Projekt z předmětu Strojové učení v Pythonu

1st Elia Pavliš
Ústav fyzikální chemie
Vysoká škola chemicko-technologická v Praze
Prague, Czech Republic
pavlisol@vscht.cz

2nd Sofie Meyer
Ústav počítačové a řídicí techniky
Vysoká škola chemicko-technologická v Praze
Prague, Czech Republic
dubovoyy@vscht.cz

***Abstract*—This study looks into a possible relationship between cause of death and quality of life (modelled by GDP per capita and air pollution). A promising relationship between certain types of fatalities and GDP was found.**

***Index Terms*—machine learning, regression models, classification, cause of death, disease, GDP, air pollution, healthcare**

I. Introduction

Globally, cardiovascular disease and cancers are the two leading causes of death. Other causes such as diarrheal infectious diseases, HIV/AIDS, malaria or pneumonia make up a much smaller portion. This is very well illustrated by figure 1. This figure is, however, merely an average and the distribution can vary wildly between countries. For example, while malaria is not a concern in most of the world, the WHO African Region is disproportionately affected, accounting for as much as 94 % of global cases. [Org]

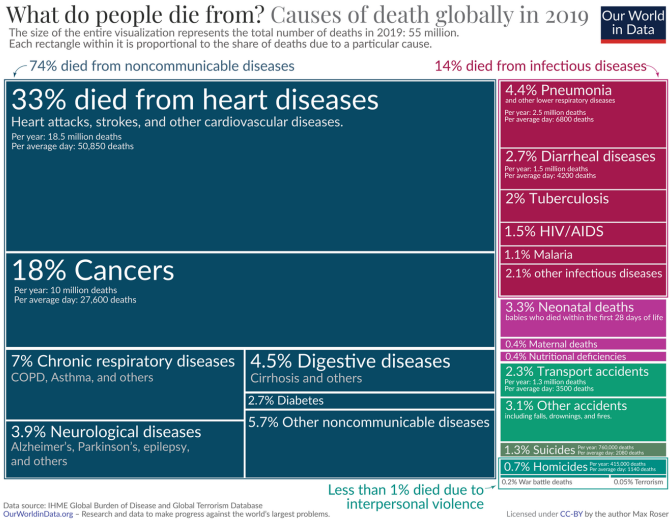

Fig. 1. Visualisation of global causes of death in 2019, published by Max Roser on OurWorldInData.org.

Our goal is to investigate the effect of economic and environmental factors on the distribution of deaths from select causes in world countries using machine learning and statistical methods. If such a link is found, machine learning methods could potentially be employed to help prepare a country's healthcare system for new challenges tied to those factors, such as rapid economic growth or a major change in environmental policy.

II. Dataset and Preprocessing Description

All datasets can be found on kaggle, an extensive online repository and community hub for machine learning. Given the limited scope of the research, it was necessary to keep methodology as simple as possible. Therefore, only the cause of death data itself and two very simple predictors were looked at.

Global cause of death data was taken from the following dataset. It is important to note that some countries lack a proper administration system for such data which leads to very uncertain estimates. Larger, multi-country regions were removed from the dataset to avoid duplicity.

As a representation of the economical situation, we chose the **GDP per capita** of each country. This choice has one major shortcoming – while it is an important indicator of economic activity, it does not reliably capture the well-being and quality of life of individuals. However, we believe it can provide a rough estimate of whether a given country could be considered "rich" (*i.e.* most people are not threatened by poverty, overall more likely to have a good quality and accessible healthcare system) or "poor".

Air pollution data was used as a rudimentary indicator of overall environment quality/safety. A possible link to respiratory fatalities was also investigated.

First of all, the datasets needed to be combined into one. The resulting data frame consisted of one row per country and year. The columns contained the number of fatalities from a given cause, GDP per capita and air pollution index. Only the years 2010-2017 were selected because global air pollution data was not available outside that timeframe. The entire final dataset was checked for non-numeric/missing data and zero GDP values.

For classification purposes, GDP and Air Pollution were sorted into 4 equal size groups ranging from `Low` to `Very High`. Total fatalities by cause and GDP Group were visualised (see fig. 2). To make interpretation easier, some causes were not included in the final dataset. The following columns were removed:

- Forces of nature, Environmental exposure, Conflict – not suitable because they are mostly determined by geography (war zones, extreme weather)

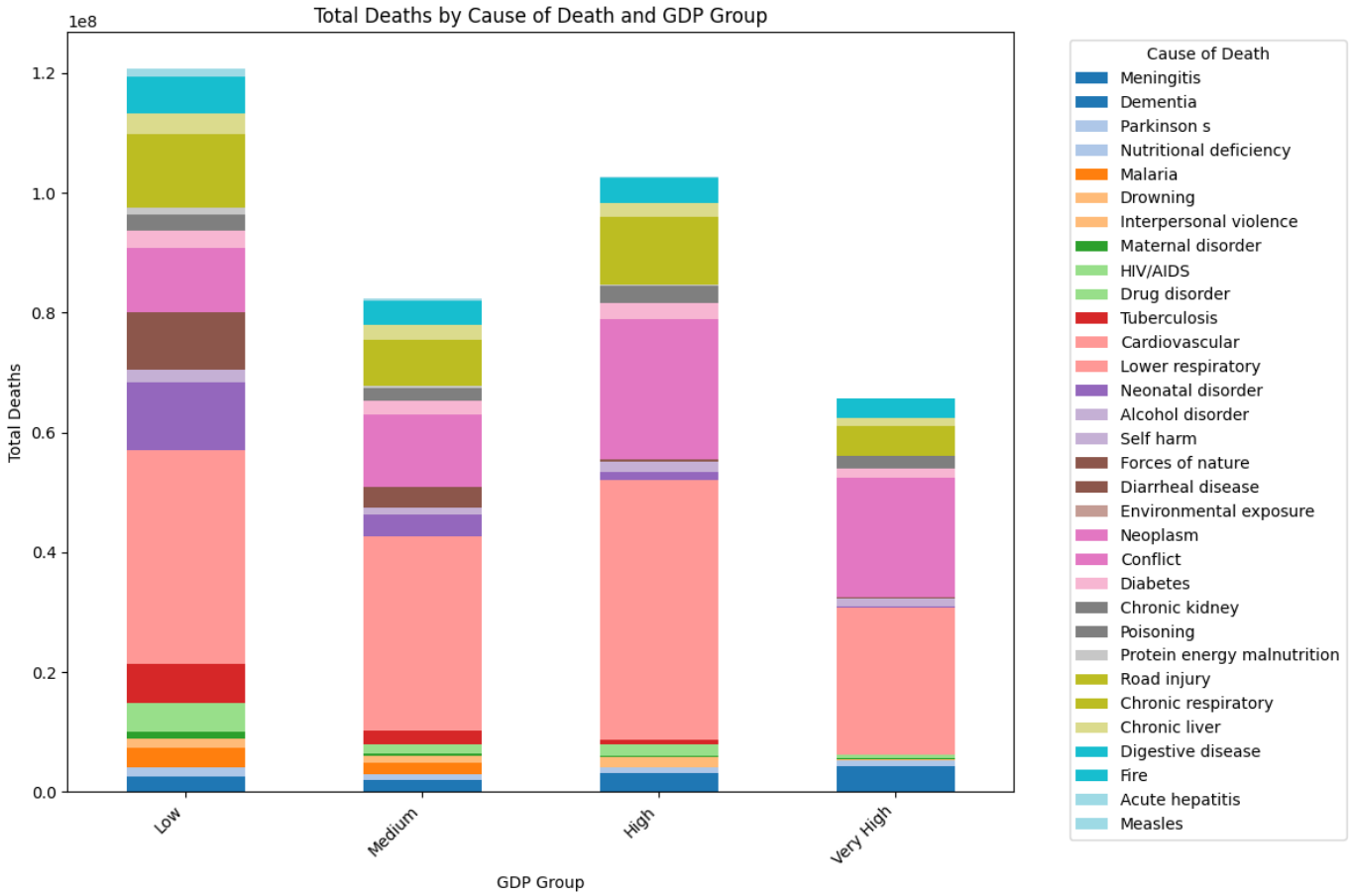

Fig. 2. Total number of fatalities in each GDP Group colored by cause of death. The proportions of different causes of death in each GDP group are visibly different – notably Malaria (dark orange), Neoplasm (magenta), Tuberculosis (red) or HIV/AIDS (light green).

- Drowning, Poisoning, Road injury, Fire – accidents and injuries that do not make up a significant portion of total fatalities and are not likely to be related to any of the factors studied
- Drug disorder, Alcohol disorder, Self harm, Interpersonal violence – these fatalities result from a very complex combination of social factors so we do not consider them appropriate for our simple model
- Meningitis, Maternal disorder, Diabetes, Chronic kidney, Chronic liver, Acute hepatitis, Measles – these make up a very small proportion of total fatalities (such as Measles) or have too much potential overlap with other conditions/too many potential causes (*e.g.* diabetes and cardiovascular disease).

The Shapiro test was performed for all columns (except Year where checking for normality does not make sense). None of the columns were even close to a normal distribution. However, some methods call for an appropriately transformed dataset because they work better with normally distributed data. The commonly used Box-Cox transformation could not be applied here because it requires strictly positive values. Therefore, the Yeo-Johnson transformation was applied and the Shapiro test was performed again on the transformed dataset. While the data still could not be considered normally distributed, the distributions were closer to normal (and not much else could be done at this point).

As an initial overview, the correlation matrix of the untransformed dataset (fig. 3) was visualised and studied. It was immediately clear that the year of collection had little impact on the other variables. Therefore, the `Year` column was removed and a new reduced dataset was formed by replacing the individual entries for each country by their mean. The correlation matrix also revealed some significant relationships between a) certain causes of death, b) GDP and air pollution, c) GDP/air pollution and certain causes of death. Those relationships served as the basis for the models discussed in section III.

A. PCA

1) General theory: For achieving dimensionality reduction and better interpretability, **Principal Component Analysis (PCA)** was applied on the cleaned and combined dataset. Dimensionality reduction is a crucial preprocessing step in machine learning because it can increase learning accuracy and improve result interpretability by removing irrelevant and redundant data[Li+18]. It achieves this by generating new uncor-

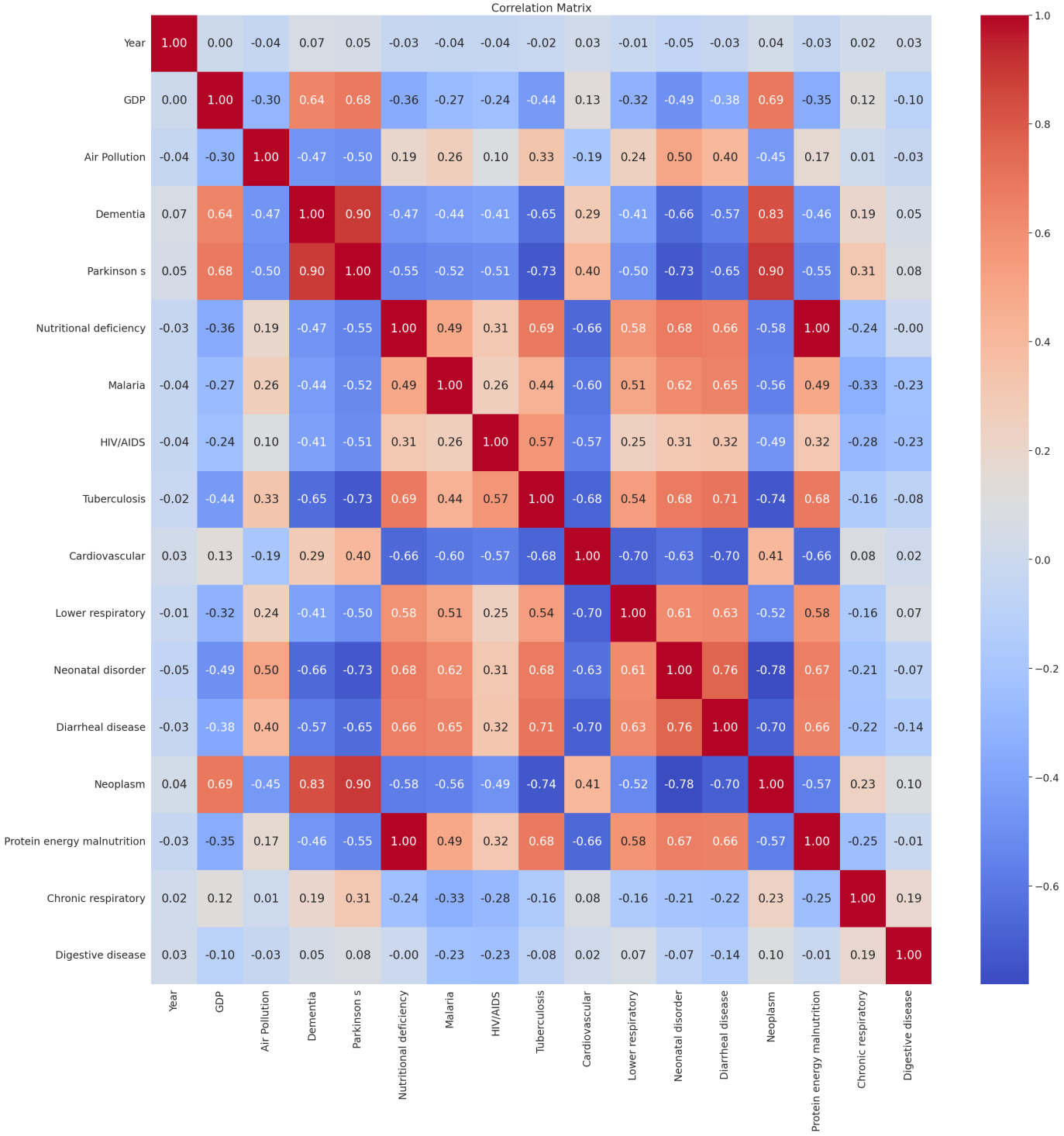

Fig. 3. Correlation matrix heatmap.

related variables, each of which is a linear combination of the original data. Identifying these new variables, known as principal components, involves solving an eigenvalue/eigenvector problem. Each principal component is chosen in a way that maximises variance in a given direction. Since these variables are defined by the dataset rather than predetermined, PCA is considered an adaptive data analysis technique[KKN14]. Standardization was implemented to improve the performance and effectiveness of PCA results.

2) Why PCA?: In order to pre-assess whether there even is a relationship worth studying between GDP (or Air Pollution, futher discussed as the *dependent variable*) and cause of death variables (further discussed as *features*), we need to visualise the dataset. In a scatter plot of two variables, this can easily be achieved via coloring the data points by the value of the dependent variable – if there is enough correlation between the dependent variable and one or both of the features, same colored points will visibly cluster. However, the number of features in our dataset is too high for this method to be applicable. At the cost of losing some information, plotting principal components instead of the original variables (figure 5) offers an easy solution. Since principal components are linear combinations of the original variables, any relationships between PCs and the dependent variable can in principle be traced back to individual features.

In summary, a PCA based visualisation simplifies interpretation in two ways:

- If the PC scatter plot shows visible clusters, there is a high chance that a clustering algorithm could also perform well on the untransformed dataset. Studying the composition of the PCs can also suggest which features separate the data into distinct groups.
- If the scatter plot shows clusters of the same color, it implies the existence of a relationship between the dependent variable and the features.

3) Parameters: Whenever PCA is used, it is necessary to set the number of principal components used for further analysis. The final number of components has to meet two criteria: **sufficient dimension reduction** (*i.e.* as few variables as possible) and **sufficient cumulative explained variance** (the cut off value usually ranges from 70 to 90 percent). Explained variance is a statistical measure of how much variation in a dataset can be attributed to each of the principal components (eigenvectors) generated by PCA. In essence, it indicates how much of the total variance is ”explained” by each component. In this case, the explained variance plots (figure 4) show that the variance in our dataset is explained mainly by the first (59.58 percent) and second component (12.95 percent). In other words, the first two components explain 72.53 percent of the total variance. As for the dimension reduction aspect, two components would be ideal. However, we wanted to make sure we did not miss anything of importance in the data. Therefore, we aimed for an explained variance of roughly 80 percent and also included the third principal component in our analysis.

The principal components were then plotted against each other and colored by **GDP Group** (figure 5) and **Air Pollution Group** (figure 6).

Coloring by **Air Pollution Group** did not show anything of importance. For **GDP Group**, the scatter plots show that the first principal component (PC1) explains the difference between the lowest and highest GDP group. However, the Medium and High groups are not very well separated. To some extent, PC2 also shows this relationship, but it is not as pronounced as in the case of PC1. The third principal component (PC3) does not seem to correlate with GDP in any noticeable way. In the first two plots, the data forms visible clusters that also seem to be differentiated by GDP value. This suggests that there might indeed be a relationship between GDP and the other variables and a cluster analysis method could provide more information.

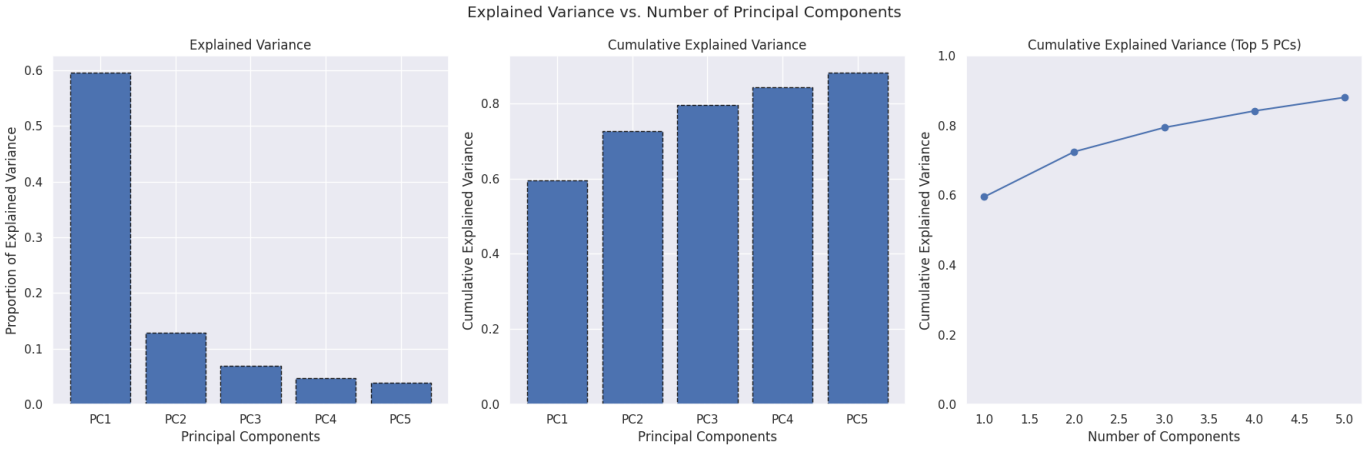

Fig. 4. PCA Explained Variance: individual (right) and cumulative (left).

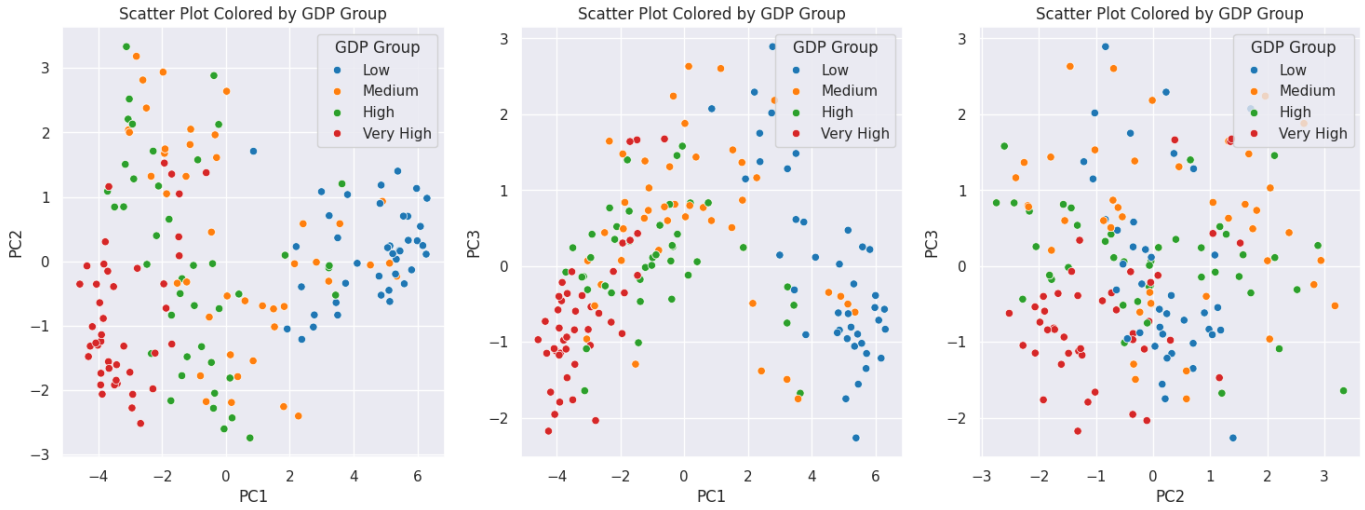

Fig. 5. Scatter Plot colored by GDP Group

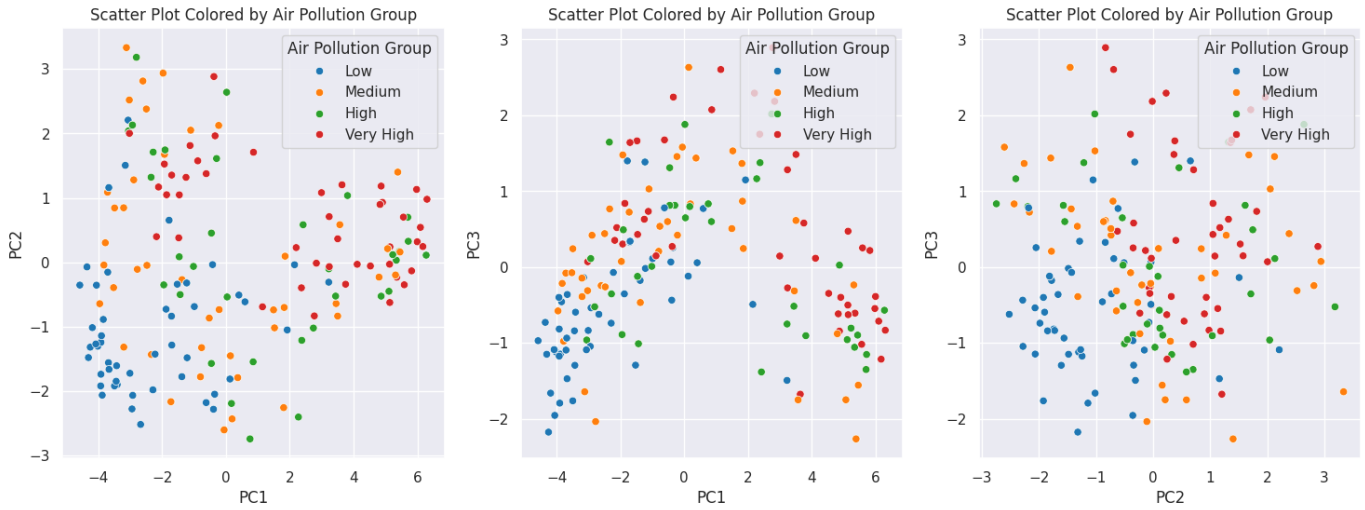

Fig. 6. Scatter Plot Colored by Air Pollution Group

III. Methods

In this work, 4 different machine learning methods were applied: K-Means clustering, Logistic Regression, Random Forest and Support Vector Classification. All methods were applied on the original variables, principal components were only used for visualisation purposes.

A. K-Means clustering

Cluster analysis is the formal study of methods and algorithms used to group, or cluster, objects based on their measured or perceived intrinsic characteristics or similarities. Unlike classification, cluster analysis does not rely on predefined category labels or class labels (unsupervised learning). [Jai10]

Different clustering algorithms are suited for different applications. Since our goal is first and foremost to identify groups of similar data points, we chose to apply the **K-Means** algorithm. This algorithm separates data points into a predefined number of clusters K. First, K centroids – points that are considered to be the centers of each clusters – are chosen. Clusters are formed based on minimizing the distance of all the points to their respective centroids. After that, the centroid value is updated and clusters are recalculated until the result no longer changes or the iteration limit is exceeded.

For our dataset, the PCA analysis suggests the optimal number of clusters to be 2 (more on this in section IV-A). To test this hypothesis, we looked at the silhouette scores (a higher score means better separated clusters) for numbers of clusters ranging from 2 to 15. The score for two clusters was indeed the highest (as shown in figure 7) so we proceeded to apply the K-Means algorithm with two clusters.

The other parameters of the scikit `KMeans` method were kept as their default values.

B. Random Forest

The **Random Forest algorithm**, introduced by Leo Breiman and Adele Cutler, constructs an ensemble of decision trees for classification tasks. This algorithm builds and combines multiple decision trees. During the construction of each individual tree, the Random Forest algorithm randomly samples the training data and randomly selects a subset of features for splitting nodes. This introduces two different levels of randomness. Once all decision trees are built, the overall classification result is derived through equal-weight voting among the trees [RCH17].

The algorithm for Random Forest is as follows (n, d and k are hyperparameters of the model):

1) **generate a bootstrap sample** by randomly selecting n samples (with replacement) from the training set
2) **construct decision trees** using the bootstrap sample:
   - randomly select d features
   - split the node using the feature that provides the best split according to an objective function, such as maximizing information gain.
3) **repeat** until the desired number of decision trees k is created
4) **combine the predictions** from each tree using a majority vote to determine the final class label [Ade+22]

In this work, the Random Forest algorithm was used to develop two classification models with GDP and Air Pollution Group as target variables. Prior to building the model, label encoding was used to assign numerical values to GDP Group and Air Pollution Group. The data was then split into training and testing sets, and a Random Forest Classifier was initialized for predicting GDP Group and Air Pollution Group separately. (Air Pollution was excluded from the GDP prediction data set and vice versa to ensure that the model would only be trained on cause of death data.) Different splitting criteria were tested with the Gini coefficient leading to best results in both cases.

The hyperparameters were set to the following values:

- number of trees: 100
- maximum depth: unlimited
- minimum samples at a leaf node: 1
- minimum samples required to split an internal node: 2
- maximum number of leaf nodes: unlimited
- split quality criterion: Gini impurity

C. Support Vector Machine

The **Support Vector Machine (SVM) classifier** is a renowned pattern classification technique, known for its excellent generalization performance even in the absence of domain-specific knowledge. This strong generalization capability is a key reason for choosing SVM as the classifier[Sha+14]. It is a specific classifier that operates on the margin-maximization principle, particularly effective with small sample sizes. It employs structural risk minimization, a concept introduced to machine learning by Vapnik, and has demonstrated exceptional generalization performance. For nonlinear problems, SVM uses a mapping function ϕ to transform the low-dimensional sample into a high-dimensional feature space. In this high-dimensional space, SVM constructs a separating hyper-plane that **maximizes the margin**[CD10].

The `'rbf` kernel was used. Different values of the hyperparameter *gamma* were tested and it was set to 0.1. All other

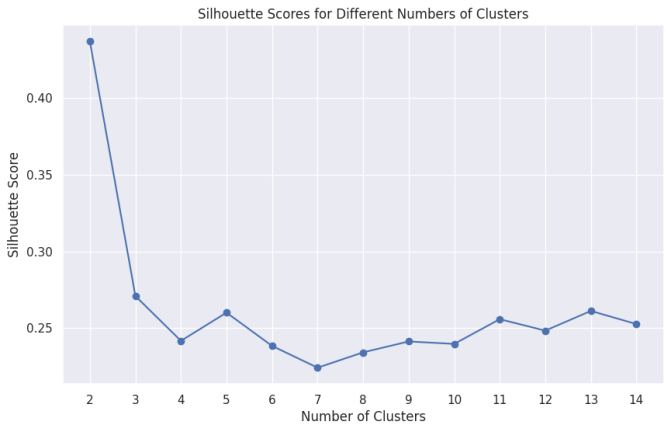

Fig. 7. Silhouette scores for different values of the number of clusters K.

parameters were unchanged from the `SVC` method's default settings.

D. Logistic Regression

Logistic regression is a classification model based on the probability of a binary (yes/no, A/B) outcome. Just like the other models, logistic regression was applied to predict GDP Group and Air Pollution based on cause of death data. Two approaches were tested: multinomial regression (a slightly altered algorithm used when data has more than two classes) and one-vs-rest regression (classic binary logistic regression where the outcomes considered are *in this class* versus *not in this class = in any of the other classes*). The confusion matrix and prediction accuracy were used to evaluate performance. The categories were encoded as integers ranging from 0 for Very Low to 3 for Very High.

Logistic regression was implemented via the `LogisticRegression` method from sklearn. The `'liblinear'` solver was used for the one-vs-rest approach. However, it cannot handle multinomial classification, so in that case we opted for the default `'lbfgs'` solver instead. The maximum number of iterations was set to 10000 in both cases. All other parameters were not changed from the default.

Validation

For all methods except clustering, the dataset was randomly split into a training set and a test set. The training set contained 80 % of the data and was used to train the model. The performance of the model on "new" data was then evaluated on the test set. In the case of SVC and Random Forest, 5-fold cross-validation was performed.

Because clustering is an unsupervised method, this method of validation cannot be carried out (there are no "correct" values to compare the results against) and therefore wasn't used.

IV. Results and Discussion

A. K-Means clustering

To visualise the clusters, PC1 and PC2 from the PCA analysis were used. Figure 8 shows that the two clusters are visibly separated, mostly by the value of PC1. This is not surprising because PCA and K-Means clustering tend to produce highly similar results [DH04]. They both minimize the mean-squared reconstruction error, only with different constraints and using different representations of the data vectors (PCA via eigenvectors, K-Means via centroids). Therefore, any cluster structures in the data will be picked up on by both methods. In K-Means clustering, the clusters will be mostly separated by principal component axes. Therefore, it is logical that the data in a PCA scatter plot will be visually separated in the same number of clusters that is optimal for a K-Means analysis.

More detailed results are shown in tables I and II. As expected, **cluster 1** contains mainly high GDP countries where the primary causes of death are those associated with high age

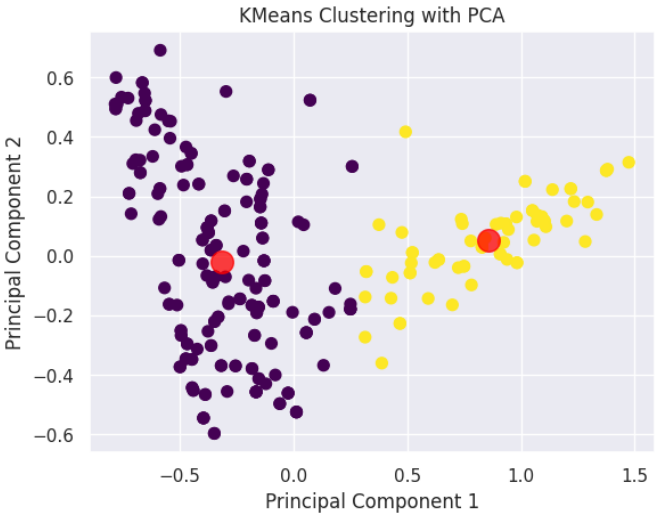

Fig. 8. K-Means Clustering with PCA

(dementia, Parkinson's, tumors). In **cluster 2**, the balance is shifted significantly towards infectious diseases (tuberculosis, malaria), poor life conditions (neonatal disorder, nutritional deficiency) and conditions that cause more fatalities when healthcare is poor quality or not readily available (diarrheal disease, HIV/AIDS). The countries in cluster 2 also have a lower GDP per capita value compared to cluster 1. This suggests a link between GDP per capita and quality of life/healthcare/life expectancy, although this analysis alone is not enough to prove it. Either way, it is clear that the two groups of countries face very different challenges in healthcare.

B. Random Forest

The model achieved an overall accuracy of **47.62% for GDP Group** and **31.90% for Air Pollution Group**, performing cross-validation to estimate how the model would perform on new data. In figure 9, the most important features for prediction of GDP Group are shown. In this model, the most important feature was **Neoplasm**, followed by Tuberculosis and Protein energy malnutrition. This further supports the theory that in countries with higher GDP per capita, death is more likely to be caused by often incurable conditions such as cancer that pose a significant challenge even to advanced healthcare systems, while countries with lower GDP are more at risk for premature death due to other health conditions. However, the accuracy of the model was so low that no conclusions can really be drawn from the results and it would not be useful for predictions. For that reason, Air Pollution Group prediction is not further discussed.

C. SVM

The SVM classification model without applying PCA and normalization showed a slightly better accuracy of **58.49%**. This result is comparable to the logistic regression model, which will be discussed further. Similar to previous cases, the

TABLE I
CLUSTER 1: SUMMARY.

	mean	std	min	25%	50%	75%	max
GDP	17079	20508	919	4959	7591	18982	113486
Air Pollution	22.691	13.374	6.455	13.261	20.549	26.469	87.022
Dementia	3.898	1.971	0.441	2.513	3.638	5.076	12.238
Parkinson's	0.945	0.338	0.235	0.743	0.881	1.158	1.935
Nutritional deficiency	0.356	0.422	0.000	0.046	0.159	0.509	1.832
Malaria	0.1155	0.5955	0.0000	0.0000	0.0000	0.0001	4.4404
HIV/AIDS	0.945	1.317	0.006	0.074	0.384	1.268	7.615
Tuberculosis	0.818	1.057	0.000	0.128	0.424	1.110	6.783
Cardiovascular	47.382	11.169	23.252	37.784	46.663	57.011	71.370
Lower respiratory	5.028	3.051	0.846	3.204	4.602	6.195	16.195
Neonatal disorder	2.519	2.541	0.057	0.330	1.820	4.069	10.735
Diarrheal disease	0.786	1.130	0.000	0.162	0.375	1.013	6.277
Neoplasm	25.273	8.063	11.457	18.228	25.106	31.257	42.683
Protein energy malnutrition	0.315	0.400	0.000	0.024	0.120	0.462	1.677
Chronic respiratory	5.570	2.825	1.390	3.375	5.078	7.502	16.262
Digestive disease	6.051	2.398	2.519	4.217	5.602	7.237	14.877

TABLE II
CLUSTER 2: SUMMARY.

	mean	std	min	25%	50%	75%	max
GDP	2269	3019	261	727	1295	2188	16329
Air Pollution	36.822	18.280	11.942	23.931	31.151	43.307	98.248
Dementia	0.944	0.447	0.246	0.652	0.810	1.170	2.581
Parkinson s	0.261	0.122	0.083	0.175	0.225	0.335	0.568
Nutritional deficiency	2.110	1.564	0.498	0.861	1.499	2.632	7.058
Malaria	7.664	7.425	0.000	0.856	5.617	14.136	25.027
HIV/AIDS	12.467	12.581	0.000	2.908	8.006	17.970	46.186
Tuberculosis	6.444	2.840	0.758	4.597	6.201	7.731	14.828
Cardiovascular	19.662	8.647	9.441	13.701	15.849	23.097	43.226
Lower respiratory	10.742	3.104	4.474	8.798	10.600	13.559	16.845
Neonatal disorder	11.093	4.095	4.492	8.233	10.647	13.439	23.971
Diarrheal disease	8.425	4.807	2.166	4.510	7.692	11.475	26.472
Neoplasm	8.946	3.106	3.792	6.769	8.746	10.771	19.472
Protein energy malnutrition	2.020	1.549	0.392	0.742	1.445	2.554	6.901
Chronic respiratory	4.282	3.809	1.551	2.331	3.049	4.059	23.899
Digestive disease	4.941	1.856	2.579	3.698	4.602	5.601	13.133

Air Pollution model achieved too low accuracy (**45,28%**) and will not be discussed.

Visualizing the **decision boundary** for an SVM model with an RBF kernel is challenging due to its **non-linear** and **high-dimensiona**l nature. To address this, data normalization and PCA were applied. The model was then visualized using the first two principal components, showing the SVC decision boundary and support vectors. Both the **One-vs-All (OvA)** and **One-vs-One (OvO)** approaches were implemented. The One-vs-All (OvA) approach achieved a lower accuracy of **49.1%** compared to the One-vs-One (OvO) approach, which attained **56.6%**. Therefore, the results of the OvA approach will not be presented in this work. Despite encountering a significant number of misclassifications (see in fig.1010) in the **SVC Decision Boundary** with the OvO approach, the model still suggests a potential relationship between GDP and death causes.

D. Logistic Regression

The accuracy was **58,49 %** for **multinomial** regression and **52,83 %** for **one-vs-rest** classification. This also shows in the confusion matrices in Figure 11 – the multinomial approach provides more correct predictions. Therefore, only the multinomial regression will be further discussed.

The overall accuracy was only slightly above 50 percent. However, a closer look at the confusion matrices reveals that the model was very successful at differentiating the extreme values (Very Low, Very High) and no such incorrect predictions were made. Moreover, all incorrect predictions involving these values were at most off by one (such as mistaking High for Very High, or Very Low for Low). This is consistent with the PCA results and again suggests that there is a relationship between very low or very high GDP per capita and cause of death.

The prediction of Air Pollution Group was not as successful.

E. Discussion and comparison

The prediction of **Air Pollution Group** was largely unsuccessful. For **GDP Group**, most models achieved at least satisfactory results. Logistic regression and SVM were similar in terms of performance with both models achieving roughly 60 % accuracy. The random forest model performed poorly in comparison. This suggests that considering the set of features as a whole is a better approach than the hierarchical structure of a decision tree. SVM, K-Means clustering, logistic

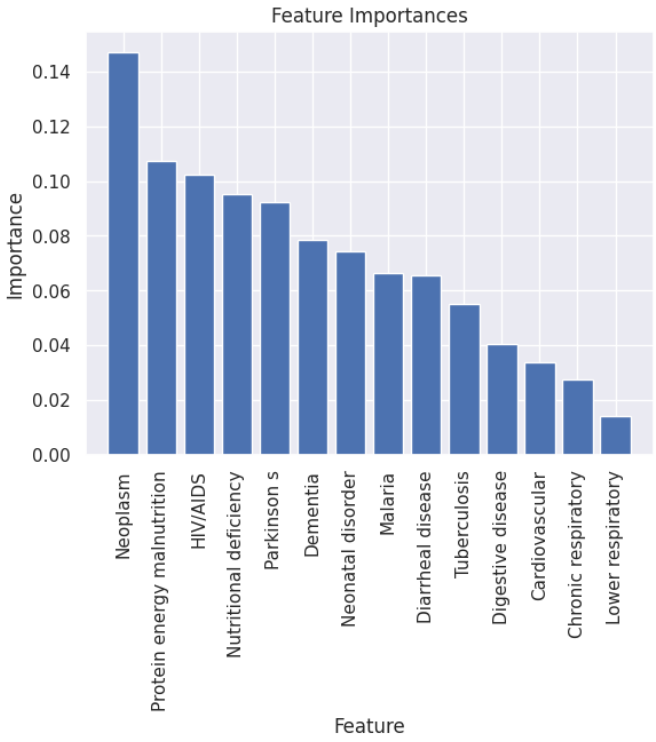

Fig. 9. GDP prediction, Random Forest model

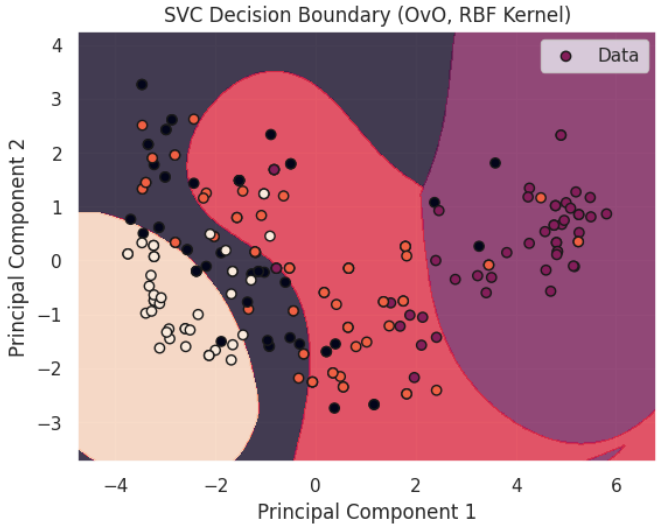

Fig. 10. GDP prediction: SVM model using RBF Kernel and PCA

regression and PCA showed a similar outcome: the countries are separable into two distinct groups. The high GDP group is characterized by more old age related and incurable disease deaths, while the low GDP countries are more prone to infectious disease and conditions that are mostly eliminated or treatable in the developed world. Clustering, SVM and logistic regression are mostly tied in terms of performance so any of them could be a good choice.

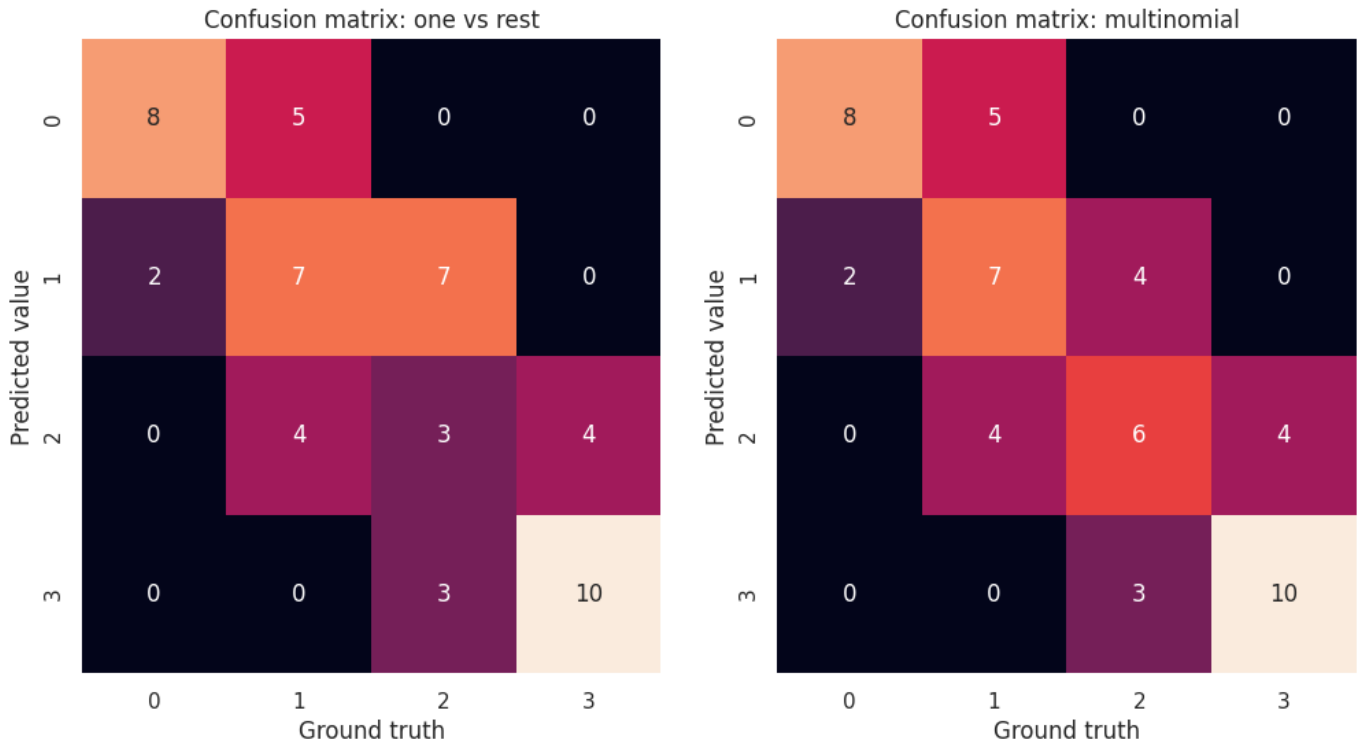

Fig. 11. GDP Group logistic regression: confusion matrices. Encoded as 0 = Very Low to 3 = Very High.

V. Conclusion

Different machine learning algorithms were applied to explore the possibility to link the biggest healthcare challenges to life conditions in a given country. Air pollution was not found to be a good candidate for this. However, some promising results were achieved with GDP prediction. The study suggests that there is a relationship to be explored between the wealth of a country and the healthcare challenges that country might face. In a further study, GDP per capita could be replaced by more variables that more closely map the personal wealth, wellbeing and quality of accessible healthcare of the citizens. Moreover, while the results of this study suggest that a model that predicts cause of death based on those indicators could be created, this needs to be confirmed by further research.

VI. Contribution

Elia Pavliš: introduction and conclusion, PCA, clustering, logistic regression

Sofie Meyer: preprocessing, PCA, SVM, random forest

References

[DH04] Chris Ding and Xiaofeng He. "K-means clustering via principal component analysis". In: *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML '04. Banff, Alberta, Canada: Association for Computing Machinery, 2004, p. 29. ISBN: 1581138385. DOI: 10.1145/1015330.1015408. URL: https://doi.org/10.1145/1015330.1015408.

[CD10] Lijun Cheng and Yongsheng Ding. "SVM and statistical technique method applying in Primary Open Angle Glaucoma diagnosis". In: *2010 8th World Congress on Intelligent Control and Automation*. 2010, pp. 2973–2978. DOI: 10.1109/WCICA.2010.5554175.

[Jai10] Anil K. Jain. "Data clustering: 50 years beyond K-means". In: *Pattern Recognition Letters* 31.8 (2010). Award winning papers from the 19th International Conference on Pattern Recognition (ICPR), pp. 651–666. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2009.09.011. URL: https://www.sciencedirect.com/science/article/pii/S0167865509002323.

[KKN14] Samina Khalid, Tehmina Khalil, and Shamila Nasreen. "A survey of feature selection and feature extraction techniques in machine learning". In: *2014 Science and Information Conference*. 2014, pp. 372–378. DOI: 10.1109/SAI.2014.6918213.

[Sha+14] Yanming Shao et al. "BOF endpoint prediction based on the flame radiation by hybrid SVC and SVR modeling". In: *Optik* 125.11 (2014), pp. 2491–2496. ISSN: 0030-4026. DOI: https://doi.org/10.1016/j.ijleo.2013.10.094. URL: https://www.sciencedirect.com/science/article/pii/S0030402613014666.

[RCH17] Qiong Ren, Hui Cheng, and Hai Han. “Research on machine learning framework based on random forest algorithm”. In: *AIP Conference Proceedings* 1820.1 (Mar. 2017), p. 080020. ISSN: 0094-243X. DOI: 10.1063/1.4977376. eprint: https://pubs.aip.org/aip/acp/article-pdf/doi/10.1063/1.4977376/13140229/080020_1_online.pdf. URL: https://doi.org/10.1063/1.4977376.

[Li+18] Wei Li et al. “Fault detection, identification and reconstruction of sensors in nuclear power plant with optimized PCA method”. In: *Annals of Nuclear Energy* 113 (2018), pp. 105–117. ISSN: 0306-4549. DOI: https://doi.org/10.1016/j.anucene.2017.11.009. URL: https://www.sciencedirect.com/science/article/pii/S0306454917303912.

[Ade+22] Abigail Bola Adetunji et al. “House Price Prediction using Random Forest Machine Learning Technique”. In: *Procedia Computer Science* 199 (2022). The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 2021): Developing Global Digital Economy after COVID-19, pp. 806–813. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2022.01.100. URL: https://www.sciencedirect.com/science/article/pii/S1877050922001016.

[Org] World Health Organization. *Malaria Fact Sheet*. https://www.who.int/news-room/fact-sheets/detail/malaria. Accessed: (16.06.2024).