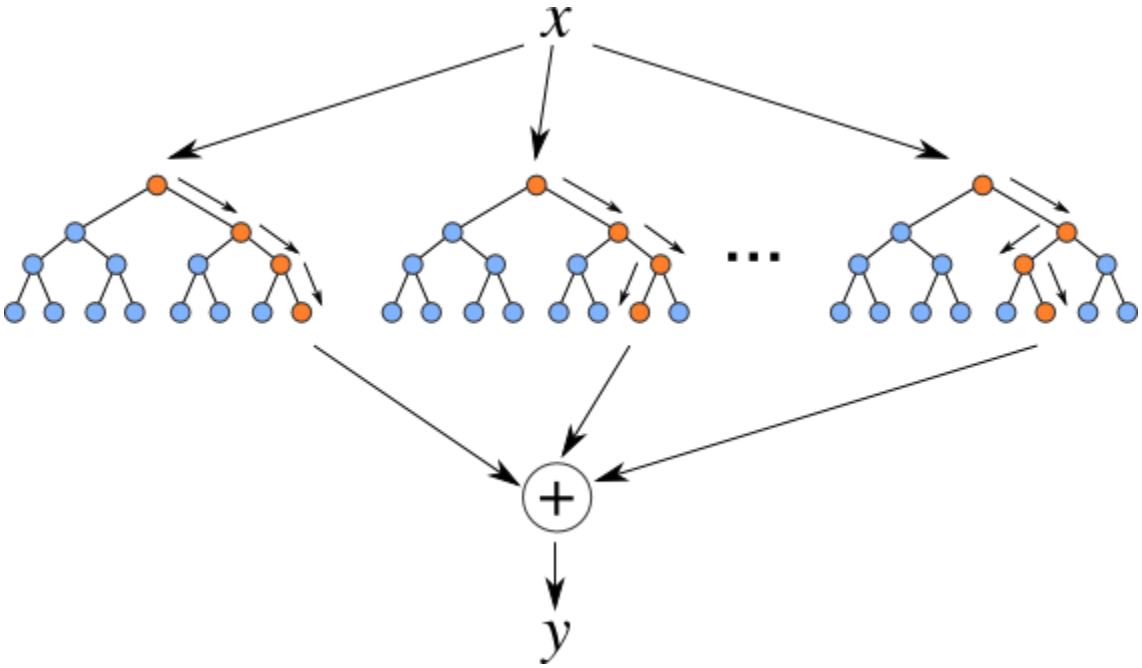
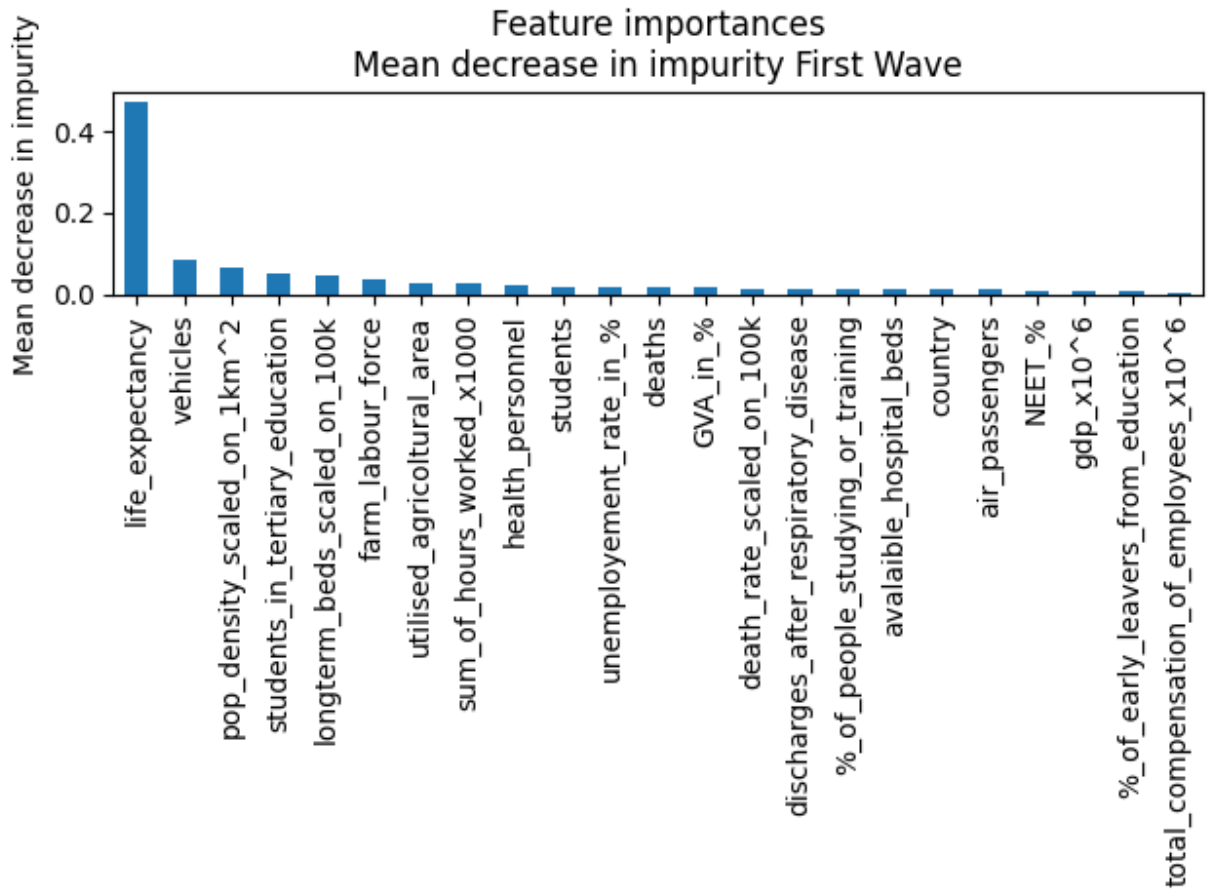
Features Selection Using Random Forests



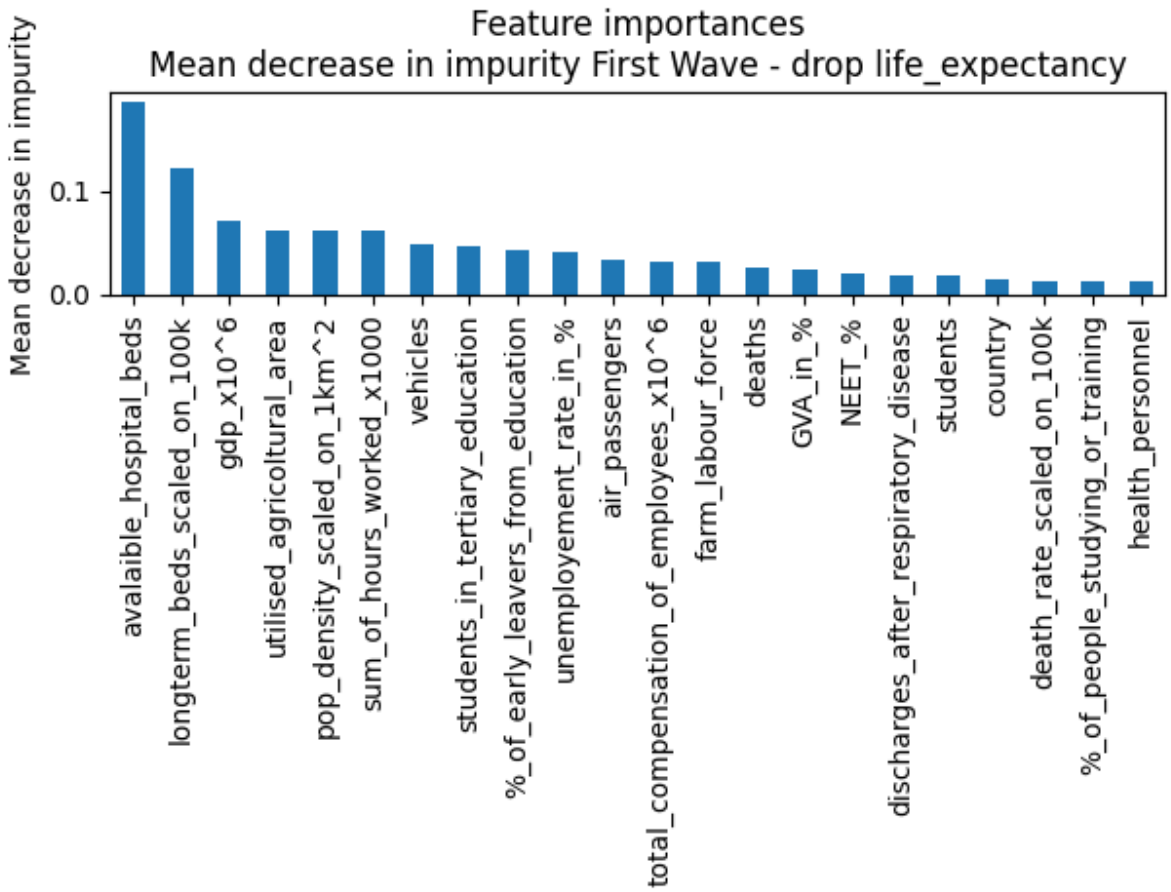
Random forests usually have high predictive power due to their low variance resulting from averaging results from multiple trees which are trained in a way to ensure that there is little dependence between them. Additionally, their attractiveness stems from pretty good interpretability since decision trees select features which carries the most information to perform prediction.

Furthermore, it can be trained on datasets with missing values, which is perfect for our scenario. For example Lasso regression also has feature selection mechanism embedded, however, it would require filling the missing data which could lead to introduction of some kind of bias into the process of feature selection.



Mean Absolute Error = 0.73

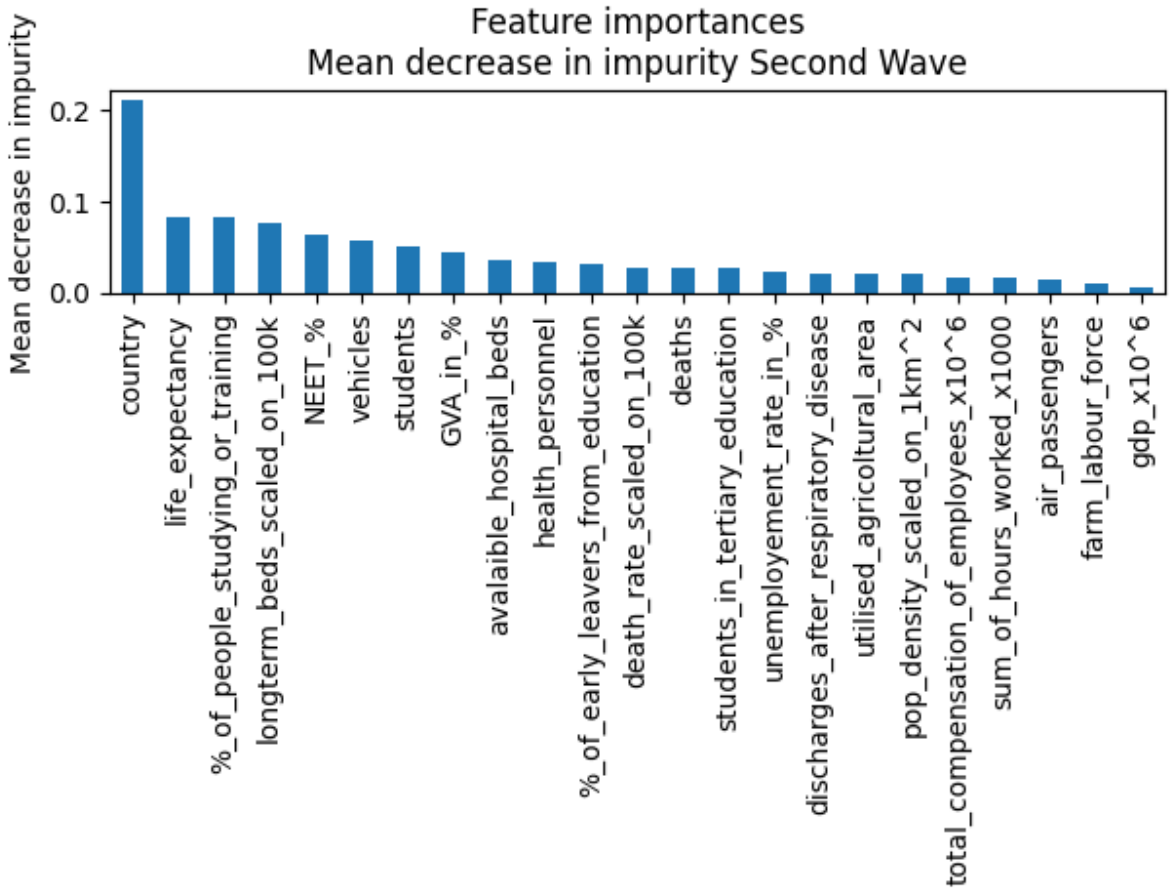
In order to predict density of cases during first wave it seems that decision of our model is mainly based on **life_expectancy** feature. It is, however, hard to distinguish another valuable features so in the next trial life_expectancy feature was deleted from the dataset and model was trained again.



Mean Absolute Error = 0.85

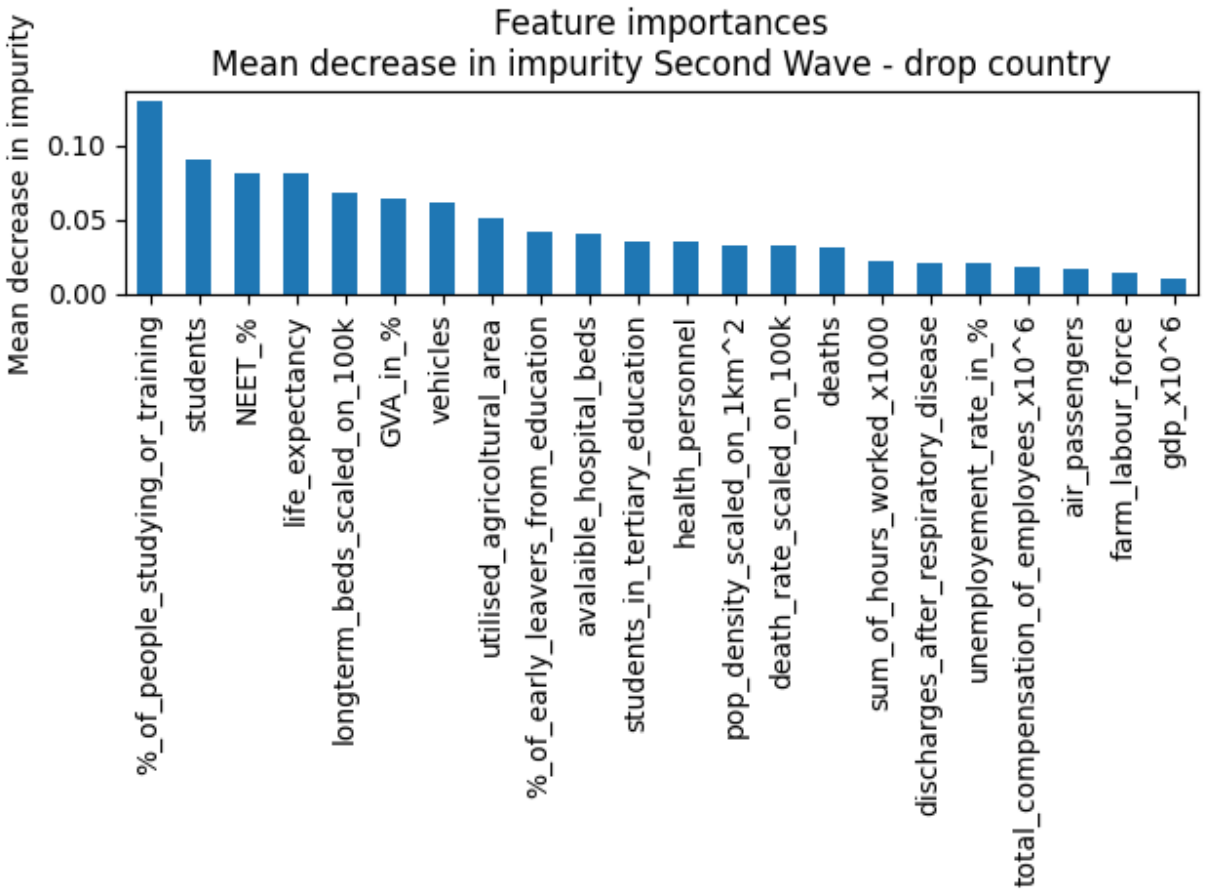
We can see that quality of prediction slightly decreased, but this experiment revealed other significant features like **available_hospitals_beds, longterm_beds_scaled_on_100k or gdp_x10^6**.

ME



Mean Absolute Error = 0.99

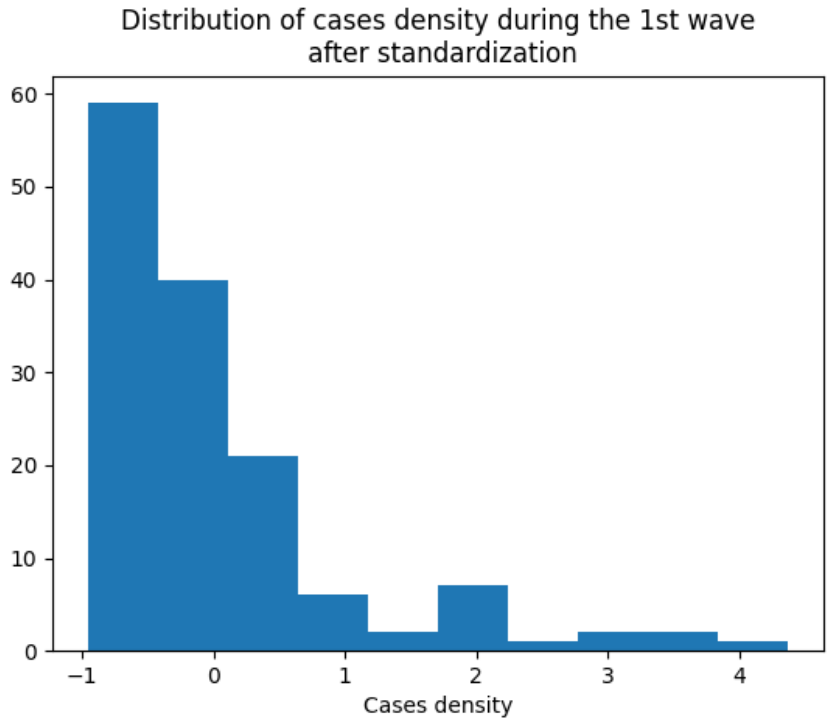
In case of the second wave the most significant feature seems to be a **country**, which might be really reasonable since range of regulations were introduced by countries (which wasn't possible during the first wave as countries weren't that prepared) which could have a determining impact on number of infections.

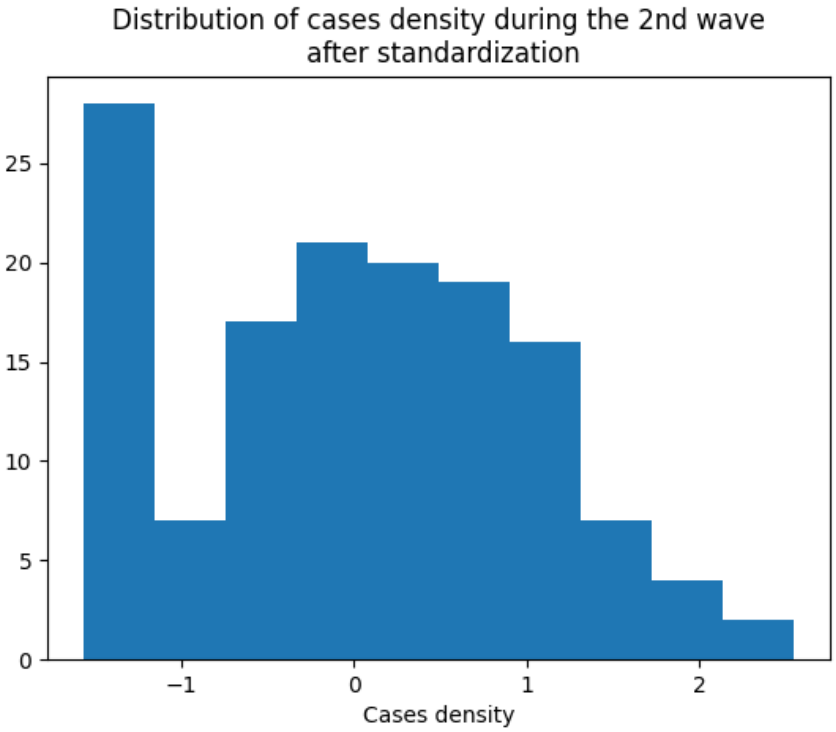


Mean Absolute Error = 1.01

We have performed the same experiment of dropping the most significant feature and checking performance of the model trained on the remaining features. This time quality of predictions didn't really decrease. It however makes sense since, as it can be seen on the PCA plot, observations from the same country tend to be close to each other in the features space.

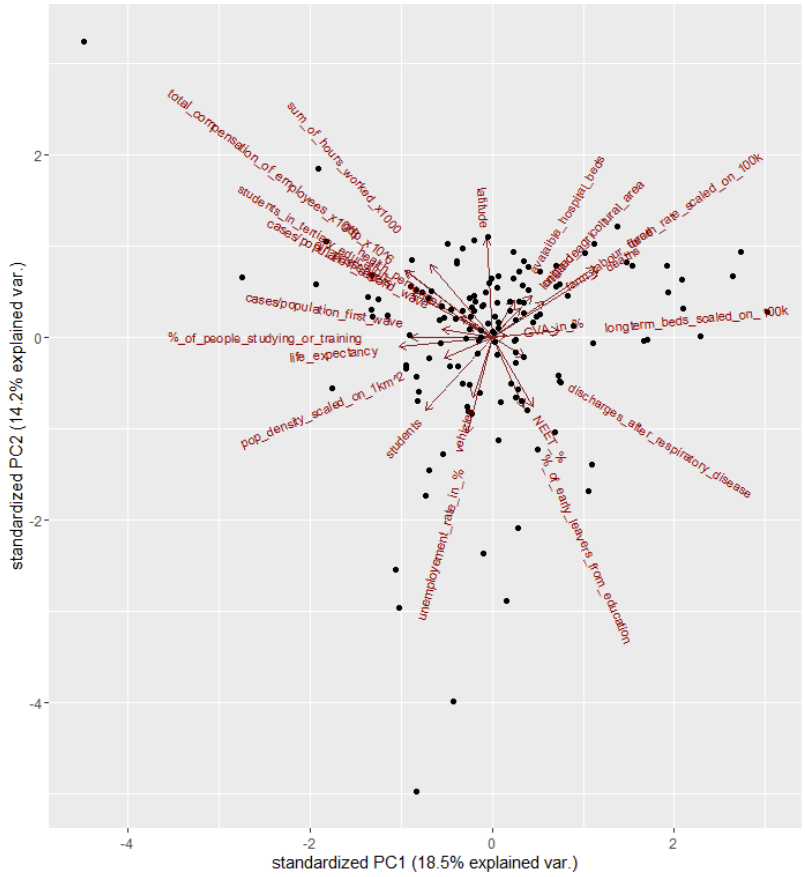
The most significant features seem to be **%_of_people_studying_or_training, students, NEET_%, life_expectancy and longterm_beds_scaled_on_100k.**





Distributions of standardized cases densities reveal that quality of predictions isn't very impressive as they tend to cover most common values in our dataset. It implies that there is still space to explain this variability by introducing additional information to our dataset.

PCA

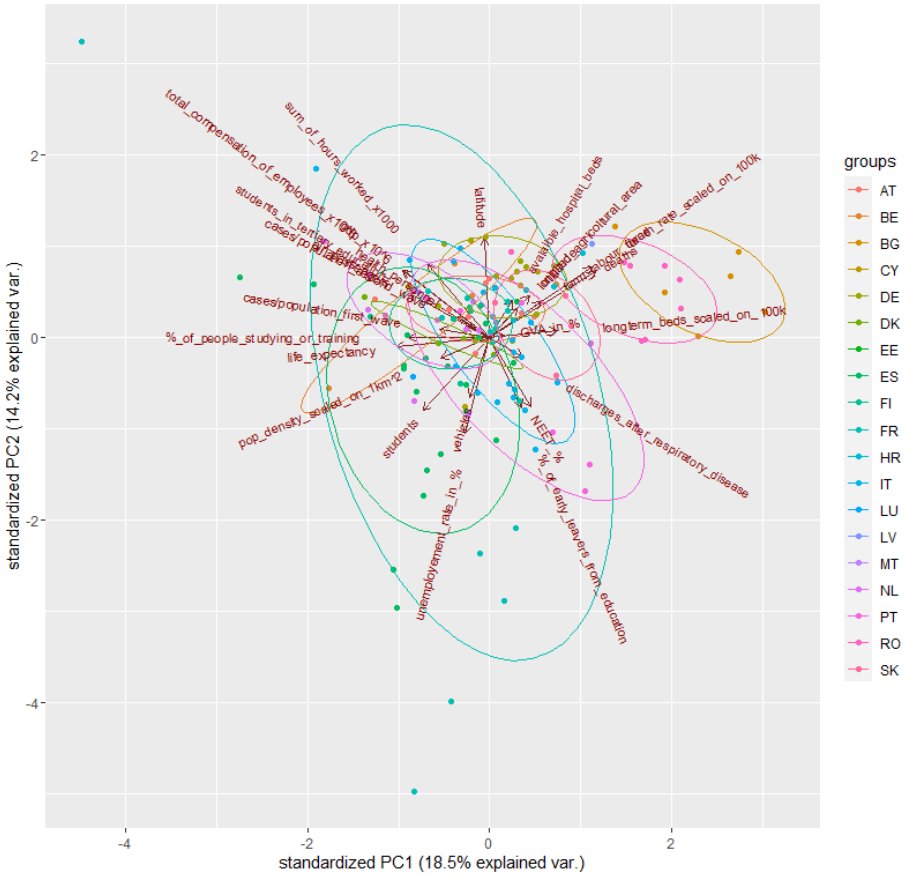


Visualization of PCA also gives a valuable insight into characteristics of the dataset. For example we can clearly see that regions with high values of **%_of_people_studying_or_training** and **life_exptectancy** tend to have high density of cases during the first wave, while these with high values of **GVA_in_%** and **longterm_beds_scaled_on_100k** had less infections.

Density of cases during the second wave is however strongly correlated with **air_passengers, students_in_tertiary_education, total_compensation_of_employees, sum_of_hours_worked**.

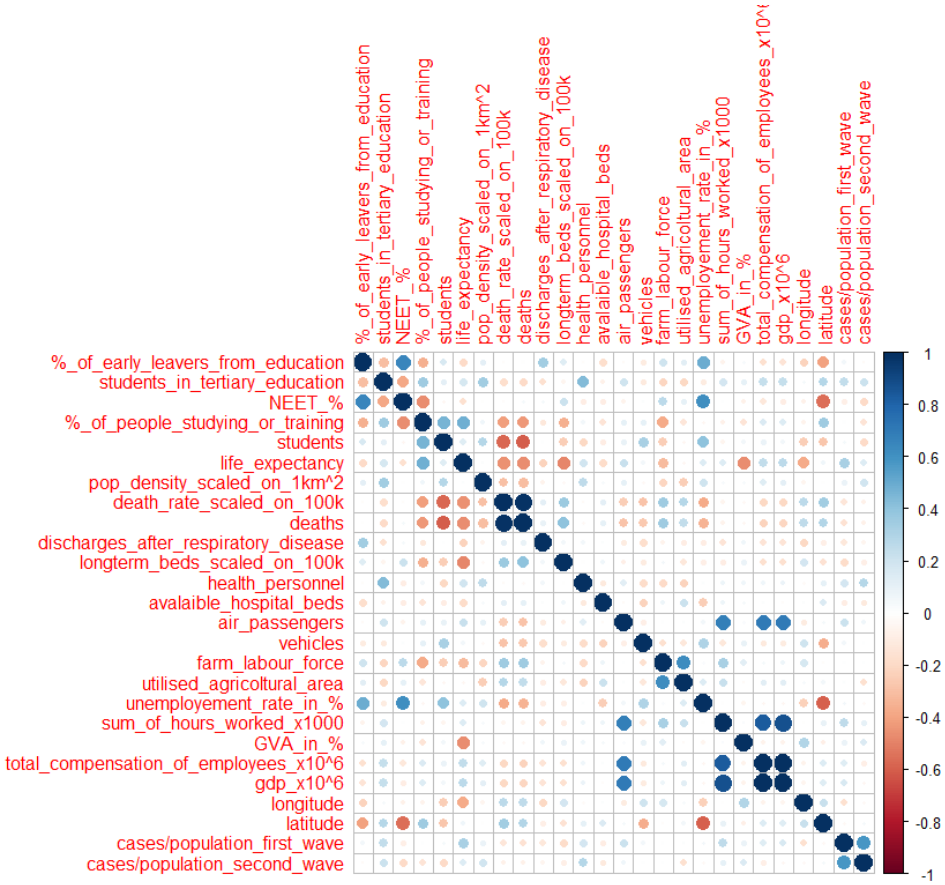
All these observations seem to have a reasonable, intuitive explanation.

PCA – observations grouped by countries



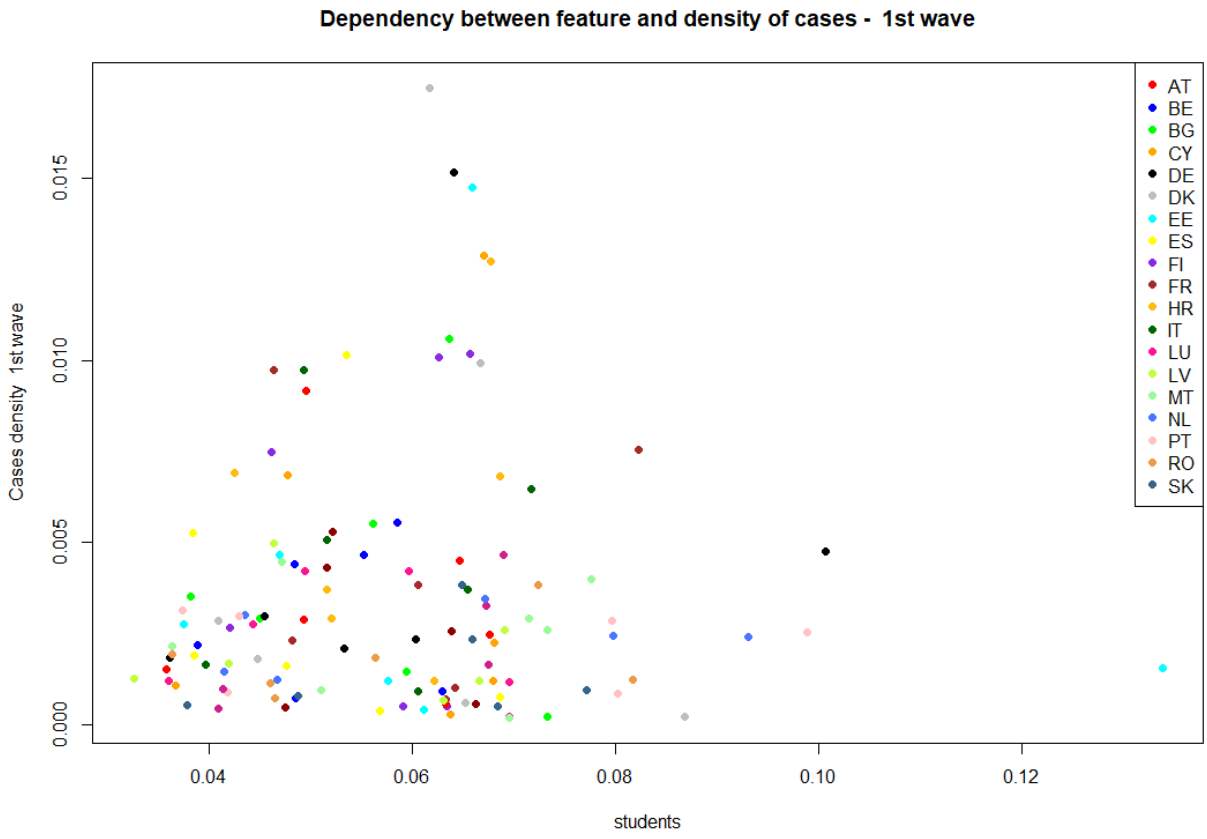
Moreover, grouping of observations by countries highlights that observations from the same countries tend to be close to each other in the features space, which already was mentioned.

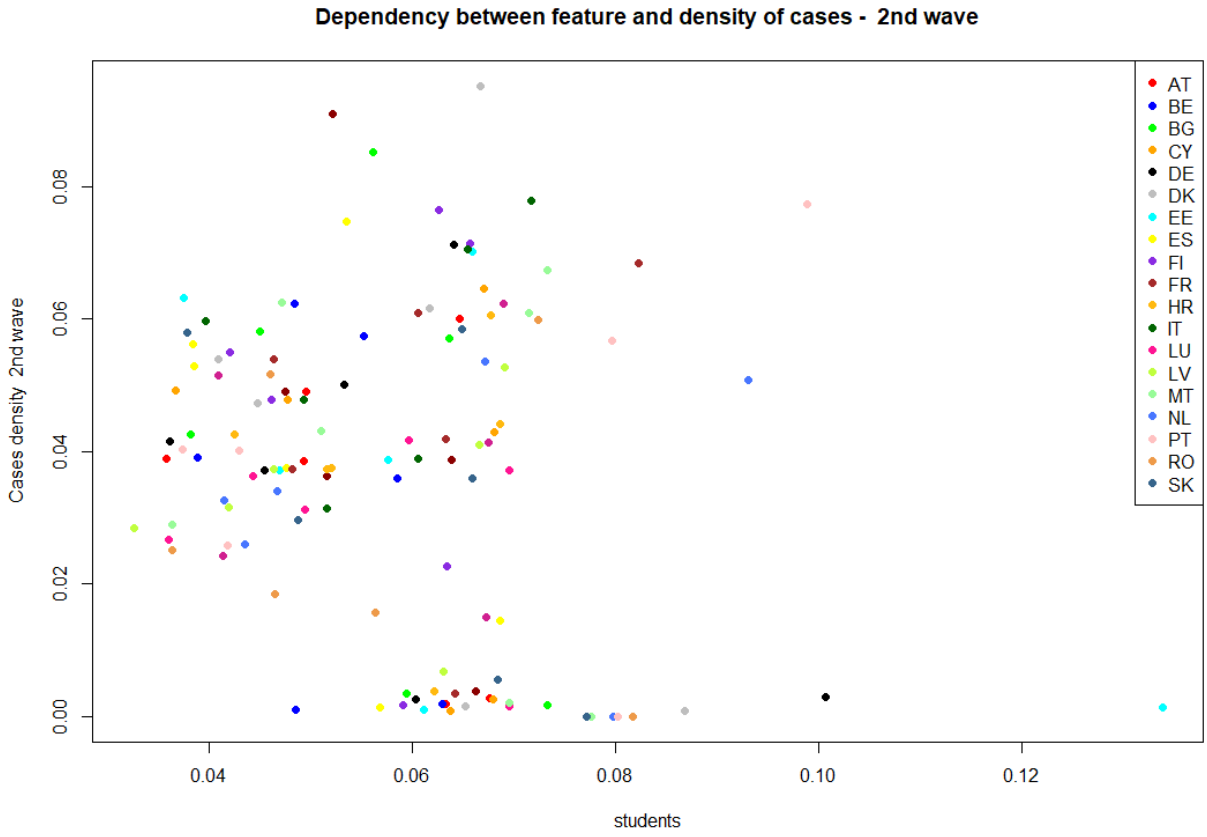
Correlation Matrix



For the sake of calculating correlation between variables, missing data has been filled with average within the country and if that wasn't possible with average within the whole column. Obviously, it was done just in order to perform initial analysis and this approach might change after some deeper consideration.

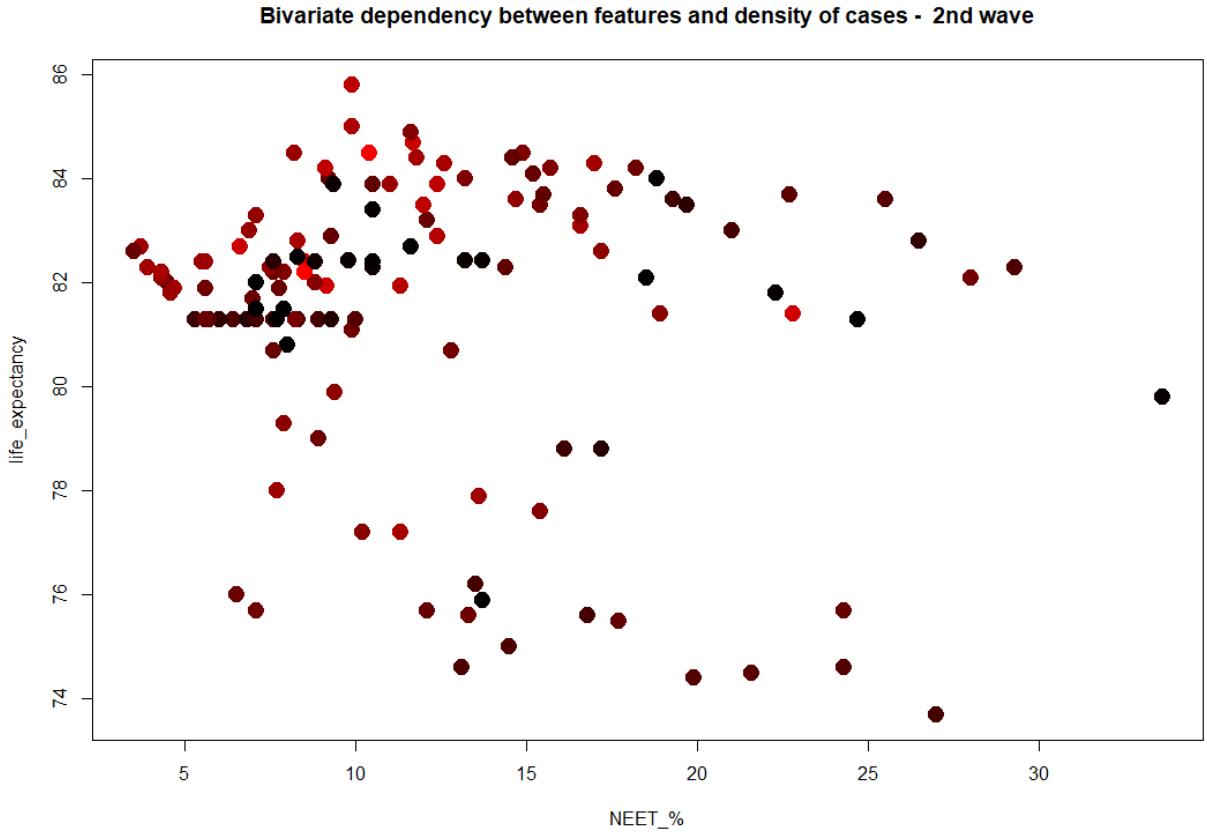
Plots of features against density of cases





For all the features which seem to be of significance plot against density of cases during first and second wave have been generated, like it can be seen above, however, no particular stronger dependency is observable.

Bivariate Analysis



In case of bivariate analysis, each pair of selected features has been plotted against each other. Value of cases density was included by introducing heatmap for observations. The more red the observation is the higher density of cases was observed in this region.

This analysis definitely requires devoting more time to observe potential interesting dependencies. One which was found is represented above. As we can see, greater life expectancy and the lower number of people which are not in employment nor training contribute to higher values of cases densities.