# CANCER RATES PER COUNTRY – DETERMINING THE IMPORTANCE OF COUNTRY LEVEL FACTORS USING RANDOM FOREST REGRESSOR

**Sandi Baressi Šegota[1], Nikola Anđelić[1], Ivan Lorencin[1], Jelena Musulin[1], Daniel Štifanić[1], Matko Glučina[2], Zlatan Car[1]**

[1] Faculty of Engineering,
The University of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia
Email: sbaressisegota@riteh.hr , nandelic@riteh.hr , ilorencin@riteh.hr, jmusulin@riteh.hr , dstifanic@riteh.hr , car@riteh.hr
[2] University of Rijeka,
Trg Braće Mažuranića 10, 51000 Rijeka, Croatia
Email: matko.glucina@uniri.hr

## Abstract

Cancer is one of the most discussed diseases in modern healthcare. Cancer rates vary by country, this indicates that there are factors that might influence the occurrence of cancer depending on the country. In this paper, the authors present the dataset which consists of cancer rates (CR) for 42 countries, with 10 possible country-level factors - Health Care Index (HCI), country population (POP), percentage of people living in the urban areas - urbanization rate (UR), nominal GDP (GDP), life expectance (LE), the birth rate per thousand (BR), the death rate per thousand (DR), $CO_2$ Emission per capita ($CO_2$), the percentage of the population that has access to the sanitation facilities (SFA), the percentage of the population that has access to a clean water source (WSA). We shall analyze the created dataset, and the influence of individual inputs is modeled and tested using the Random Forest (RF) algorithm. The results indicate that $CO_2$ emissions, BR, and the GDP have the highest influence according to the applied RF feature importance analysis.

**Key words:** cancer rates, correlation analysis, feature importance analysis, machine learning, random forest algorithm

## 1 Introduction

Cancer is a group of diseases that involve abnormal cell growth and have the potential to invade or spread to other parts of the body [1]. Cancer is caused by mutations of the cell DNA, which may be triggered by many environmental factors such as radiation, viruses and infections, type of diet, physical activity, smoking and tobacco [2, 3], etc. The occurrence of cancer differs from country to country and is expressed as the country cancer rate [4]. Due to it being an ensemble tree-based algorithm, the RF algorithm has been used in many fields to determine the importance of individual factors and has shown to be a robust algorithm for the task [5]. In this paper, the authors employ the RF algorithm to determine the influence of different country-level environmental and governmental factors on the cancer rates of that very country. The authors construct the used dataset, from data sourced from different sets describing various factors that can be considered to have an influence on the country's cancer occurrence rate. First, the created

dataset will be presented and discussed, and then followed by a brief description of the RF algorithm applied. Finally, results will be presented and discussed, with conclusions drawn at the end of the article.

## 2 Methodology

In this section our goal is to create the dataset, along with basic statistical analysis, followed by a brief description of the applied machine learning methodology.

### *2.1 The dataset*

The dataset is constructed by combining a number of datasets together. The data for cancer rates are provided by World Cancer Research Fund at American Institute for Cancer Research [6], while the values for the other inputs: Health Care Index, population, urbanization rate, GDP, average life expectancy, birth, and death rates, $CO_2$ emissions, and access to sanitation facilities and clean water sources are provided from various datasets collected from Our World in Data [7]. The provided data covers 50 countries [6], 8 of which are eliminated due to the lack of input values. The final dataset consists of 42 countries. The histograms of each individual have been observed and they show that the distributions of data are relatively clean and there are not any large outliers present in the data, signifying it can be used for further modeling. The descriptive statistics - minimal value, maximal value, range, median, and standard deviation have been calculated for each of the inputs. This shows that there are variations between individual inputs used, signifying the need for data scaling. Data is scaled using a fitted scaler which guarantees that the scaled data will have a zero mean and a unit variance, which will allow for easier data modeling using the RF algorithm [8]. The final dataset analysis performed is the correlation analysis to determine the apparent influence of the selected inputs on the CR. Correlation coefficients $r_{x^1,x^2}$ are calculated using:

$$r_{x^1,x^2} = \frac{\sum_{i=0}^{n}(x_i^1 - \frac{1}{n}\sum_{j=0}^{n} x_j^1)(x_i^2 - \frac{1}{n}\sum_{j=0}^{n} x_j^2)}{\sqrt{\sum_{i=0}^{n}(x_i^1 - \frac{1}{n}\sum_{j=0}^{n} x_j^1)^2 \sum_{i=0}^{n}(x_i^2 - \frac{1}{n}\sum_{j=0}^{n} x_j^2)^2}}, \tag{1}$$

where $x^1$ and $x^2$ represent the two datasets the correlation coefficient is calculated for, $n$ is the number of elements in $x^1$ and $x^2$, and $x_i^1$, $x_j^1$, $x_i^2$, $x_j^2$ represent the individual elements of the datasets $x^1$ and $x^2$. The highest correlations to CR are related to $CO_2$, while there is very little correlation between BR and WSA with CR. DR has a negative correlation with CR, with the value of -0.15. Other correlation values are positive and larger than 0.1.

### *2.2 RF application for feature importance*

The RF algorithm works by generating decision trees. Each node of a tree contains a condition where a decision is made to go down a level to the following node. This process is repeated until a leaf of a tree is reached, where a solution value (in our example the predicted CR) is reached. RF is a so-called ensemble method, which generates a large number of individual decision trees and bases the output of the model on the combination. For this paper, the number of trees generated within the method is 1000, with no limit to their depth or width, and other hyperparameters such as a decision to split a node, or a minimum set of samples per node are set to automatic [8]. There are two ways

of determining the feature importance: mean decrease in impurity (MDI) and feature permutation (FP). Impurity is defined as the measure of homogeneity of the labels at each node. The impurity measure used in the presented research is variance, which is the standard regression impurity measure. If the $N$ is the number of instances, and $y_i$ is the individual regression label value, impurity $I$ is defined as [8]:

$$I = \frac{1}{N}\sum_{i=0}^{N}(y_i - \frac{1}{N}\sum_{i=0}^{N}y_i)^2. \tag{2}$$

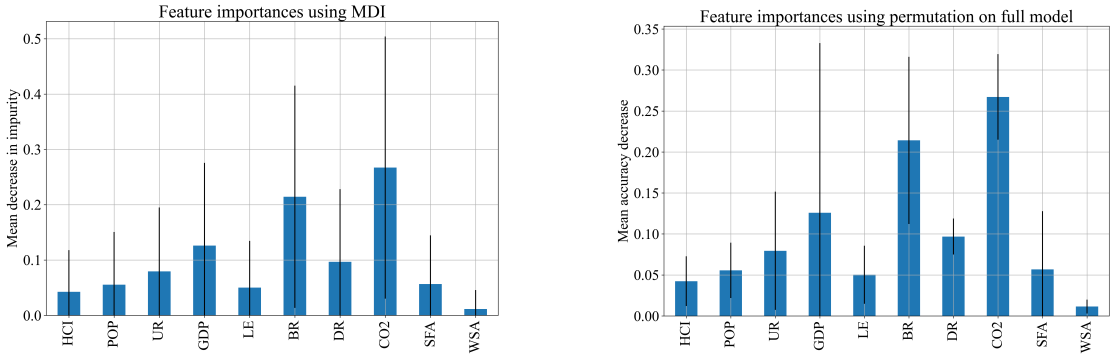
Impurity is an important metric as it defines the information gain (IG) between two levels of the tree. The IG allows us to determine how much knowledge is obtained inside the model due to the separation being made. If we assume the dataset $D$, of size $N$, which is split due to the decision $s$ into $D_{left}$ of size $N_{left}$ and $D_{right}$ of size $N_{right}$, the $IG$ is calculated with [8]:

$$IG(D,s) = I(D) - \frac{N_{left}}{N}I(D_{left}) - \frac{N_{right}}{N}I(D_{right}) \tag{3}$$

Using MDI can be misleading when high cardinality of features (many unique values) are present, such as is the case in the used dataset. For this reason, we shall also introduce the FP feature importance. The FP algorithm works by first training the algorithm with the given dataset and obtaining a score. Then, the training process is repeated $N_f$ times, where $N_f$ is the number of features. In each iteration, one of the features is randomly shuffled. The difference in regression scores between the score on the original dataset and the one with a feature shuffled defines the importance of the feature, where the higher score difference indicates a more important feature [8]. Both of the feature metrics have been used in the presented research to allow for the comparison and cross-validation between the collected results.

## 3 Results and discussion

Figure 1 demonstrates the results for feature importances determined with RF algorithm, using MDI approach in subfigure 1a, and FP approach in subfigure 1b. Besides somewhat larger importance given to individual features when using FP, the results between the two methods are similar. The most important features are shown to be $CO_2$, BR, and GDP.




(a) $MDI$ (b) $FP$

Fig. 1: Feature importance derived from RF algorithm, using (a) MDI, and (b) FP approaches.

The importance of $CO_2$ emissions shown in the initial correlation analysis has been confirmed by the RF feature importance analysis. Still, the higher importance of other features - such as BR and GDP is only revealed when observing the RF results. It should be noted that there is a high deviation in the results. This is most probably caused by a relatively small dataset size, and this is the reason why various regressors in the RF ensemble achieve varied results.

## 4 Conclusions

The authors designed a dataset for 42 countries and trained an RF ensemble model, using it to determine features. The most important factors that influence CR, determined using this approach are $CO_2$ emissions, BR, and the GDP of the country. The data shows the importance of BR which wasn't present in the correlation analysis, showing the quality of the RF feature importance analysis. Some of the results vary, but this can be remedied by using a larger data set in future research.

## Acknowledgments

This research has been (partly) supported by the CEEPUS network CIII-HR-0108, European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS), project CEKOM under the grant KK.01.2.2.03.0004, Erasmus+ project WICT under the grant 2021-1-HR01-KA220-HED-000031177 and University of Rijeka scientific grant uniri-tehnic-18-275-1447.

## References

[1] Lorencin, I et al., On urinary bladder cancer diagnosis: utilization of deep convolutional generative adversarial networks for data augmentation. Biology 10.3:175, 2021.

[2] Baressi Šegota, S et al., Semantic segmentation of urinary bladder cancer masses from CT images: a transfer learning approach. Biology 10:11:1134, 2021.

[3] Musulin, J et al. An enhanced histopathology analysis: An ai-based system for multiclass grading of oral squamous cell carcinoma and segmenting of epithelial and stromal tissue. Cancers 13.8: 1784, 2021.

[4] Qi, Y Random Forest for bioinformatics. Ensemble machine learning, Springer, 2012.

[5] Blagojević, A et al. Artificial intelligence approach towards assessment of condition of COVID-19 patients - identification of predictive biomarkers associated with severity of clinical condition and disease progression. Computers in biology and medicine 138:104869, 2021.

[6] WCRF International. Global cancer data by country. Available at: , 2022.

[7] Our World in Data. "Our World in Data". Available at: , 2022.

[8] Pedregosa, F et al. Scikit-learn: machine learning in python. The journal of machine learning research 12: 2825:2830, 2011.