**Motivation**

From “WHO methods and data sources for country-level causes of death 2000-2019” published in December 2020, WHO provided the Global Health Estimates (GHE) data for years 2000, 2010, 2015 and 2019 for selected regional groupings of countries, areas and territories. As great changes and developments took place in various aspects, we are interested in studying the trends and determinants of life expectancy over the past twenty years for each continent. Specifically, we are trying to find the dominant factors, including years, genders and categories of diseases, that affect life expectancy, and how these categories have changed over different continents. Our results aim to reflect a global-level trend for essential factors for life expectancy and potentially provide insightful inferences for focus of future health research and service.

**Data**

The original (uncleaned) data used was obtained from the [Kaggle posting](https://www.kaggle.com/datasets/adamsmith852/life-expectancy-data-gho) which contains combined data from various World Health Organization datasets, including life expectancy scores at birth, mean BMI, alcohol consumption per capita, and mortality and global health estimates by country. All links are in the appendix.

While investigating the life expectancy dataset, we found that it had 147 columns. In order to explore the dataset more efficiently, we split the columns into groups of 25 which resulted in some columns that were related remaining grouped. Next, we made correlation plots and pairplots to see how correlated the columns were and found that there were groupings of highly correlated columns. We investigated this further and found a document for the WHO data and found that there were larger categories we could group the columns into. For example, a bunch of features that related to different kinds of lymphomas and cancers were highly correlated.



In total, the original data had 147 columns or features. Out of the 147, we chose “Life expectancy at birth” as our dependent variable.

In the original dataset, a total of 420 values were found to be missing. Out of these, 384 values were missing for "BMI", 20 for "Life expectancy at birth", and 16 for "Country". Since our dataset has 1464 rows, it would not be prudent to drop all the missing values, which constitute approximately 27% of the entire dataset. As most of the missing data pertains to "BMI", we decided to replace the missing "BMI" values using the average BMI values for the corresponding country and gender, which helped reduce the number of missing values from 420 to 60. Finally, we removed the remaining 60 missing values.

Due to the high dimensionality of the data, which would be increased if we one-hot-encoded or dummy-coded our 2 categorical variables (country, gender), we decided to modify the original dataset.

We categorized the features into larger categories, and added the sum of deaths per capita 1000 for each feature based on the categorizations provided by WHO on page 41 of their [global health estimates document](https://cdn.who.int/media/docs/default-source/gho-documents/global-health-estimates/ghe2019_cod_methods.pdf?sfvrsn=37bcfacc_5). The categories include: infectious and parasitic diseases (instead of tuberculosis, STDs excluding HIV, HIV/AIDS, etc.), respiratory infectious (instead of lower respiratory infectious, upper, etc.), and so on, such that we reduced 141 features to merely 22.

In addition, since “country” contained 183 unique values (i.e. 183 unique countries), and hence, a high dimensionality if one-hot-encoded, we decided to alter this column and convert the country column to continents. This further reduced our number of features in “country” (once dummy-encoded) from 183 to 6.

Our final altered dataset, once dummy-coded, contained 35 columns or features including our dependent variable. Our primary focus was to identify the top factors that affect life expectancy for each continent. To achieve this, we divided the altered dataset into six subsets based on their respective continents, namely Africa, Asia, Europe, North America, South America, and Oceania. Subsequently, we further divided each of these subsets into training and testing data using a 70/30 split, resulting in a total of 12 datasets (six for training and six for testing). Our training data comprise 997 rows in total, while our testing data includes 431 rows.

**Analytics Models**

OSR^2 for all models below:

|  | South  America | Europe | Africa | Asia | Oceania | North America |
| --- | --- | --- | --- | --- | --- | --- |
| Linear Model | -0.094742 | 0.40104 | 0.410436 | 0.40647 | -3.79033 | 0.4472618 |
| CART | 0.3256 | 0.5113 | 0.436257 | 0.54063 | 0.87839 | 0.44844 |
| Random Forest | 0.6955 | 0.705 | 0.743239 | 0.84263 | 0.91237 | 0.7773532 |

[**South America**] *- all feature significance scores in the appendix*

The OSR^2 for the linear model for South America turned out to be a negative number. This indicates how terrible the linear model is in fitting the test data. Compared to the R^2 value for the training data, we can see that the much lower value from the OSR^2 indicates that overfitting exists. So, the linear model is a bad model to use.

For the CART model, the model is still not good, but better than the linear model, for estimating the relationship between factors and life expectancy. This is indicated by the low OSR^2 value.

Finally, the Random Forest model looks to be the best model for estimating the relationship due to the OSR^2 being fairly high. According to the Random Forest model, the top factor for affecting life expectancy at birth the most is the disease titled neurological conditions. Followed closely to the top factor is musculoskeletal disease. As you can see, gender and BMI still play a role in being closely associated with life expectancy, but not as much as musculoskeletal disease and neurological conditions. We also have endocrine blood immune disorders and alcohol following after that which round out the top factors that affect life expectancy the most. The rest of the factors played a minor role.

**[Europe]***- all feature significance scores in the appendix*

Overall, we have trained three different models for analyzing the most important factors that affect life expectancy in Europe: the linear model, the CART model, and the RF model. Out of the three models we have trained, we decided to focus on the results given by the RF model as the RF model for Europe reached the highest OSR^2 on our test data. Specifically, the linear model for Europe had an OSR^2 of 0.4, the CART model achieved an OSR^2 of 0.51, and the RF model reached the highest OSR^2 of 0.705.

Looking at the relative feature importance plot for Europe given by the RF model, we discovered that the top five most important factors for Europe are: “alcohol”, “endocrine blood immune disorders”, “cardiovascular disease”, “oral conditions”, and “digestive diseases”. However, by taking a closer look at the importance score, we found that “alcohol” plays a main role among these top factors as the importance score for “alcohol” is significantly higher than other factors. Specifically, the actual feature importance score for “Alcohol” is 0.53 while it’s about or less than 0.05 for other factors.

**[Africa]** *- all feature significance scores in the appendix*

The linear model coefficients show that skin diseases have the highest weight on life expectancy, with coefficient 16.2557 and p-value 0.413. While the model’s R^2 is 0.9, the OSR^2 is 0.41 which implies overfitting. We also tried CART regression, which used BMI as the feature with highest importance, but the OSR^2 for this model is quite low, around 0.43. Finally, we tried an ensemble by using a random forest regressor. This increased our OSR^2 to 0.74 and also showed that the most important feature for predicting life expectancy in Africa was indeed BMI. Both the CART decision tree and random forest had the same top factors but the random forest had better OSR^2 overall.

The top 5 factors for Africa using the Random Forest Regressor were BMI, infectious and parasitic diseases, nutritional deficiencies, malignant neoplasms, and alcohol. The size of the dataset especially after splitting the data by continent could be limiting so in the future specifically searching or collecting life expectancy data for each continent could improve the accuracy of our results.

**[Asia]** *- all feature significance scores in the appendix*

The linear model for Asia shows a particularly high impact of sudden\_infant\_death\_syndrome (coefficient: -16.5601, p-value: ​​0.001) on life expectancy, which aligns with common sense - infants have much shorter length of life that pulls down the overall result. Otherwise, most coefficients are close to 0, with an R^2 of 0.600 and an OSR^2 of 0.406.

We then consider switching to trees for better reflection of non-linearity. The CART model achieves an OSR^2 of 0.541, and ranks the top 5 important factors as musculoskeletal, neonatal conditions, infectious and parasitic diseases, neurological conditions and the year 2000.

The RF model achieves the highest OSR^2 of 0.843, but outputs the 5 most important factors similarly as CART, though involving no year: neonatal conditions, infectious and parasitic diseases, musculoskeletal, neurological conditions, and BMI. Considering the size of data in Asia, RF returns a reasonable feature importance ranking and outperforms others in terms of OSR^2, thus is a competitive choice for analytical modeling.

**[Oceania]** *- all feature significance scores in the appendix*

The linear regression model had an OSR^2 of -3.790, indicating that the model was extremely overfit and that the model was generally bad at predicting on new data. As such, we will not consider the linear regression model.

Comparing the CART regression model to the Random Forest model, results were very similar. The OSR^2 of random forests were slightly higher at ~0.912, while the CART tree was ~.878, though the MAE of random forests was around ~0.023 higher than the CART tree.

Despite these minor differences, there were some major differences in the models. The top factors under the CART tree were endocrine\_blood\_immune\_disorders, Infectious\_and\_parasitic\_diseases, and genitourinary\_diseases, which had feature significance scores of 0.6824, 0.1199, and 0.0880 respectively. Under the random forest model, top factors were endocrine\_blood\_immune\_disorders, neurological\_conditions, malignant\_neoplasms, musculoskeletal, and Infectious\_and\_parasitic\_diseases with scores of 0.2797, 0.2030, 0.1518, 0.1081, and 0.0694 respectively. For both models, the rest of the features were either insignificant or had a score of <0.06.

Comparing the 2 models, it is clear that the top factors across both models were endocrine\_blood\_immune\_disorders and Infectious\_and\_parasitic\_diseases, though I would also argue that musculoskeletal and genitourinary\_diseases are also somewhat important, given that these factors have above 0.1 feature significance in both models.

**[North America]** *- all feature significance scores in the appendix*

The linear regression model for North America had a high R^2 of 0.9 but an OSR^2 of 0.4473, indicative of overfitting.

So we moved on to use a CART decision tree regressor and achieved a slightly higher OSR^2 of 0.4484. This model ranks the most important features as unintentional injuries followed by gender, malignant neoplasms, respiratory infections, and digestive diseases.

Under the Random Forest model, we achieved the highest OSR^2 of 0.77735. This model ranks the most important features as unintentional injuries followed by BMI, gender, alcohol and neonatal conditions.

Between the CART and Random Forest model, there is a significant difference in the OSR^2 and some differences in the rankings. Both models agree that the most important factor in determining life expectancy at birth is unintentional injuries. In both models, it seems like gender does play an important factor as it is ranked second and third in CART and Random Forest respectively. They disagree on the other top factors but it may be interesting to note that unintentional injuries is the most relatively important feature in both models, indicating that life expectancy in North America is largely impacted by the number of unintentional injuries inflicted.

**Impact**

Our study aimed to identify the key factors, such as BMI, alcohol consumption, and different disease categories, that have the most significant impact on life expectancy for six different continents. Our findings indicate that while some diseases have a considerable impact on life expectancy across all continents, there are also other factors and diseases that are unique to specific regions. By understanding these variations, policymakers and public health practitioners can prioritize their efforts and allocate resources more effectively towards interventions that target the most pressing and the most serious health issues in their countries. Furthermore, our study can inform and empower the general public to make informed decisions about their health and lifestyle choices by increasing their awareness of the factors that have the most significant impact on their health and life. Overall, our analysis has the potential to contribute to a global effort towards improving health outcomes and life expectancy for diverse regions and populations.

In the future, it could also be interesting to expand the scope of our analysis to consider factors that affect life expectancy estimates across different years, such that we can compare factors that impact life expectancy in Oceania in 2015 versus in 2019, for example. This analysis would help us better understand the general trends in life expectancy across different years, and help us better understand what factors may become more important in future test data (future years).

The impact of our model may vary depending on our subpopulation of interest, such as country of origin. While the data from the WHO is gathered from countries around the world, we ended up grouping the countries by continent for project purposes. However it is important to note that each of the countries within the continent may be drastically different from each other. Using Asia as an example, the factors that affect life expectancy in China may differ from that of Thailand, perhaps due to differences in the economy, culture, environment, etc. This can affect the impact of our model because it may overlook important differences between countries that are relevant to our discussion regarding factors that determine life expectancy.

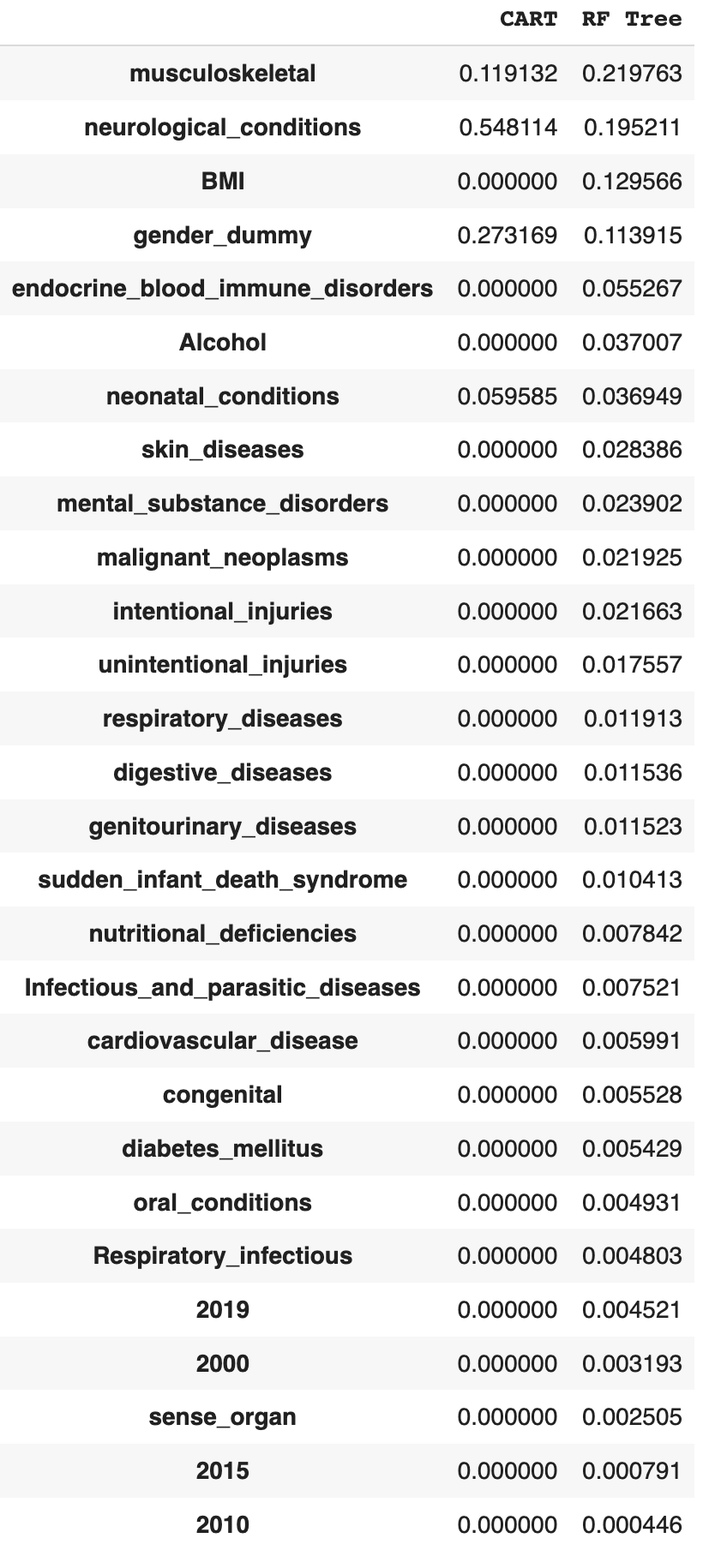
**Appendix/Works Cited**

Data Sources

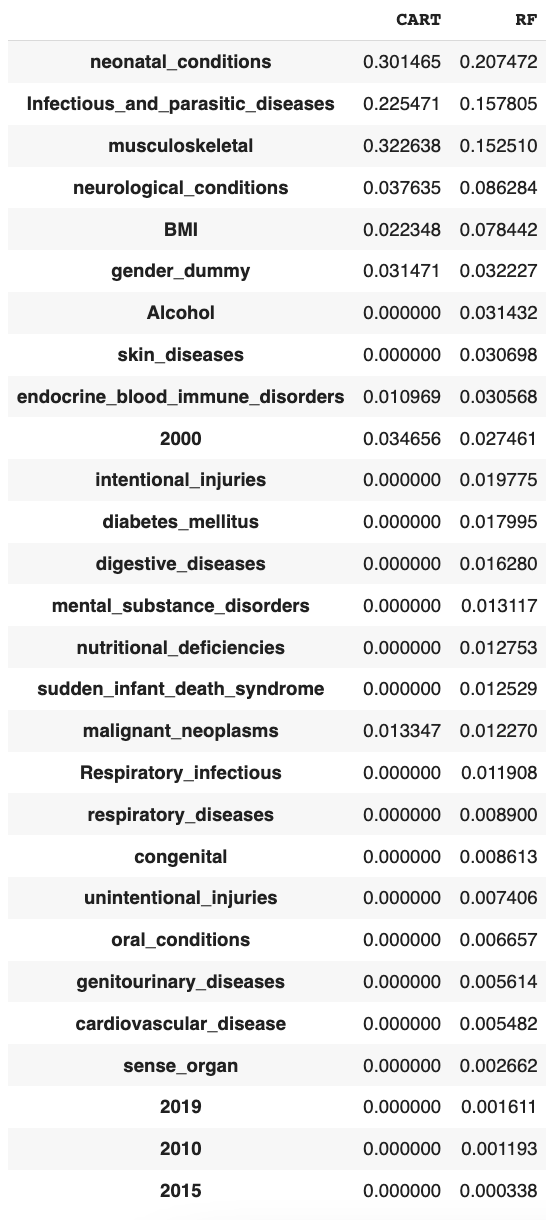
* Life Expectancy at Birth
  + <https://www.who.int/data/gho/data/indicators/indicator-details/GHO/life-expectancy-at-birth-(years)>
* Mean BMI (kg/m²) (crude estimate):
  + <https://www.who.int/data/gho/data/indicators/indicator-details/GHO/mean-bmi-(kg-m-)-(crude-estimate)>
* Alcohol, total per capita (15+) consumption (in litres of pure alcohol):
  + <https://www.who.int/data/gho/data/indicators/indicator-details/GHO/total-(recorded-unrecorded)-alcohol-per-capita-(15-)-consumption>
* The rest of the factors:
  + <https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death>

CART and Random Forest Feature Importance Tables

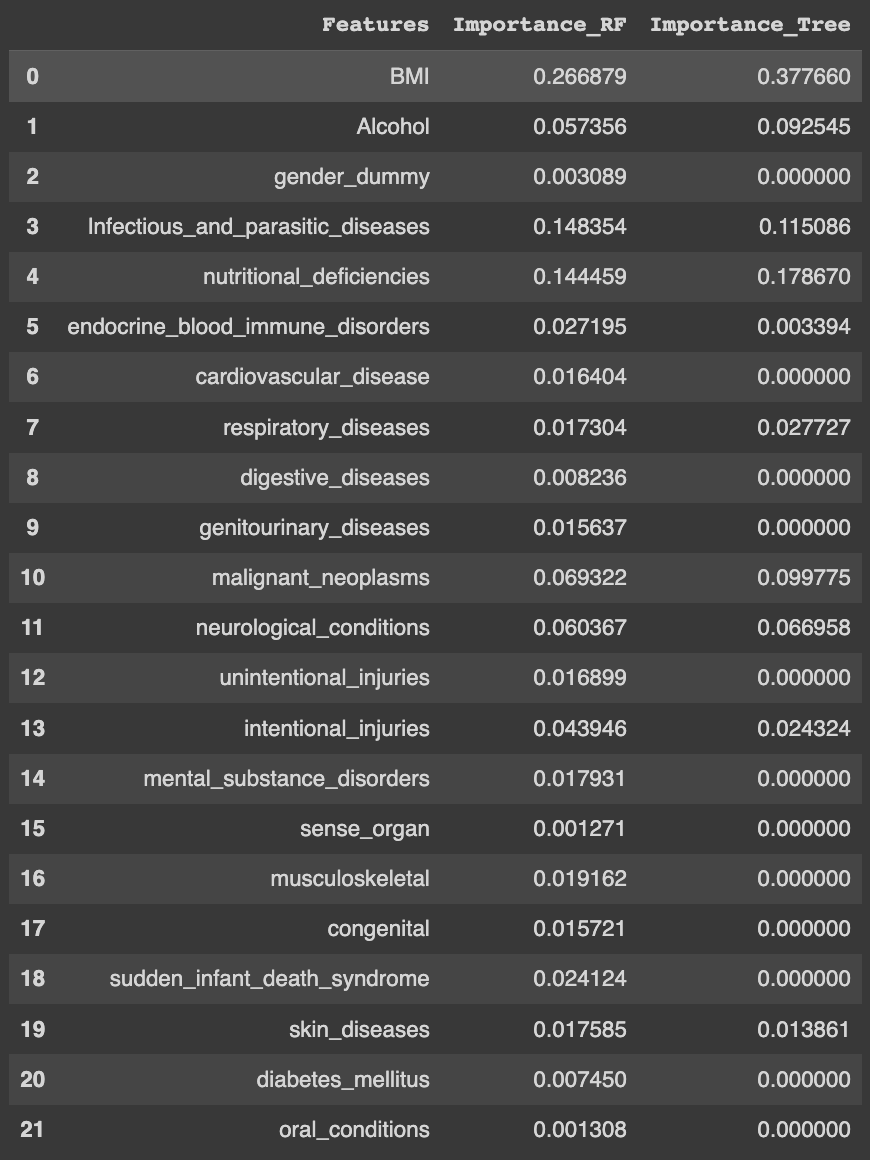
South America:



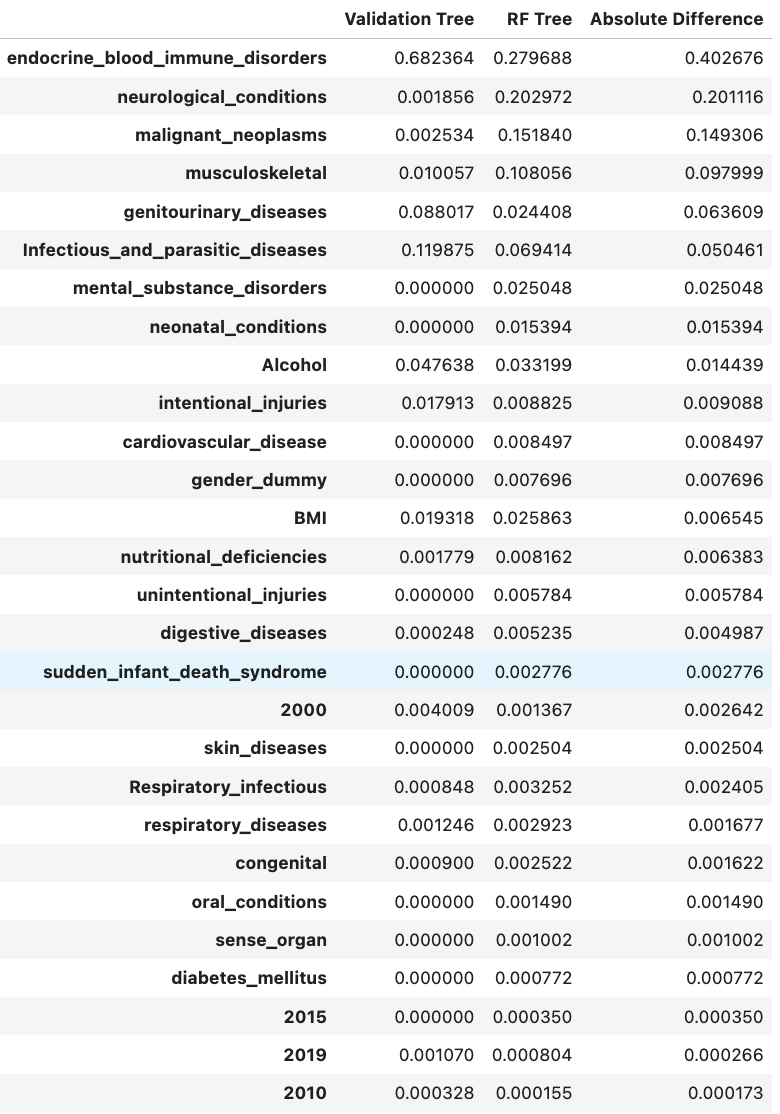
Asia:



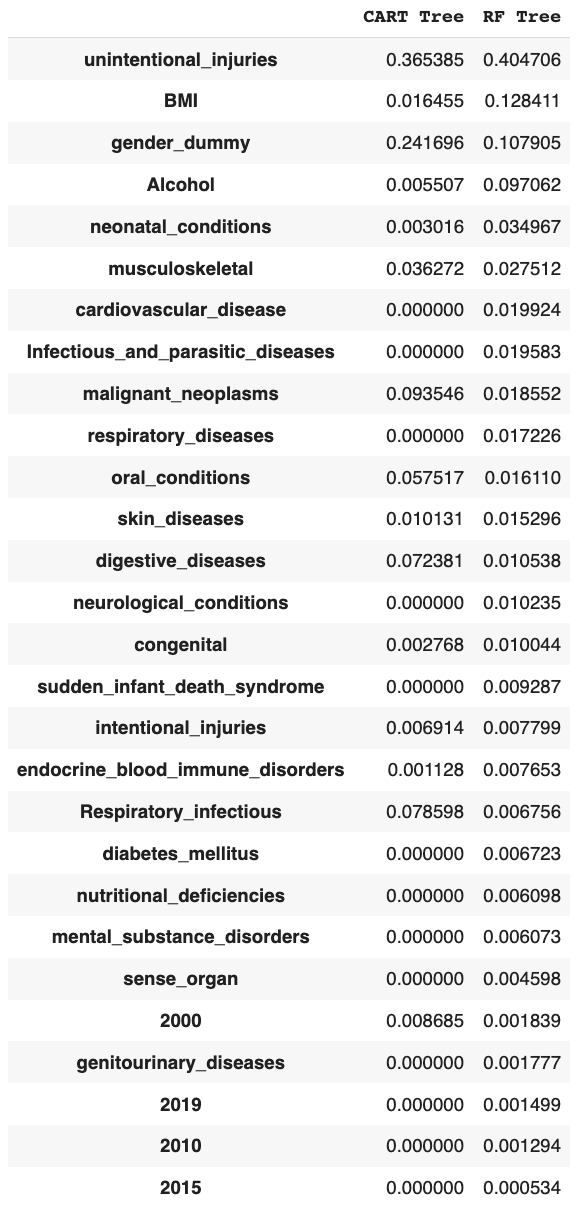
Africa:



Oceania:



North America:

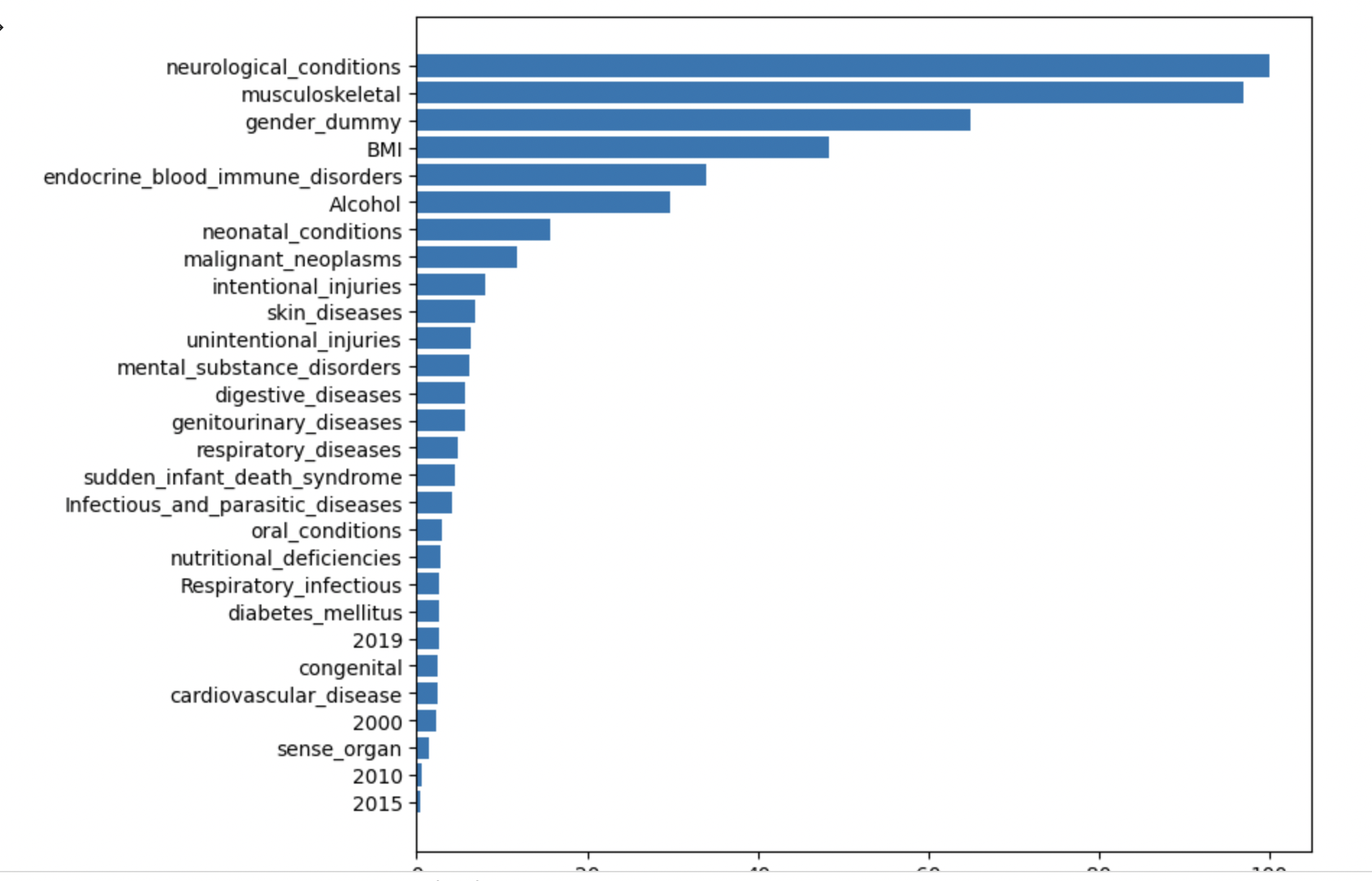


Europe:

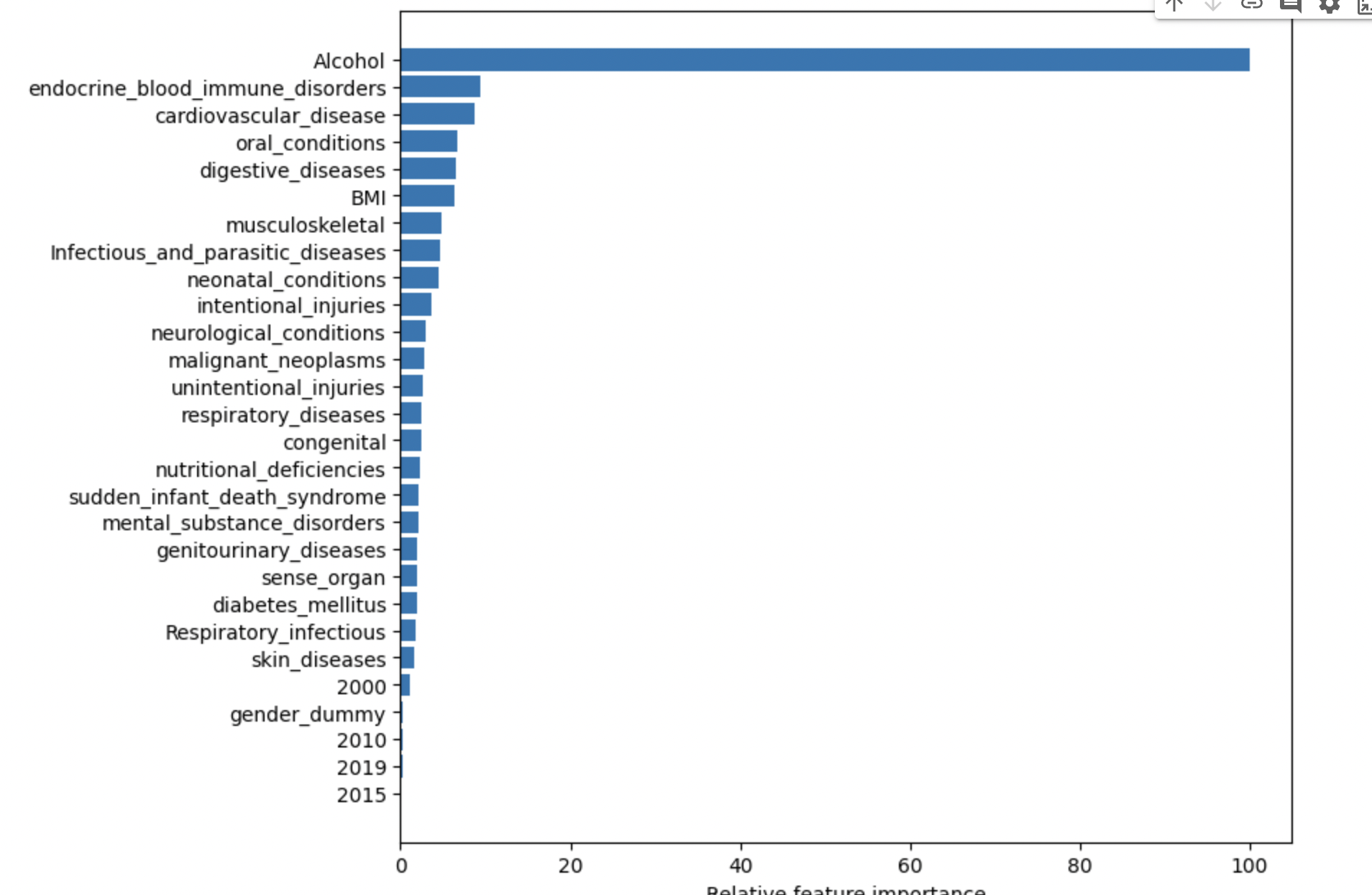


Random Forest Relative Importance Plots

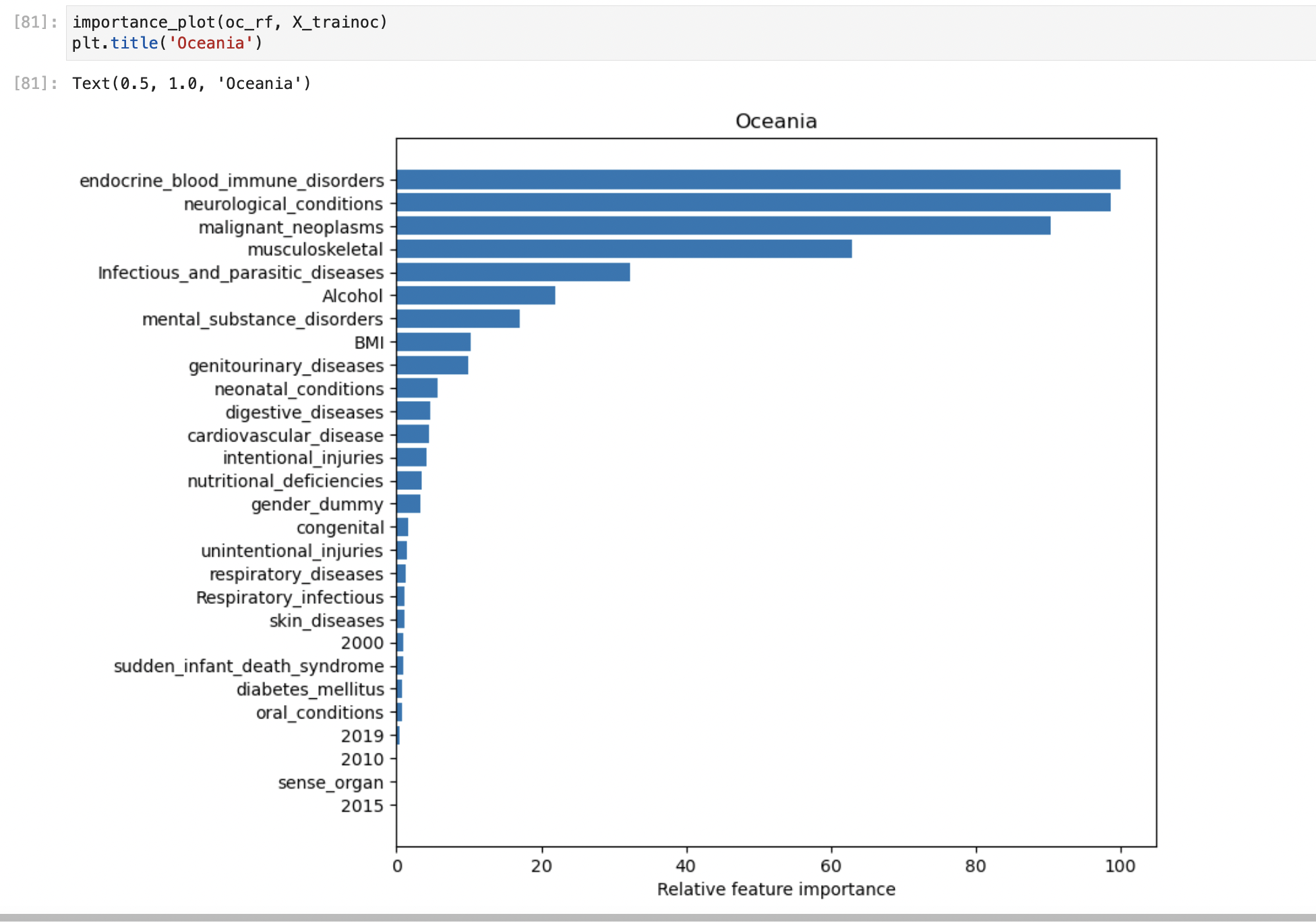
South America: (OSR2: 0.6955008880964537)



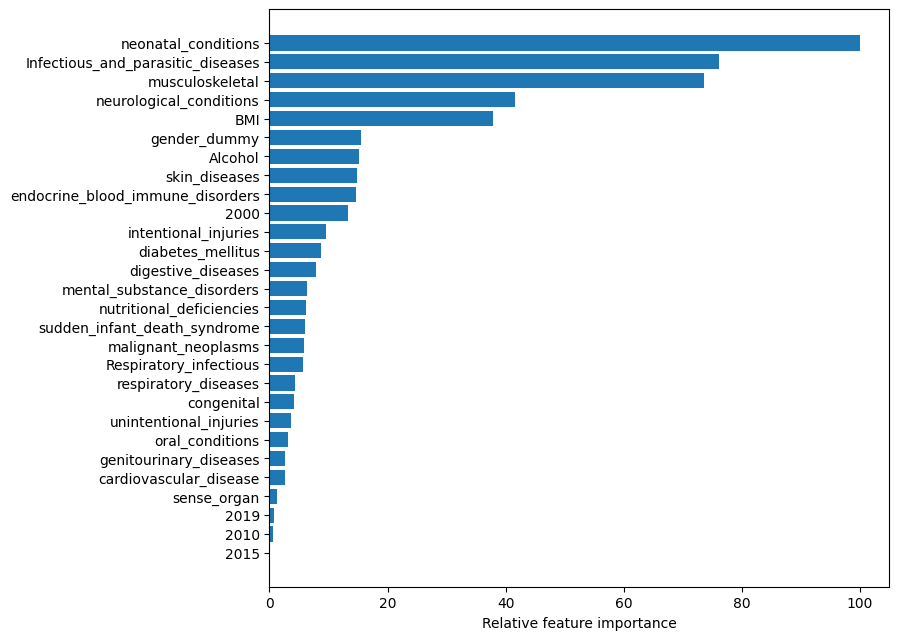
Europe: (OSR2: 0.6962466797844868)



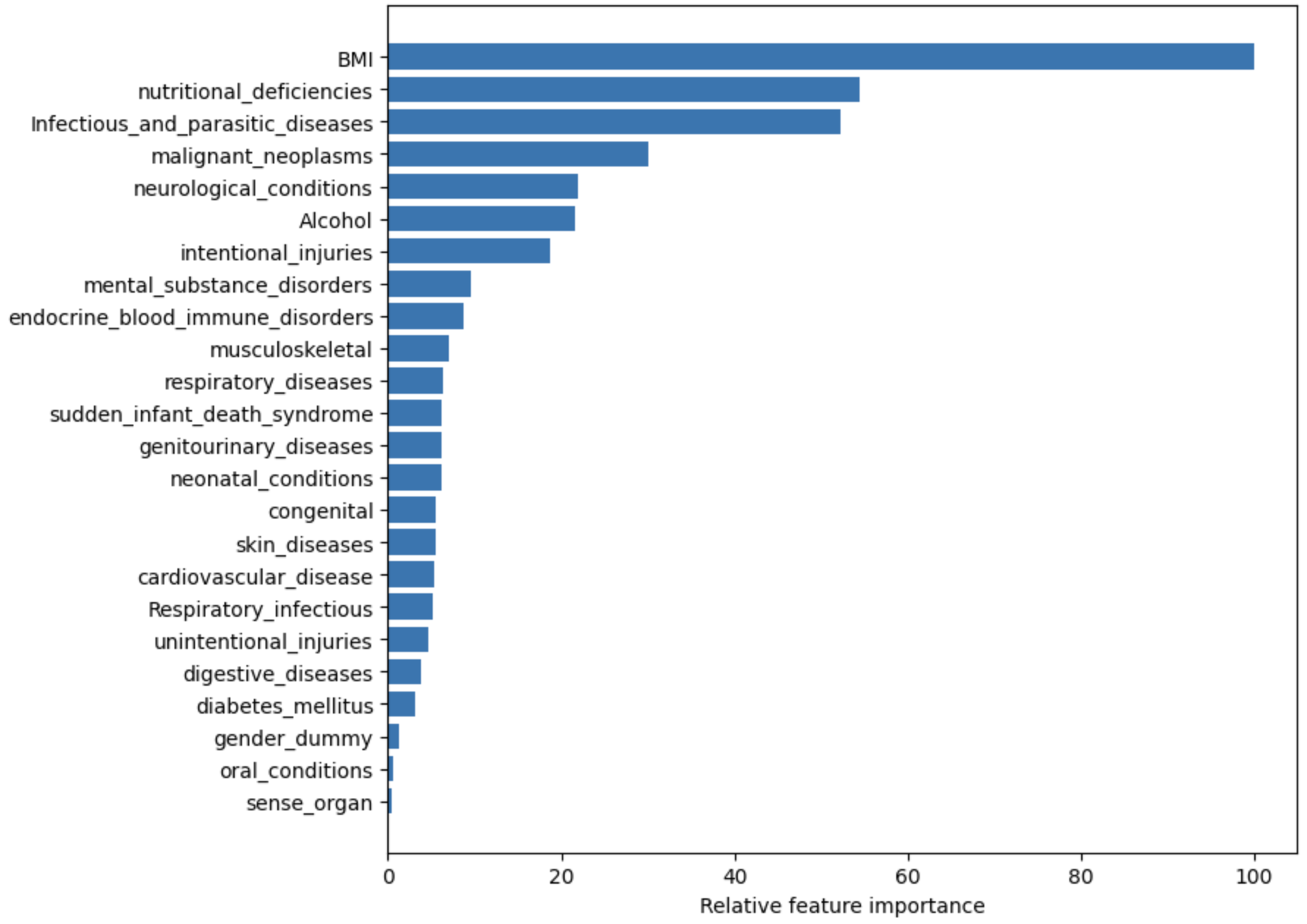
Oceania: (OSR2: .906)



Asia: (OSR2: 0.8426283173297445)



Africa: (OSR2: 0.7488290262832173)



North America: (OSR2: 0.7773532)

