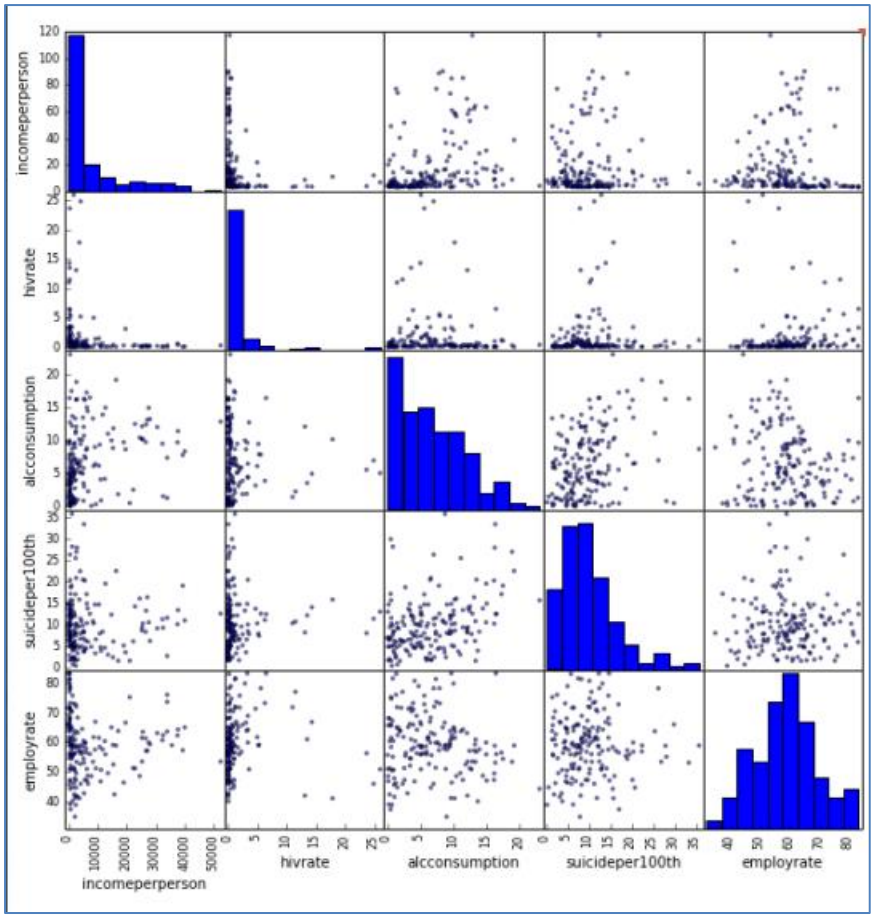
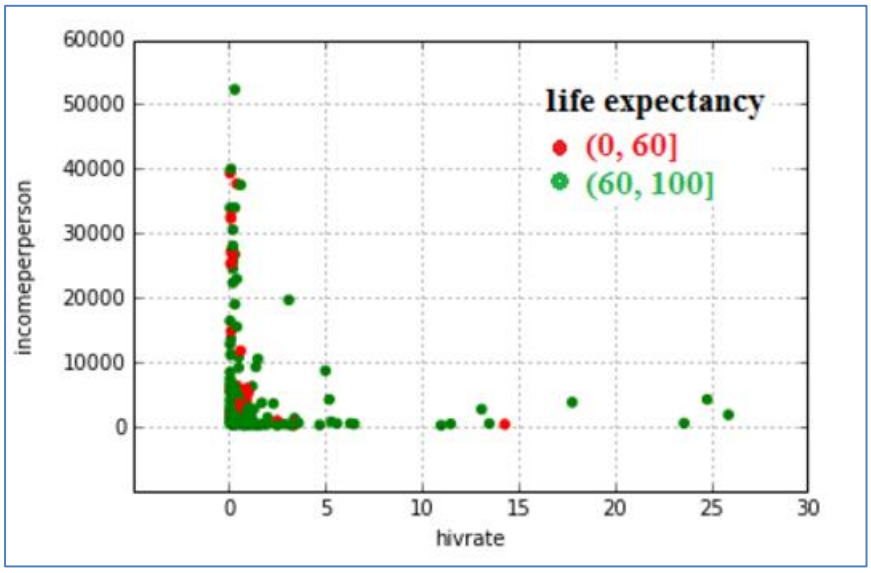
Finding the most important predictors to predict life expectancy and making predictions with the GapMinder dataset with Random Forest and ExtraTree Forest Ensemble Classifiers using Python Scikit Learn and Pandas

Ensemble learning using **Random Forest** and **ExtraTree Forest** were performed to evaluate the importance of a series of explanatory variables in predicting a binary, categorical response variable.
The following explanatory variables were included as possible contributors to a classification tree model evaluating **life expectancy** (my response variable, which is a continuous numeric variable but was binned into 2 categories: (0-60] and (60-100]), income per person, alcohol consumption, armed forces rate, breast cancer per 100th, co2 emissions, female employment rate, hiv rate, internet use rate, oil per person, polity score, relectric per person, suicide per 100th, employment rate, urbanization rate.

The following figure shows relations in between some of the predictors used and some exploratory visualizations:

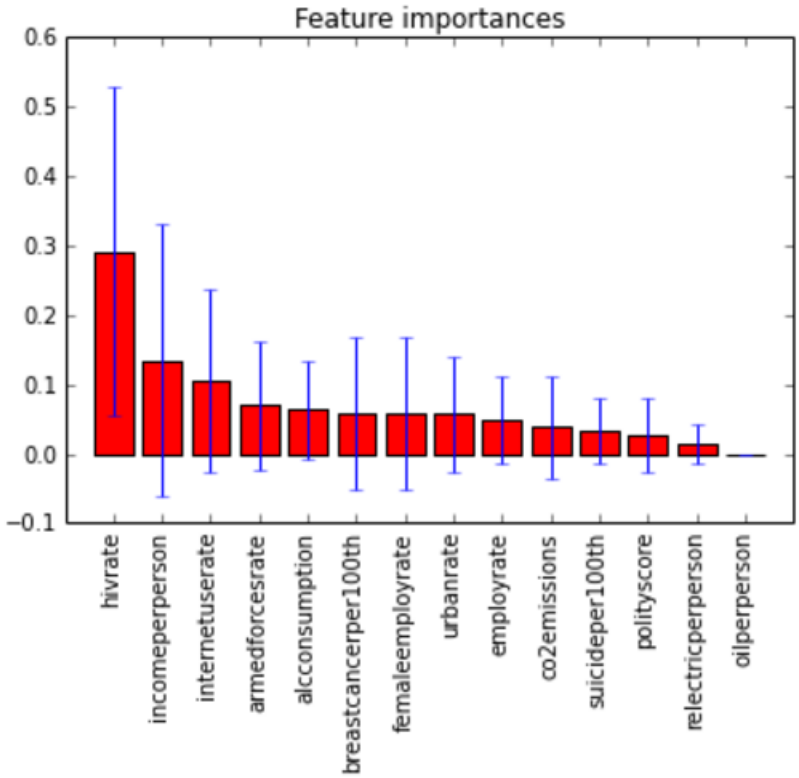




After removal of the NA values in the life expectancy variable, the predictor variables in the original dataset were imputed, the missing values in the numeric columns were replaced with median values. Then the dataset was divided (by taking a random sample of size 60% of the entire dataset) into training dataset with 114 data tuples and test dataset with 77 data tuples. Then the Random Forest and Extra Tree classifiers were trained on the trainign dataset and the models were used to predict on the test dataset. Ans ensemble of 25 decision trees were used to build the random forest predictor and gini index measure was used for the best feature selection at each round for the decision trees.

As can be seen from the 3 most important predictors selected by the ExtraTree Forest model were: hiv rate, income per person and internet user rate.

The model learnt from the training dataset was used to predict the life expectancy for the countries in the test dataset. The confusion matrix (contingency table) on the test dataset is shown below, which shows that we obtained ~**90.9% accuracy** on the held-out unseen dataset.



Predicted	(0, 60]	(60, 100]
Actual		
(0, 60]	13	7
(60, 100]	0	57