**The Smith Parasite**

*Group Project*

**MACHINE LEARNING 2022/2023**

# 

[**Introduction**](#_6bf0ftv7220h) **3**

[**Exploration**](#_mcumyktv9nf4) **3**

[**Preprocessing**](#_g1sm27w10qkh) **4**

[Data cleansing and outlier treatment](#_iguaehy24tt3) 4

[Creation of new variables](#_9ljxllwpvcm7) 5

[Univariate analysis, grouping and variable transformation](#_84myndqi0tor) 5

[Feature Scaling](#_33gos376jjko) 5

[Feature Selection](#_gieut0qpolo3) 6

[**Modeling**](#_42bkee2bt0u2) **6**

[**Assessment**](#_hr2djmm0t7r0) **8**

[Metrics](#_cxwi7sa7m5mk) 8

[Assessment methods](#_qv0e4anjek6b) 9

[Results](#_d7ij3gc4g9oa) 9

[**Conclusion**](#_61irqwkc1ks2) **10**

[**References**](#_7gtgzx1bjl3n) **11**

[**Annexes**](#_pe7foo92y3k) **12**

# Introduction

This report presents the process to create a predictive model to discover patient vulnerability towards infection with the Smith Parasite. Our objective is to identify a pattern for this infection, which has already affected over 5,000 unrelated people. To conduct this study, we evaluated and engineered three sets of data: sociodemographic, health, and habit data, consisting of 17 original variables. Multiple supervised learning models are deployed in order to discover the most accurate predictions and answer the question: "Who is more likely to suffer from the Smith Parasite?".

# Exploration

In this section, the underlying datasets are presented. This includes a brief description of the variables, as well as their unique characteristics. The data exploration is supported by tables and graphical analysis, referenced in the annexes. The data is structured as follows, with variable names printed in bold letters:

***Dependent variable*** - whether or not the patient contracted the Smith disease (**Disease**).

***Sociodemographic Data*** *-* containssocial and demographic data, including **Name**, **Birth\_Year**, **Region** and **Education**.

***Health Related Data -*** contains data on the fundamental health indicators, such as the **Height**, **Weight**, routine checkup pattern (**Checkup**), whether or not they or their family members have diabetes (**Diabetes**), their cholesterol and blood pressure levels (**High\_Cholesterol** and **Blood\_Pressure** respectively), as well as the number of the days in a month when they feel mentally or physically handicapped (**Mental\_Health**, **Physical\_Health**).

***Habits Related Data -*** presents behavioral information on patients, which includes smoking, drinking, exercising, fruit eating, and water drinking habits (**Smoking\_Habit**, **Drinking\_Habit**, **Exercise**, **Fruit\_Habit** and **Water\_Habit** respectively).

As seen in *Annex 1,* the training data has 800 observations, including the dependent variable. The test dataset consists of 225 observations without any information about the disease outcome. The 17 independent variables include ten categorical and seven numerical variables. The numerical variables and their basic statistical data are presented in *Annex 2* and *Annex 3*. The detailed view on percentiles allows one to interpret the distributions and identify outliers or null values. While Education exhibits some 13 null values, High\_Cholesterol, Birth\_Year, and Physical\_Health show unusual values to the extremes. More details are described in the preprocessing section.

Furthermore, box plots, histograms and pairwise scatter plots are investigated for continuous variables (refer to *Annex 4 and 5*). In the diagonal line in *Annex 4*, one can see histograms of the individual variables’ distribution. Particularly **Blood\_Pressure**, **Physical\_Health** and **Height** are distributed unevenly or skewed. Visually, most of the variables are weakly pairwise correlated, the only recognizable moderate correlation being between Height and Weight. Bar graphs are used to evaluate the categorical variables (*Annex 7*). Data pertaining to habits such as drinking, smoking, and water consumption revealed vastly disparate classifications.

The dependent variable Disease is distributed almost equally with positive and negative cases in approximately the same amounts. The average disease rate in the sample is 0.51. This simplifies the analysis, as no stratification has to be applied and enough values are available to understand the characteristics of both states of the infection.

# Preprocessing

The data is processed in multiple ways to be usable for modeling. This chapter describes the null values and outlier treatment, groupings, new variable creation, scaling, and feature selection. The goal of this is to avoid data anomalies, and to prepare the data for the characteristics and requirements of the classifiers. These steps help to increase the accuracy of the classifiers within our sample and beyond.

## Data cleansing and outlier treatment

Strong outliers with extremely high values are reported in the variable **High\_Cholesterol**. The three outliers even had the same value in **High\_cholesterol** (568), which likely stems from a measurement error and are excluded from the analysis. This prevents the models from being biased and should lead to higher accuracy, as it now fits the majority of the data better. This is supported by such outliers not being found in the test dataset (see *Annex 3*).

**Education** has 13 observations reported as null and is the only variable with missing data. Null values in **Education** are not excluded from the analysis, as a grouping heuristic was applied for the variable to keep more information. No missing data is detected in the test dataset for **Education**.

Anomalies are seen in **Birth\_Year**, which has a large gap in its interval and eleven surprisingly low values (*Annexes 4 - 5*). See the distribution of the **Birth\_Year** on the train and test datasets on the same figure in *Annex 6.* The shape and range of the distribution of those eleven values within the 19th century is visually the same as the distribution of **Birth\_Year** within the 20th century. These outliers appear to be the result of a single-digit error. This assumption is backed by 19th century observations that include blood pressure information, while the blood pressure measuring device was not invented until the 1880s. The outliers are adjusted 100 years upwards to fit into the distribution. This allows us to keep more observations for modeling.

Extreme values in **Physical\_Health** are not considered to be outliers as this variable has reasonable lower and upper limits between 0 and 30. The few patients reporting the value 30 can be interpreted as the natural distribution being strongly skewed towards lower values.

## Creation of new variables

Two new variables were introduced in the dataset - **Gender** and **BMI** (Body Mass Index). Gender is directly derived from the variable **Name** as it contains a title - “Mr.” or “Mrs”. **Name** becomes redundant after this operation as it cannot provide any additional information for the model anymore [1]. The BMI is a convenient measure used to broadly categorize a person as underweight, normal weight, overweight, or obese, based on tissue mass and height. Major adult BMI classifications are underweight (under 18.5), normal weight (18.5 to 24.9), overweight (25 to 29.9), and obese (30 or more). The **BMI** is calculated with:

|  |  |
| --- | --- |
|  | [[2] |

## Univariate analysis, grouping and variable transformation

To make the categorical data usable for modeling, they need to be one-hot encoded, leading to an excessive amount of total features. To reduce dimensionality, the categories are systematically grouped. Each category within a variable is assessed based on their mean disease rate, logical relatedness and quantity. This also allows to merge small categories, which reduces imbalances, sample bias, and makes the overall model more economical. The ten categorical variables in the dataset had 42 different categories within them, which are reduced to 24 through grouping. Only **Gender**, **Exercise**, and **Smoking\_Habit** are not changed. Continuous variables are also analyzed for a transformation to a binary variable. In case of evident levels within a variable’s relationship to disease, or unbalanced, skewed distributions, the data is adjusted. The continuous variables **BMI** and **Physical\_Habit** are transformed to a binary and 4-category-variable respectively. Graphs about the variables before and after data transformation, as well as written reasoning for the most important decisions, are presented in *Annex 8* for categorical and in *Annex 9* for continuous variables.

## Feature Scaling

Feature scaling can be crucially necessary when using distance-, variance- or gradient-based methods (KNN, PCA, neural networks, regularized regressions), because depending on the case, it can improve the quality of results or the computational effort. However, in some cases (tree-based models in particular, logistic regression with no regularization), scaling has no impact on the performance [3]. As we are going to explore different modeling techniques, some scaling is necessary to apply.

There are three major feature scaling techniques commonly used in Machine Learning: Robust Scaling, Standardization (Z-score normalization), and Min-Max Scaling. Robust scaling is useful when the data contains outliers. Standardization provides better results under the assumption that the feature distributions are having a Gaussian distribution. As the outliers were already removed from the data and also none of the distributions are having a Gaussian shape, Min-Max-Scaling was applied for all continuous variables. This is also backed by the fact that most of the features are naturally bounded within some reasonable range and cannot have extreme values (for example, Weight and Height). Min-Max Scaling transformation preserves the distance in between observations, but scales its values in a range between 0 and 1.

## Feature Selection

Despite the reduction of features through grouping, it is still necessary to validate the processed features. There are three main categories of feature selection algorithms: wrappers, filters and embedded methods [4]. We decided to explore at least one method from every category - Recursive Feature Elimination (RFE) with Logistic Regression as a wrapper-style method, Lasso Regression as an embedded method, both Chi-squared Test and Spearman’s correlation as filter methods.

Both continuous and categorical variables were tested according to selection algorithms suitable to their scaling [5]. Spearman’s correlation was applied to each continuous feature singularly, while a Chi-squared test was applied to each single categorical variable. A p-value of below 0.05 was required for a significant result. In order to capture the behavior of variables amongst one another, Logistic Recursive Feature Elimination and Lasso Regression were applied on the whole featureset. Categorical dimensions with more than two categories were transformed into dummy variables. The medium category was excluded before any analysis to avoid perfect collinearities and redundancy [6]. When deciding which variables to keep, a voting heuristic is applied. All variables require to have at least two out of three possible positive results in the selection algorithms. As a result, **Education\_grouped\_Bad**, **Water\_Enough**, and **Diabetes\_I don't have diabetes, but I have direct family members who have diabetes** were excluded from the final modeling.

Additionally, a correlation matrix is set up to investigate possible multicollinearity **(**see*Annex 11***)**. If two predictor variables are strongly correlated, their information is overemphasized in the models and could result in a bias. Especially the variables **Weight**, **Height**, **Gender**, and **BMI** are very related. Since **BMI** is deemed to be more informative about a patient than **Weight**, which can be reconstructed with the **Height** dimension, **Weight** is also dropped from further analysis. All in all, 24 variables were reduced to 20, which are kept for the final modeling. The results are seen in *Annex 10*.

# Modeling

With the reduced featureset, a diverse set of models for supervised machine learning is applied to predict the propensity for infection with the Smith Disease. We followed an approach of building models with successively increasing complexity to test their predictions against the testing dataset later. At first, the base models Decision Tree, Gaussian Naive Bayes, Logistic Regression, K-Nearest Neighbor (KNN), Multi-Layer Perceptron, Support Vector Machine, and Linear Discriminant Analysis are constructed. These models are diversified further by using a Bagging approach. Additionally, other Ensemble methods are applied. These are Random Forest, Ada Boosting and Gradient Boosting. Finally, meta classifiers, a Voting and a Stacking Classifier, are implemented. Due to the models being extensively covered in class, we decided not to explain their methodologies in detail and focus on the training approach and hyperparameter tuning.

Early prediction assessments reveal that many chosen models easily predict the data correctly. To make sure that this is not induced by overfitting, the training data is first split into a 20% validation set and a 80% training dataset. This allows us to have a second checkmark for making model decisions, particularly for tree-based learners who are prone to overfitting on the training data.

Each model underwent grid search or random search in combination with cross validation to find the best specifications for each algorithm, according to the F1 score. Grid search was used when algorithms had an overseeable range of possible parameters. It “works through different combinations of parameter tunes, cross validates each and determines which one gives the best performance”[7]. When complexity grew too large, such as in the Bagging methods or the Multi Layer Perceptron, random search was applied. Random search is preferably used to randomize training instead of using every single feature combination. As some hyperparameters are more useful than others, this can lead to finding better model specification faster [7]. By training the model with different combinations of hyperparameters and evaluating them using K-fold cross-validation, we make sure that the best parameters do not overfit on one subset of the data.

**Ensemble Methodology**

In combination with basic models, empirical evidence has shown performance improvement through bagging [8]. We replicated this with our dataset and fitted a bagging model on each base learner. Refer to the assessment section to learn about their performance.

As it quickly became clear that the existing models performed very well, we wanted to generalize the final model even further. Since higher model complexity is not a disadvantage in disease prediction, we aggregate the best models into a combined meta ensemble classifier. Every model with a cross-validated F1 score of over 0.95 was considered. When models that embody different approaches to prediction are combined, we can increase diversity and possibly make better decisions on unseen data [9]. Due to the strong individual model performances, which already include combined weak classifiers, a non-trainable combiner was chosen. The almost equally strong model inputs make weighted voting of each classifier obsolete, which allows for a simplified majority vote classifier. Contrary to the default method on *sklearn*, the pretrained models make up the classifier and are not trained again.

For another alternative to the voting classifier, a stacking classifier is implemented. Contrary to the previous method, all previously built models are included. In a second step, the models’ predictions work as inputs for the default logistic regression. Simplified, the regression then calculates the most meaningful weights for the final prediction. Same as in the Voting Classifier, the models of the first layer are not trained again and correspond to the pretrained ones.

# Assessment

In this section, we will elaborate on the evaluation of our models. Firstly, we discuss why assessment of the model is important, followed by description of the metrics we decided to use and why we used them. We also report on how we conducted the assessment, and finally, discuss the results. The goal of evaluating a model in Machine Learning is to “determine the usefulness of our learned classifiers on various collections of data sets” [10], by assessment of prediction performance. Ideally, this evaluation should represent how the model would perform in a real life scenario [11].

## Metrics

Only metrics suitable for classification problems are considered. The idea is to use basic metrics, such as precision and recall, as well as the summary metrics accuracy and F1. As mentioned in Keller et. al., in order to fully comprehend how a model is performing, it is recommended to use multiple performance measures[12]. In addition to selected metrics, we analyze the time performance of our models. Our primary metric of choice is: The F1-score, which is used to measure accuracy of the model.

F1 Score is the “harmonic mean of precision and recall”[13]. The score emphasizes the performance on the positive level, which in our case, where it is more important to determine if a patient is sick than if he is not, is valuable.

We also consider accuracy, which is the proportion of events correctly identified (positive or negative) on all events[14]. Accuracy is considered to be a simple indicator of model performance. It works well with balanced classification tasks. We use it as a reference metric, i.e. compare it with the F1 score.

Additionally, we useRecall (True Positive Ratio) that tells us how confident we can be that “all the instances with the positive target level have been found by the model.[15] Recall is associated with costly False Negatives. Again, referring to a medical problem example - when falsely informing a sick patient that he/she is healthy can be dangerous, i.e. if sickness is contagious.

The last metric is Precision, which encapsulates the “proportion of Predicted Positive cases that are correctly Real Positives”[16]. High costs are associated with False Positives. It is less relevant when it comes to the sick patient example, however, also important. Someone who was diagnosed with cancer, while not actually having one, might be detrimental.

## Assessment methods

We have used evaluation methods in various ways throughout the project. After hyperparameter tuning on the training data, we evaluate all model performances. Then, by using Stratified K-Fold, a cross validation method that uses stratified sampling, we score the models. In this step, we include both training and validation data. The additional validation data helps to estimate model accuracy on previously unseen data. Stratified K-Fold is considered to be a solid approach when testing models in classification[17]. In the next part, we fit the pretrained models with fixed hyper parameters on three different datasets: the dataset transformed with all features, the dataset after feature selection, and a dataset containing the 15 most relevant variables as defined by the Random Forest most\_important\_features method. This is done to additionally validate the feature selection. One can find the feature importance in *Annex 12*withbehavioral, demographic and health features equally represented. We observe how performance changed with cross validation between the datasets, focusing on the F1 score. The figures 1,2,3,4 in *Annex 13*present the outcome. As observed, the 15-features-dataset creates slightly better performance. The final evaluation with the Stratified-K-Fold method consists of all models using the 15 most important features, by calculating all selected metrics and training time.

## Results

The final evaluation produced the results, as seen in *Annex 15***,** which depict basic models and bagged models. All four metrics hold different significance and provide meaningful insights into the performance of models. However, as the problem is in the medical domain, specifically the detection of sick patients, it is reasonable to place more emphasis on the F1 score, followed by Recall. As mentioned earlier, these metrics focus on the positive level of prediction, which for our problem area is more significant. Just taking into account the results of the 15-feature dataset, various models perform well. Random Forest, Gradient Boost, ADA Boost, the Voting Ensemble, Decision Tree, Stacking, and the Bagged KNN method scored at least 0.95. Overall, the Bagged KNN model is chosen as the one to deploy. It shows a consistent score of 0.96 across all metricsOverall, the models showed similar performances with slight differences in results every time we ran them. Even though the Stacking Method and the Random Forest have a score of 0.97 in some metrics, the first one is computationally much more expensive while the latter one shows only 0.95 in Recall. The Voting Classifier also shows the same results as the Bagged KNN method, but is much more heavy in its setup, due to the inclusion of the MLP. Long training times are particularly seen in the Bagged Logistic Regression, Multi-Layer Perceptron, and the Ensemble Voting Classifier. Hyperparameter tuning is especially long-bearing for the (Bagged) MLP methods, which make the algorithms unattractive for further use.

So, taking the computational efficiency of the model into account, we were led to make a decision between bagged KNN and ADA Boost. After careful consideration, we decided to deploy the ADA Boost model because it consistently showed the highest accuracy on both the validation and training sets

Overall, it can be observed that the other large ensemble methods provided a clear performance boost. Interestingly, in the first training run (Annex 14), it can be seen that the Ensemble and AdaBoost methods achieved almost a perfect F1 score. However, in subsequent runs, the scores were lower. This highlights the importance of cross-validation when determining the best performing model, as some models may be more sensitive to overfitting and small changes in the data can significantly affect their performance. Methods such as bagging and boosting are designed to reduce these effects. In our bagged methods, we observed small improvements in the mean scores over the basic models in some cases. The Support Vector Machine, Multi-Layer Perceptron, Decision Tree, and Naive Bayes performed worse with bagging, but not significantly. However, the bagging methods performed much better in the first training run, indicating a stronger degree of overfitting. Ultimately, we found that the ADA Boost model was preferable due to its consistently high accuracy on both the training and validation sets, as well as its reasonable training time and low requirement for hyperparameter tuning.

# Conclusion

The objective of this report was to identify a pattern for the Smith Parasite infection and to answer the question of "Who is more likely to suffer from the Smith Parasite?" To achieve this, we evaluated and engineered three sets of data: sociodemographic, health, and habit data. We then applied multiple supervised learning models and found that the Bagged KNN model was the best fit, according to a cross-validated F1 score of 0.96.

Through our exploration and preprocessing of the data, we found several important predictors of the Smith Parasite infection, including a last Checkup being more than three years ago, birth year, mental health, high cholesterol, bad fruit habits, diabetes, or blood pressure. By understanding these patterns, we can identify groups of people who are at a higher risk of contracting the Smith Parasite and take preventive measures to protect them. Further research is needed to confirm these findings on a larger scale internationally and to identify potential interventions that can reduce the risk of infection in vulnerable groups.

Overall, this report has provided a comprehensive analysis of the Smith Parasite infection and has identified several key factors that can help predict its occurrence. We hope that this information will be useful in the efforts to combat this disease and protect vulnerable populations.

# 

# References

**[3]** Ahsan, M., Mahmud, M., Saha, P., Gupta, K., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies, 9(3), 52. <https://doi.org/10.3390/technologies9030052>

**[7]** Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281-305.

**[8]** Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. doi:10.1023/A:1010933404324

**[1]** [*The SuRF Report 2*](https://apps.who.int/iris/bitstream/handle/10665/43190/9241593024_eng.pdf) (PDF). The Surveillance of Risk Factors Report Series (SuRF). World Health Organization. 2005. p. 22. [Archived](https://ghostarchive.org/archive/20221009/https://apps.who.int/iris/bitstream/handle/10665/43190/9241593024_eng.pdf) (PDF) from the original on 2022-10-09

**[6]** Dormann, C. F., Elith, J., & Bacher, S. (2012). *Collinearity: A review of methods to deal with it and a simulation study evaluating their performance* (Vol. 36, Ser.1,pp.27-46,Rep.).Ecography.doi:<https://doi.org/10.1111/j.1600-0587.2012.07348.x>

**[17]** Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, *4*, 972421.

**[15,16]** Powers, David M W (2011). ["Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation"](https://web.archive.org/web/20191114213255/https://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf) (PDF). *Journal of Machine Learning Technologies*.

**[12,13,14]** Kelleher, J. D., Namee, M. B., & D’Arcy, A. (2020). Fundamentals of Machine Learning for Predictive Data Analytics, second edition: Algorithms, Worked Examples, and Case Studies. The MIT Press.

**[11]** Gaël Varoquaux, Olivier Colliot. Evaluating machine learning models and their diagnostic value. Machine Learning for Brain Disorders, In press. ffhal-03682454v3f

**[10]** Japkowicz, N. (2006). Why Question Machine Learning Evaluation Methods ? ( An illustrative review of the shortcomings of current methods).

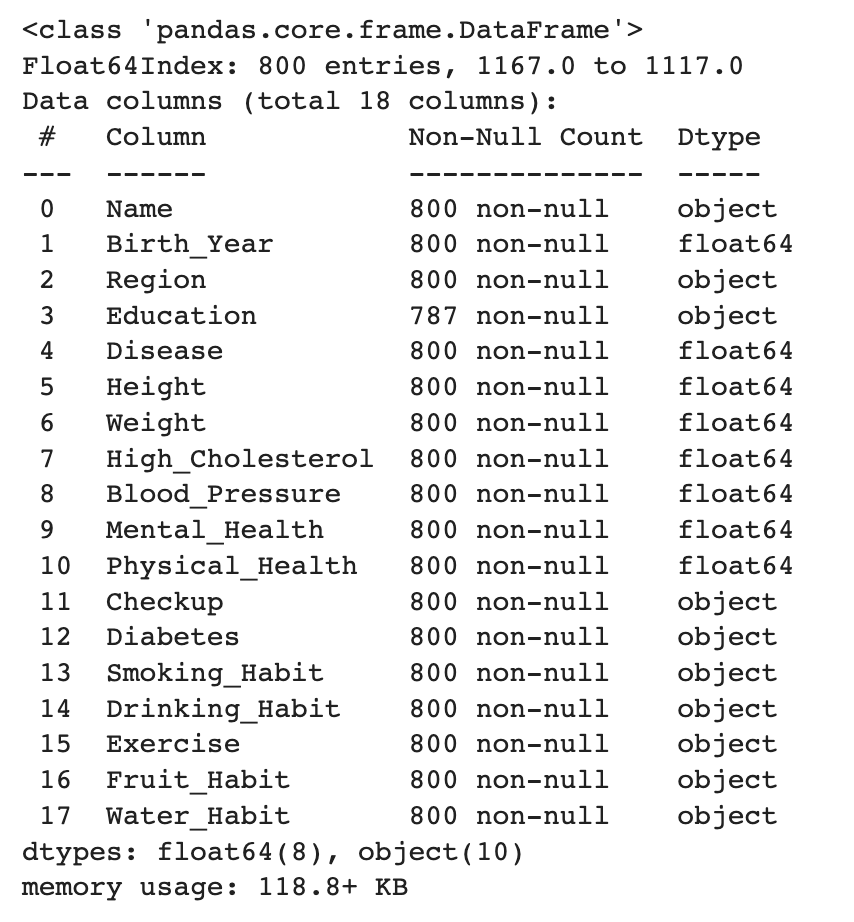
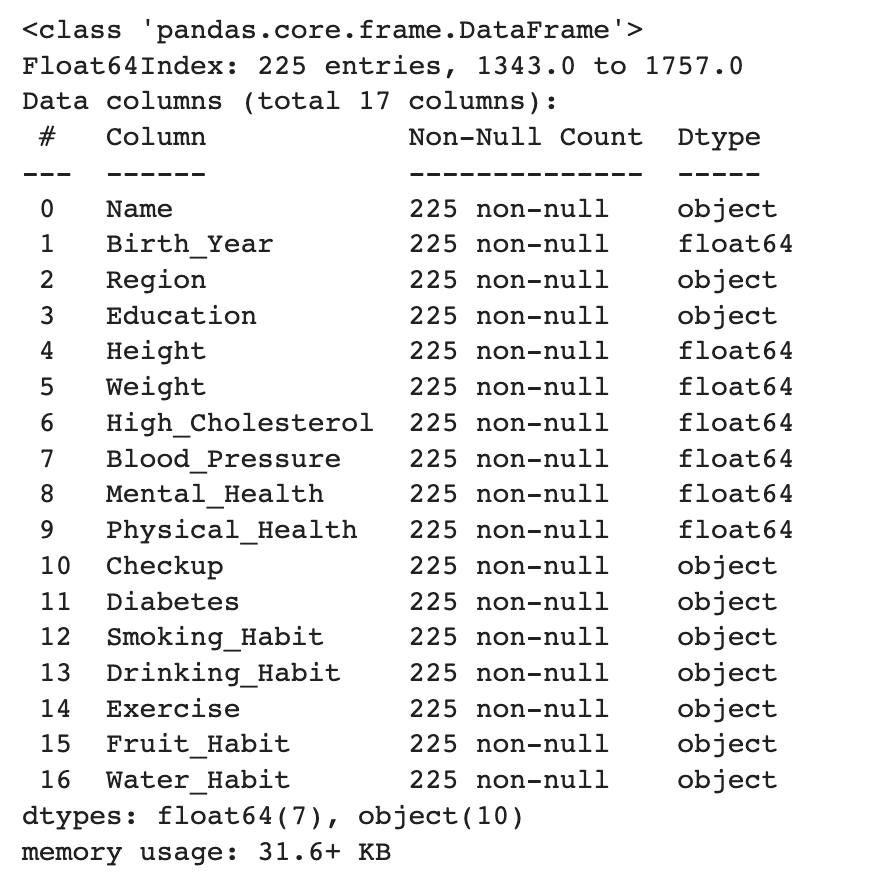
**[9]** Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

**[2]** M. Wang, W. Fu, X. He, S. Hao and X. Wu, "A Survey on Large-Scale Machine Learning," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 6, pp. 2574-2594, 1 June 2022, doi: 10.1109/TKDE.2020.3015777.

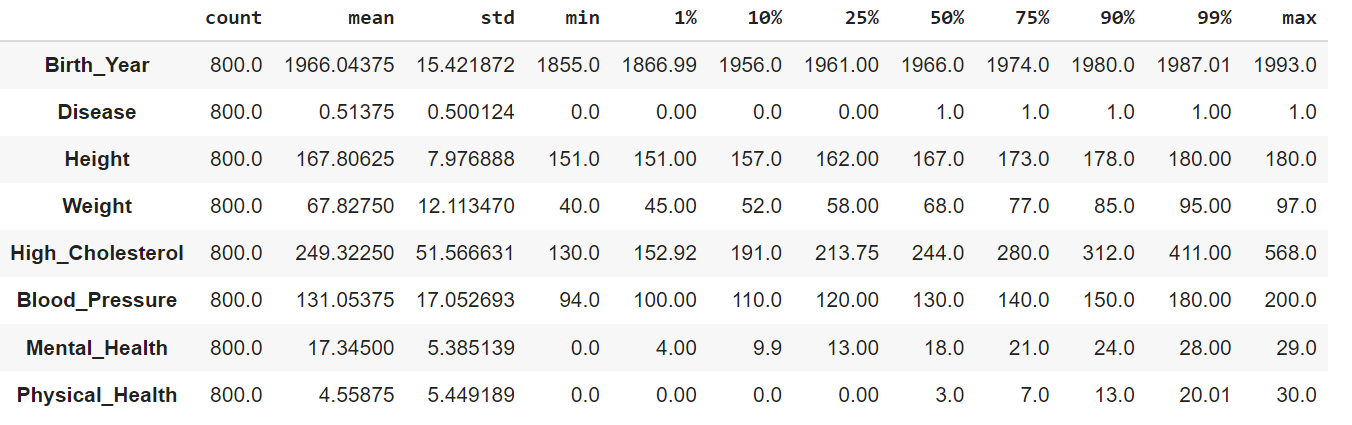
**[4,5]** Guyon, I., &amp; Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research,3,1157–1182. https://doi.org/https://elearning.novaims.unl.pt/pluginfile.php/141018/mod\_resource/content/0/guyon03a.pdf

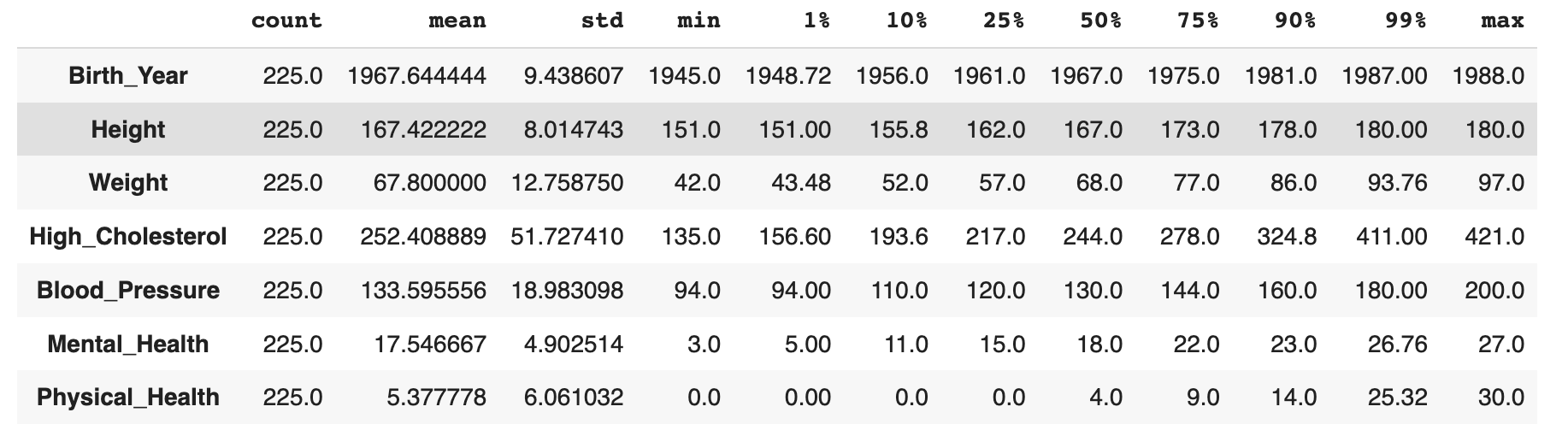
# 

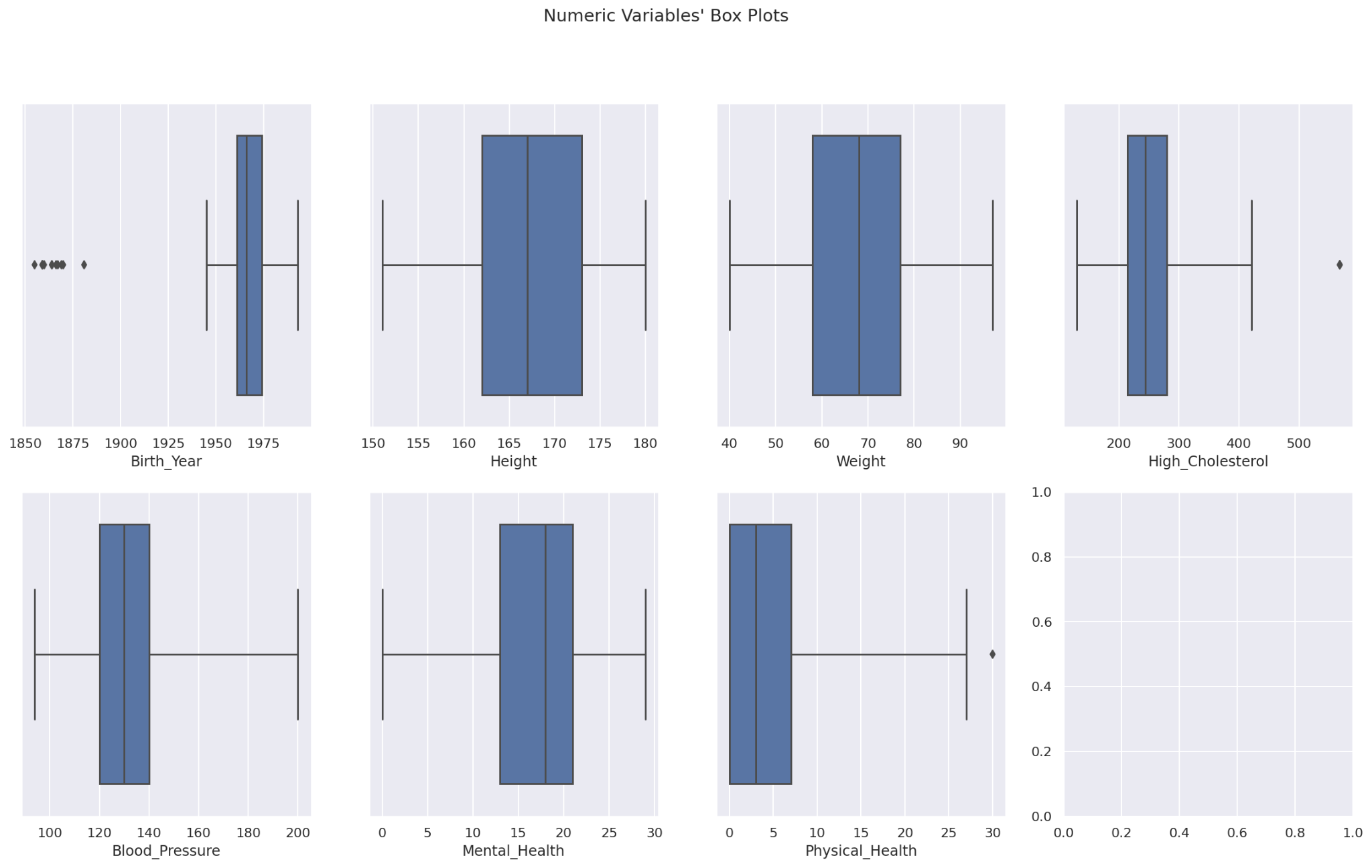
# Annexes

*Annex 1.* **Summary of of the train and test datasets**

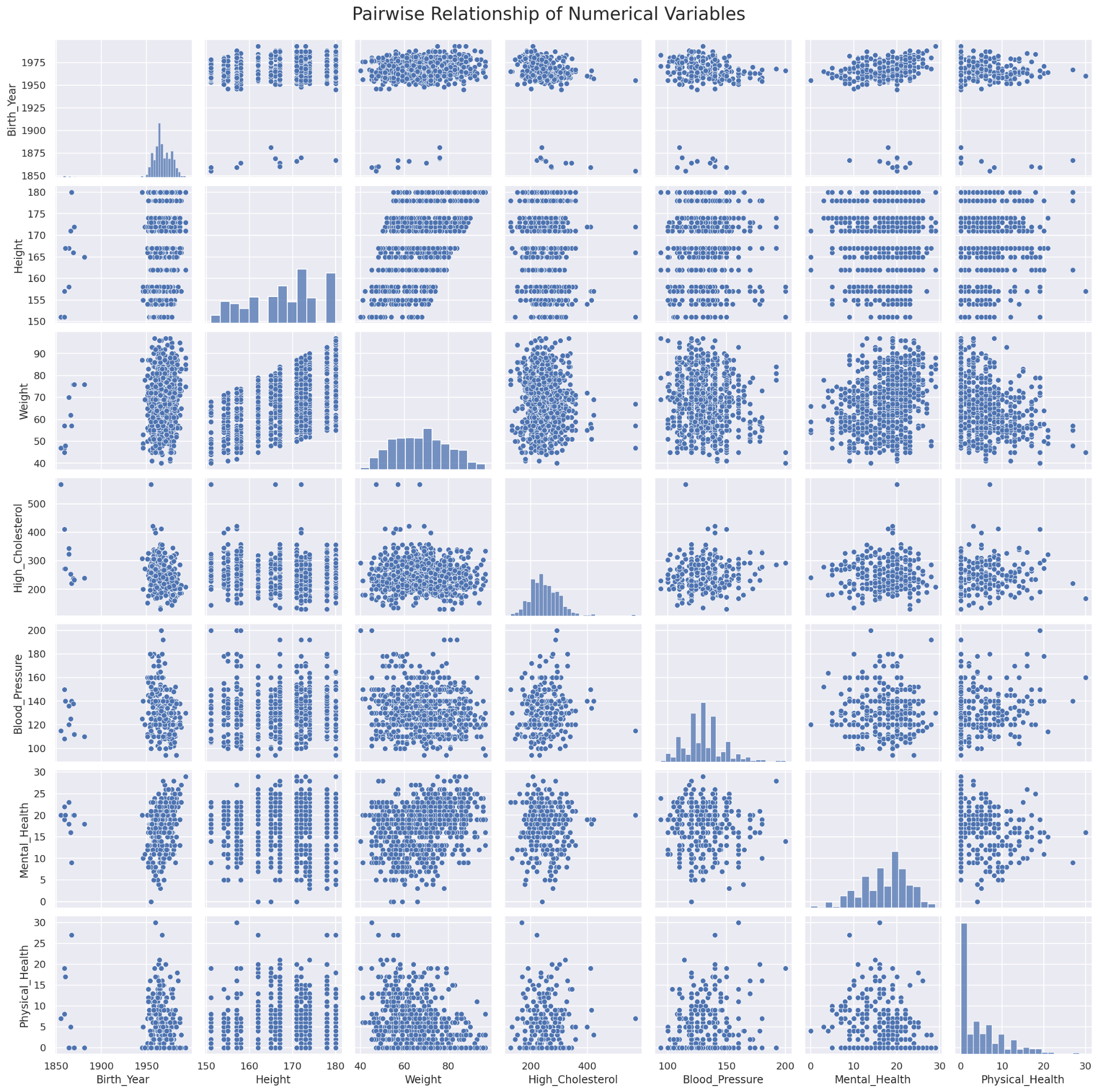
*Annex 2.* **Basic statistics of numeric variables on the train dataset**

*Annex 3.* **Basic statistics of numeric variables on the test dataset**

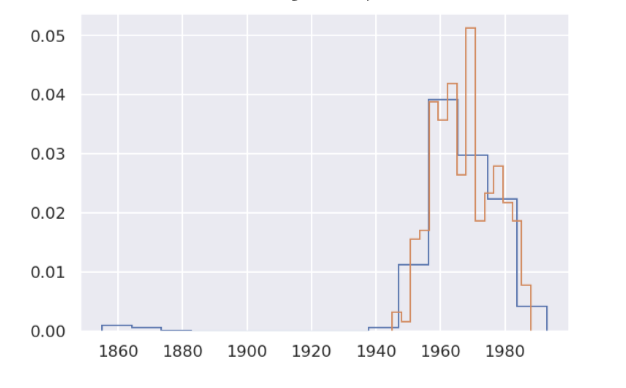


*Annex 4.* **Boxplots of numerical variables**  


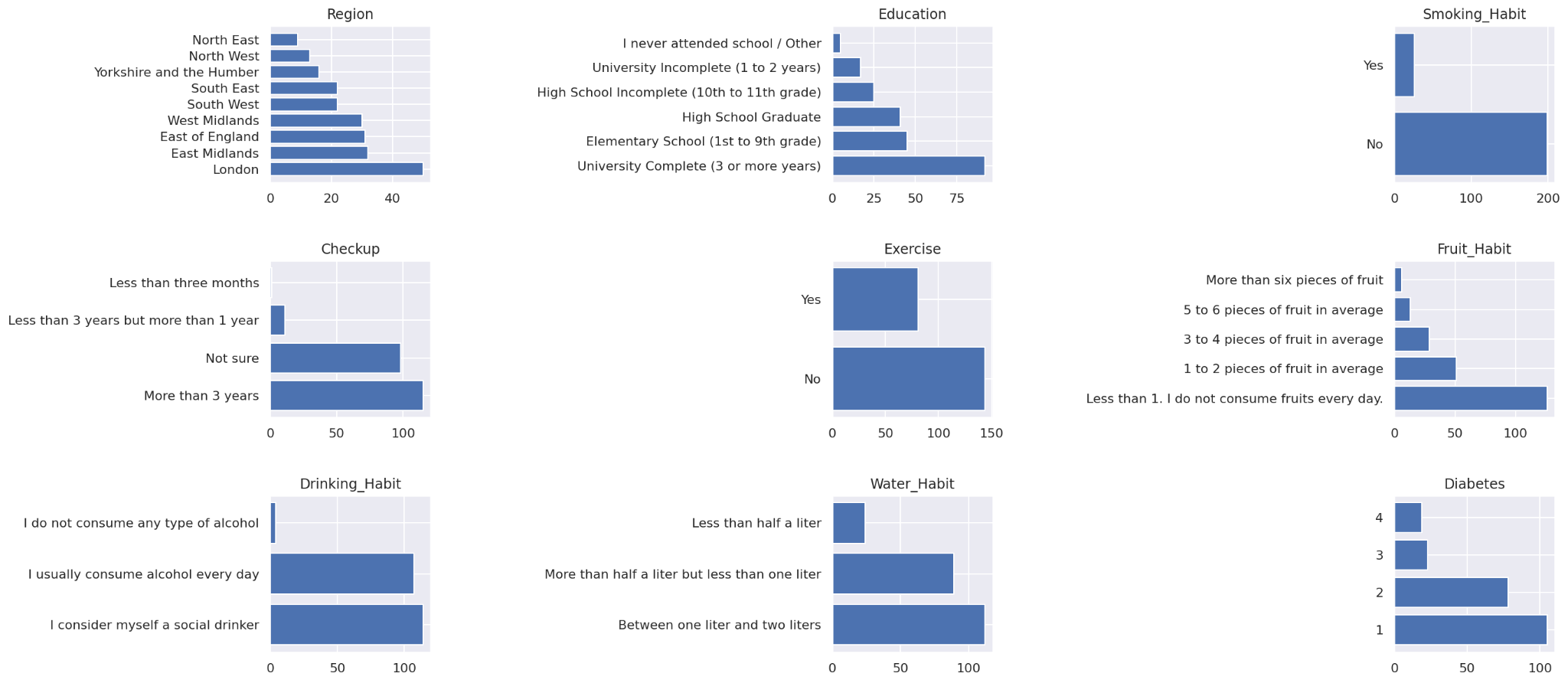
*Annex 5.* **Pairwise relationship of raw numerical variables along with their histograms**



*Annex 6.* **Birth year distribution on the train (blue line) and test (orange line) datasets**



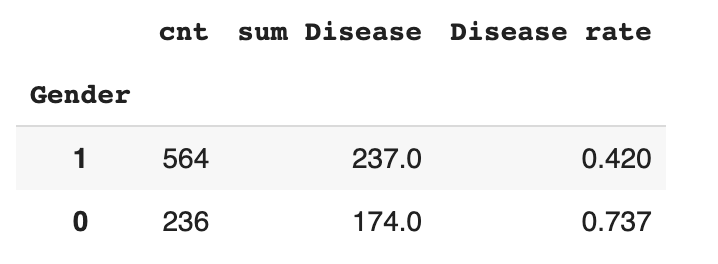
*Annex 7.* **Plotting categorical variables**

The bar plot for the variable "Diabetes" is presented through numbers 1 to 4 due to the lengthy titles, representing the following answers: **1 - Neither I, nor my immediate family had diabetes   
2 - I have had pregnancy diabetes or borderline diabetes   
3 - I do have diabetes  
4 - I don't have diabetes but I have direct family who has diabetes)**

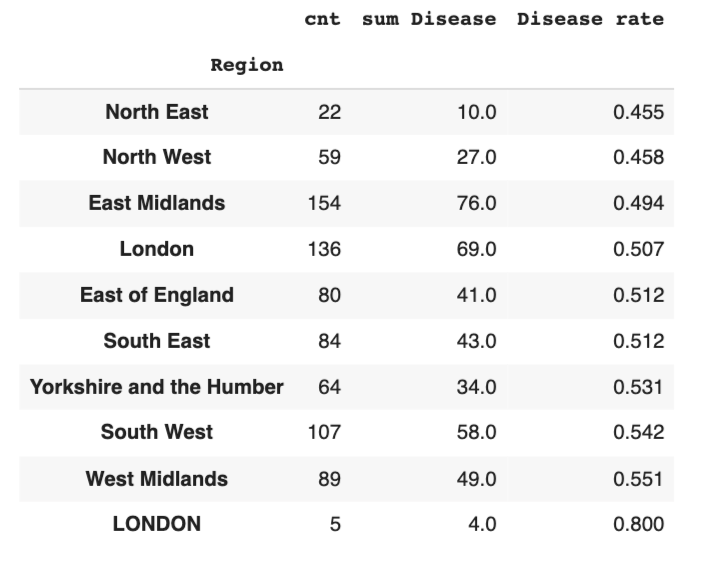
*Annex 8.* **Univariate analysis of the raw categorical variables**

***Sociodemographic variables***

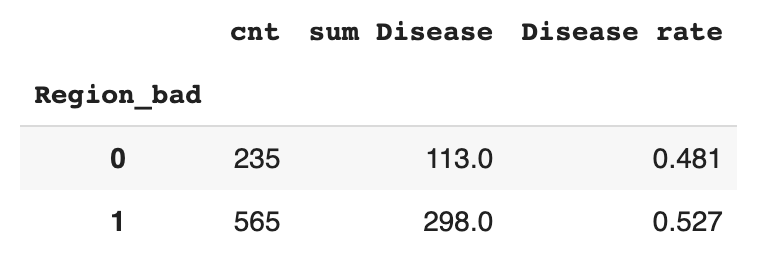
***Gender***



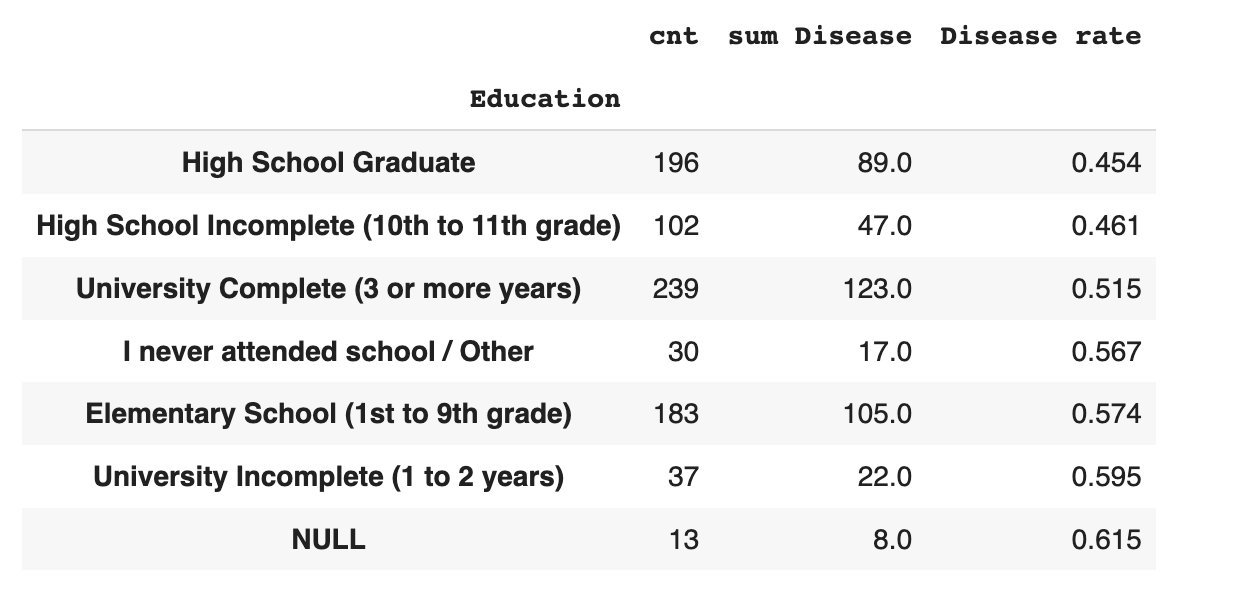
***Region***

******

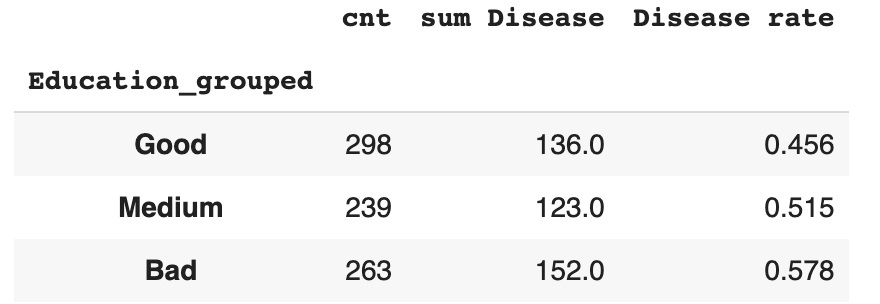
The dimension does not have strong differences in its infection rates and contains few misspellings. The region London is contained in uppercase letters five times together with its normally spelled category, which leads to a merge of both. Even though the Northern regions have slightly lower rates and the West shows slightly higher rates, there is not one region with a substantial association with the disease. Two large groups (Good and Bad) are created to reach a more statistically significant distinction. The more lowly infected North East, North West, and East Midlands are grouped into one region. All other regions are merged together, leading to the binary variable Region\_bad.



***Education***

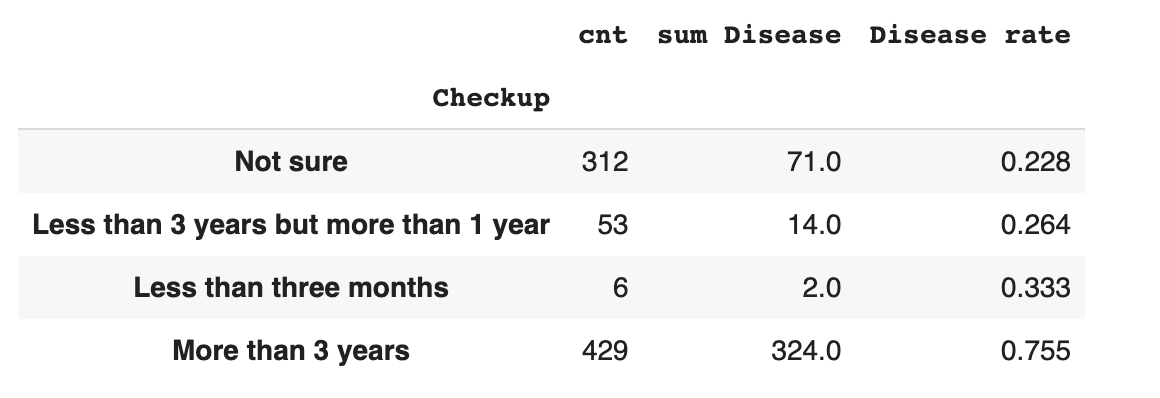


Those who never attended university but attended high school have a lower chance of contracting a disease. *High School Graduate* and *High School Incomplete (10th to 11th grade)* groups are merged into one group (“Good”) as they have very similar disease rates. *University complete (3 or more years)* exhibits a medium infection rate and forms a single category (“Medium”). All other groups (including NaN’s) form the last group, as they are of mostly small quantity and have above-average disease rates.

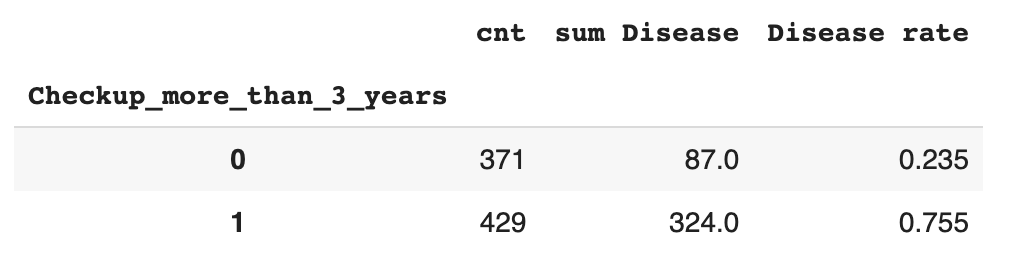


***Health related variables***

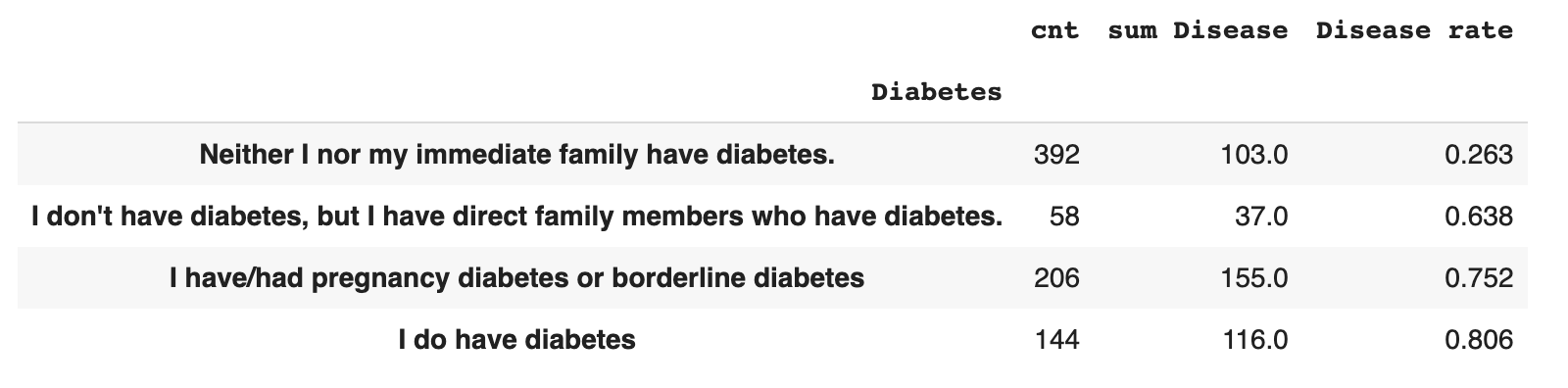
***Checkup***

******

The group *More than 3 years* is substantially worse than others. The other three categories are grouped based on very similar rates and some small sample sizes.

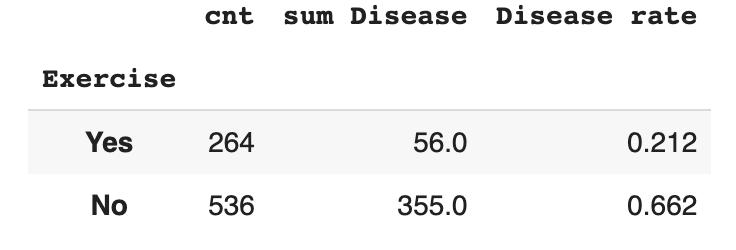


***Diabetes***

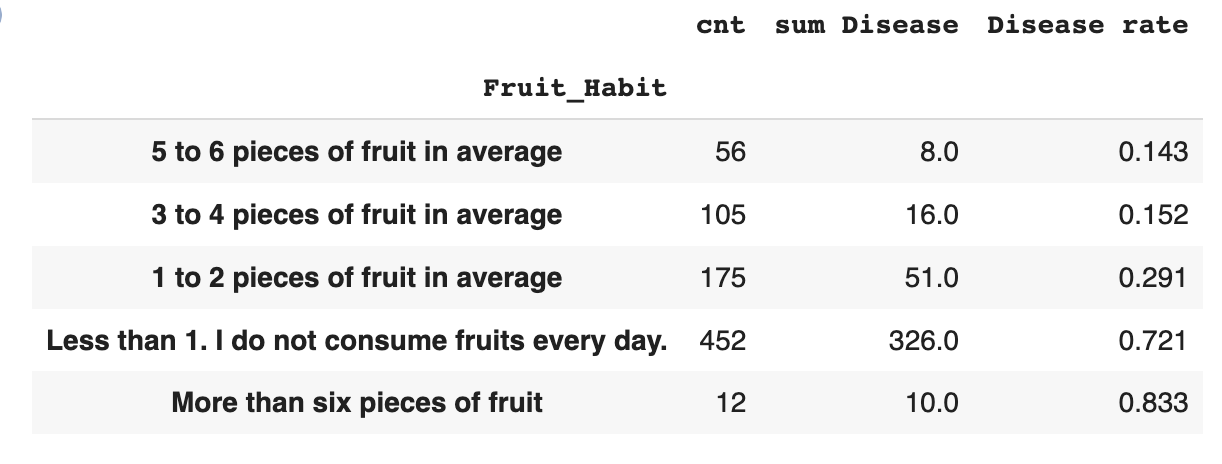


***Habits related variables***

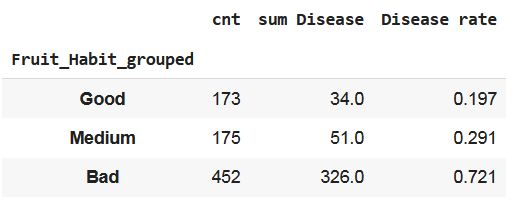
***Exercise***



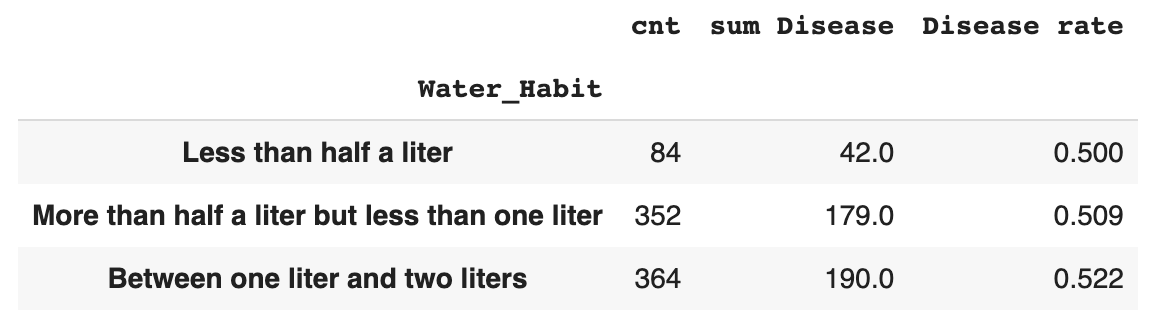
***Fruit\_Habit***



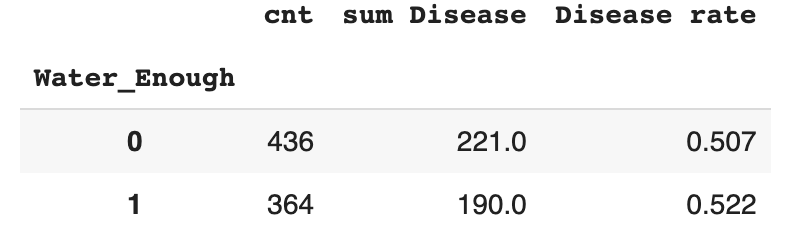
Fruit\_Habitis grouped into three categories to balance classes. It appears that not eating fruits at all increases the risk of having a disease and eating more than two pieces of fruit considerably reduces the risk. Although eating more than six pieces of fruit has the highest disease rate, the number of patients in this group is small and likely a result of sample bias. As it is logically more related to higher fruit consumption, it is grouped with increased fruit intake. The final grouping considers Less than 1 piece of fruit per week (Bad), one to two pieces of fruit per week (Medium) and more than two pieces of fruit per week (Good).



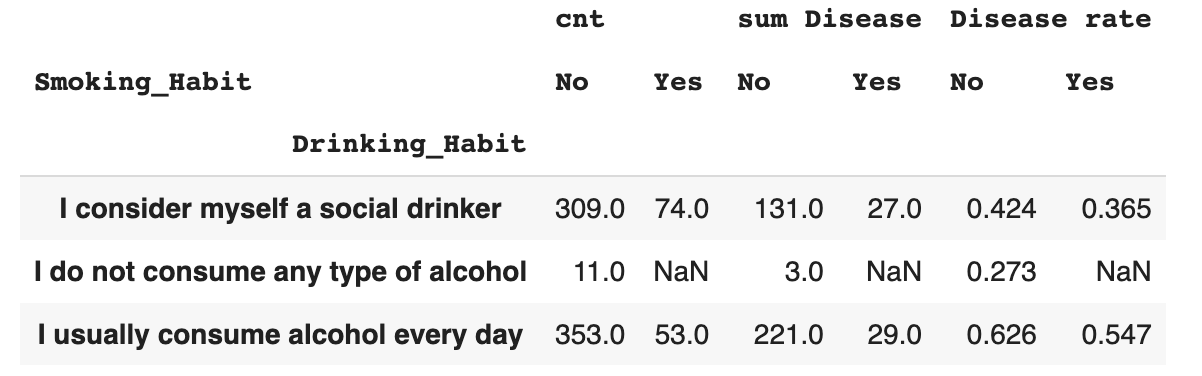
***Water\_Habit***



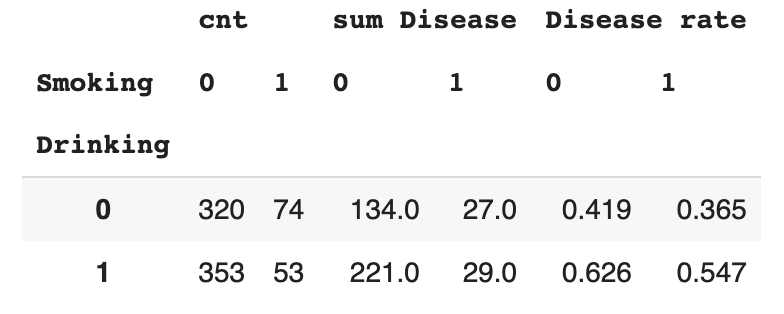
Even though the correlation is weak, we decrease the number of dimensions by grouping the small category *Less than half a liter* with *More than half a liter but less than one liter.* Both categories have very closely fitting disease rates.



***Drinking\_Habit*** and ***Smoking\_Habit***



Being an alcoholic appears to be clearly linked to the disease. There are almost no never drinking people in the sample (11 patients). Those are grouped together with social drinkers based on the close relationship to disease rate. As there are hardly any smokers in the sample, the dependence is not clear. As a result, the smoking habit variable remained unchanged. The distribution of the final transformations can be seen below

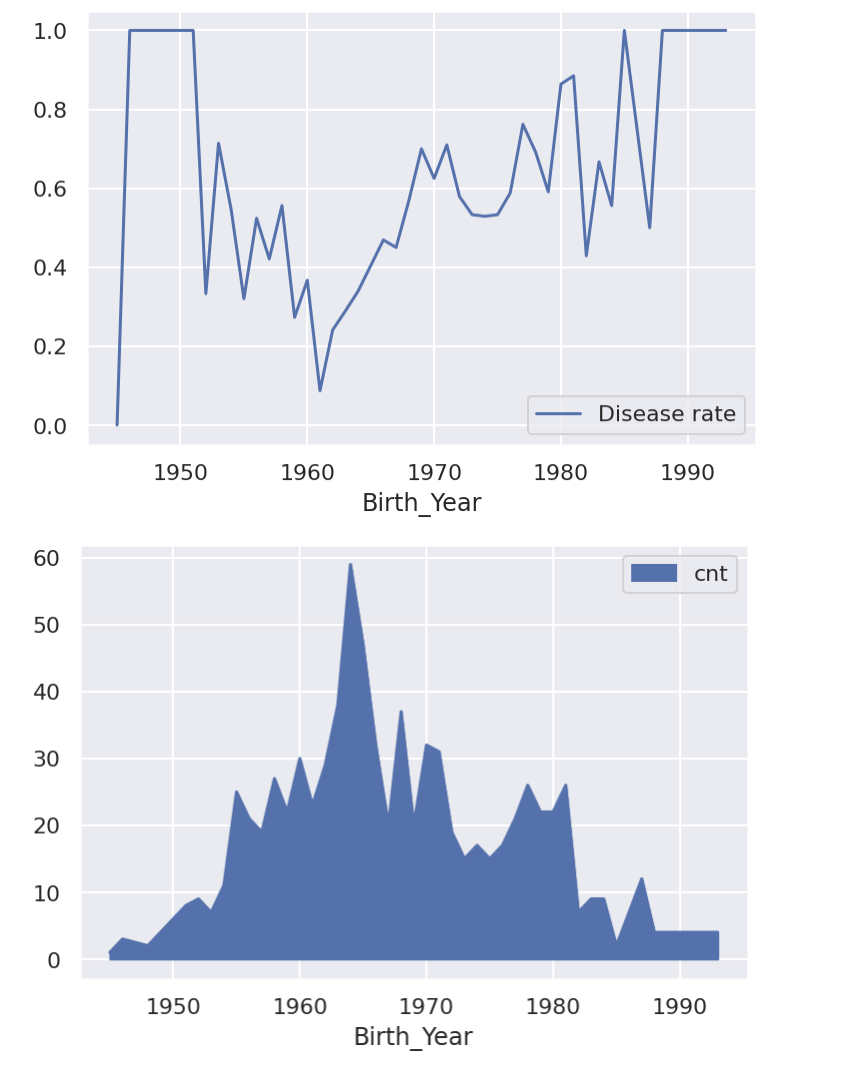
**

*Annex 9.* **Univariate analysis of the raw numerical variables**

***Sociodemographic variables***

***Birth\_Year***

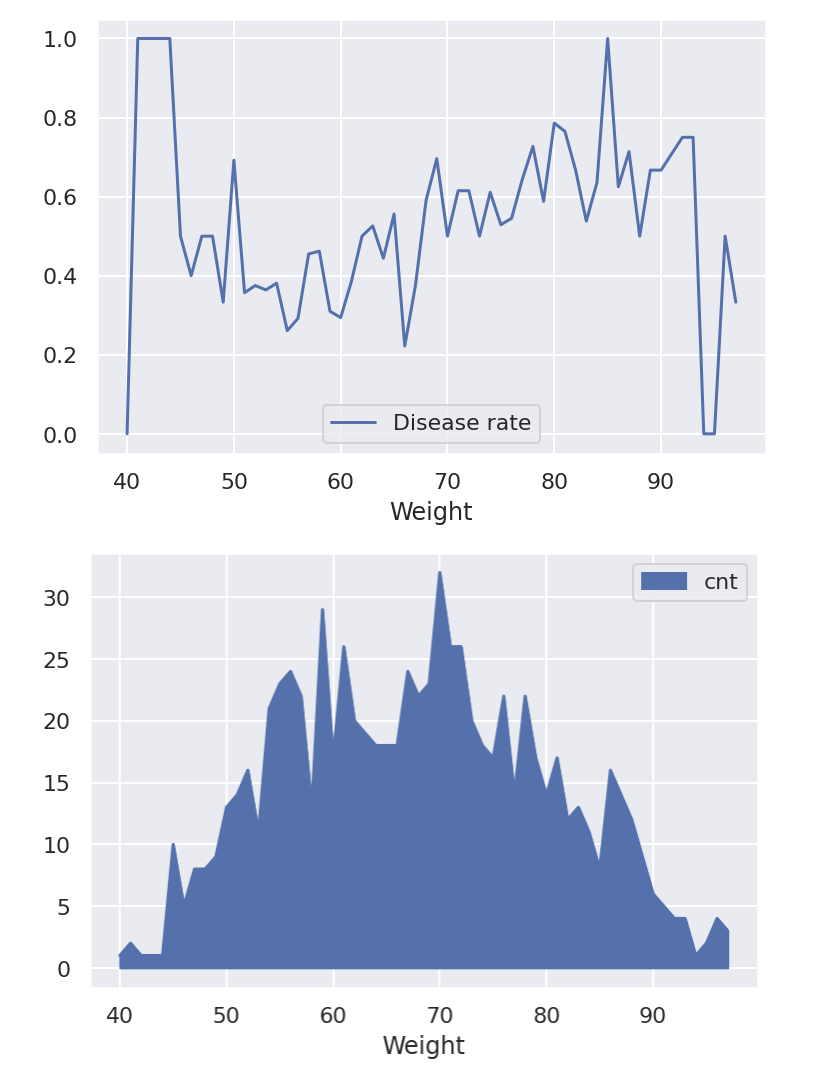
The dependency of birth year and disease rate is strong, but not monotonous. This can be a problem for most algorithms, especially logistic regression and decision trees. Disease rate decreases before 1961 and increases after with almost equal rates. However, as this could be potentially related to confounding factors, the feature was not adjusted for the modeling.

******

***Health related variables***

***Weight***

The dependency is not monotonous as the left part of the distribution (Weight<50) is having higher disease rate as well as the right part (Weight>55). Despite this fact, this variable is kept as it is. The left part of the distribution is having not so many observations to make strong conclusions towards transforming this variable.

******

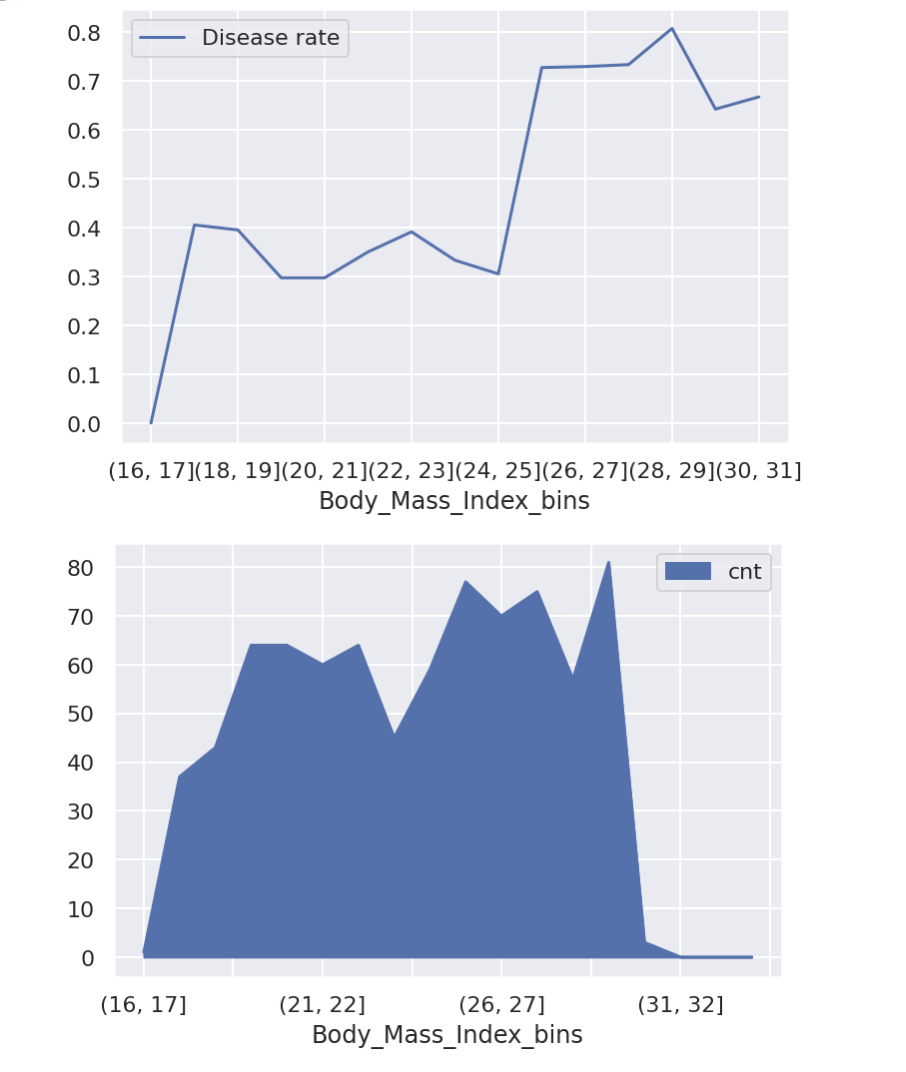
***Height***

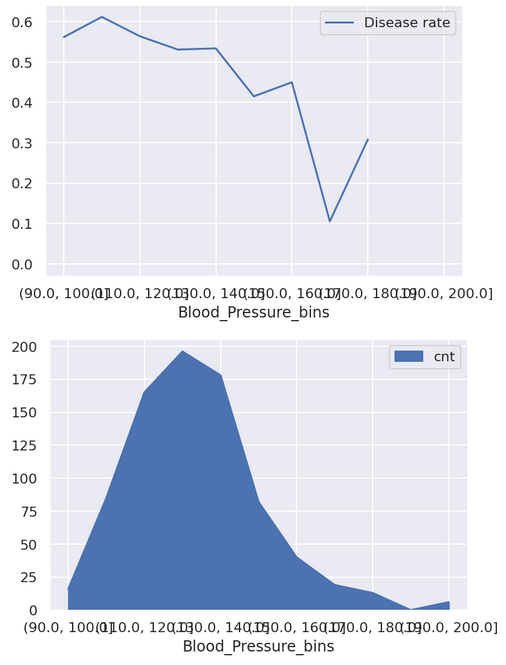
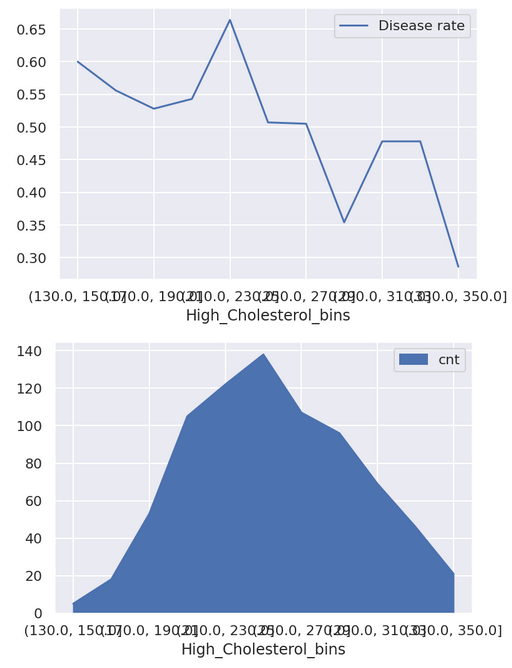
Similar to Birth\_Year, the volatile relationship towards disease rate might be a result of confounding factors. There is no obvious way to treat the data to make the result clearer, which is why the data is not adjusted any further.



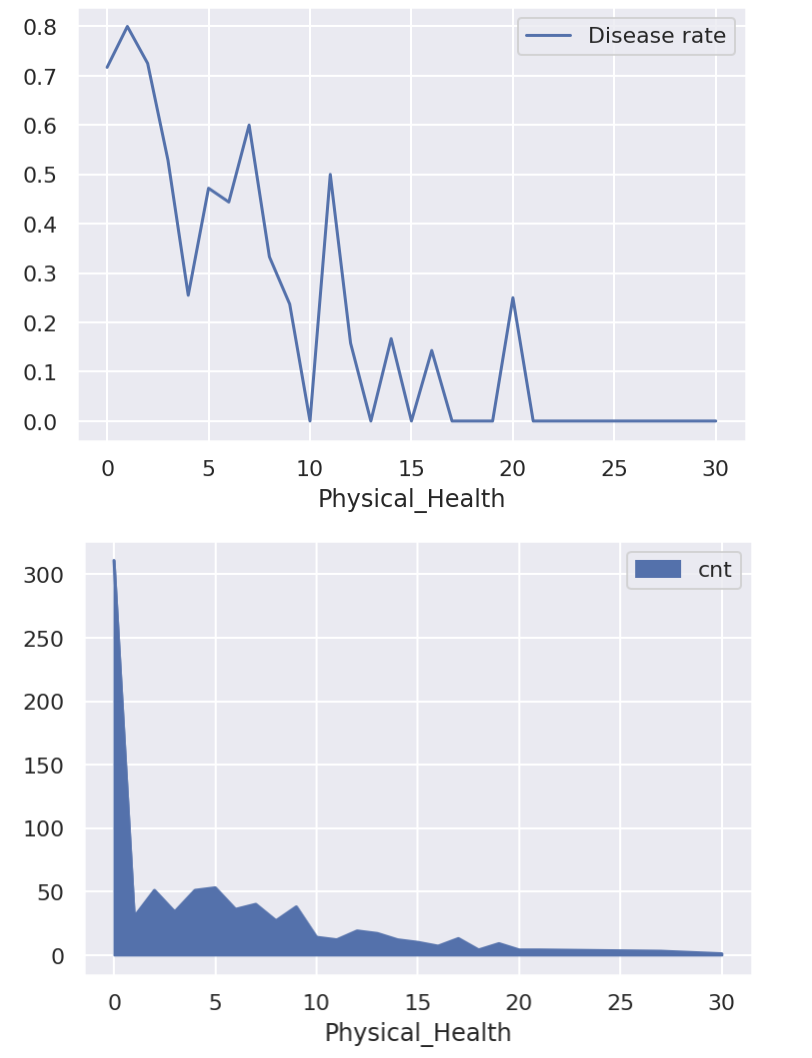
***BMI***

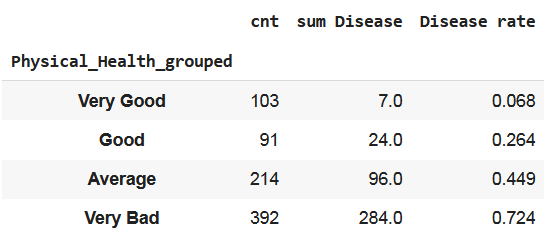
After binning, the BMI variable shows a greater divide after approximately 25. Both before and after this threshold, the relationship towards disease is stable but different. Thus, we resort to making the attribute binary and creating a High\_BMI variable.



***Blood Pressure, Cholesterol & Mental\_Health***

**Physical\_Health.** Physical health also shows strong spikes across its distribution, in combination with being highly skewed. This leads to applying groupings to normalize this behavior. Values of below three, below eight, below twelve, and higher or equal to twelve are grouped separately.

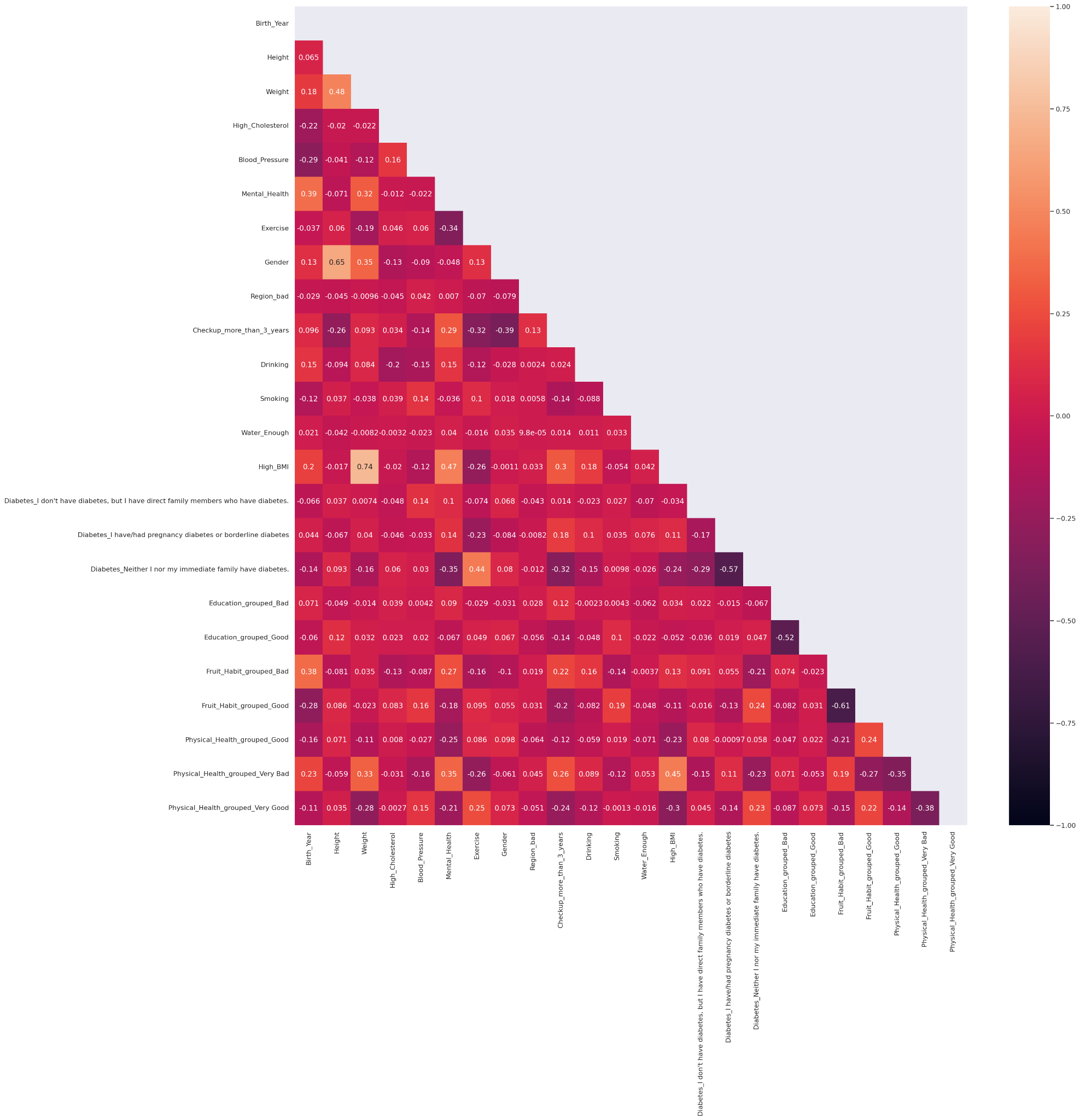
******

******

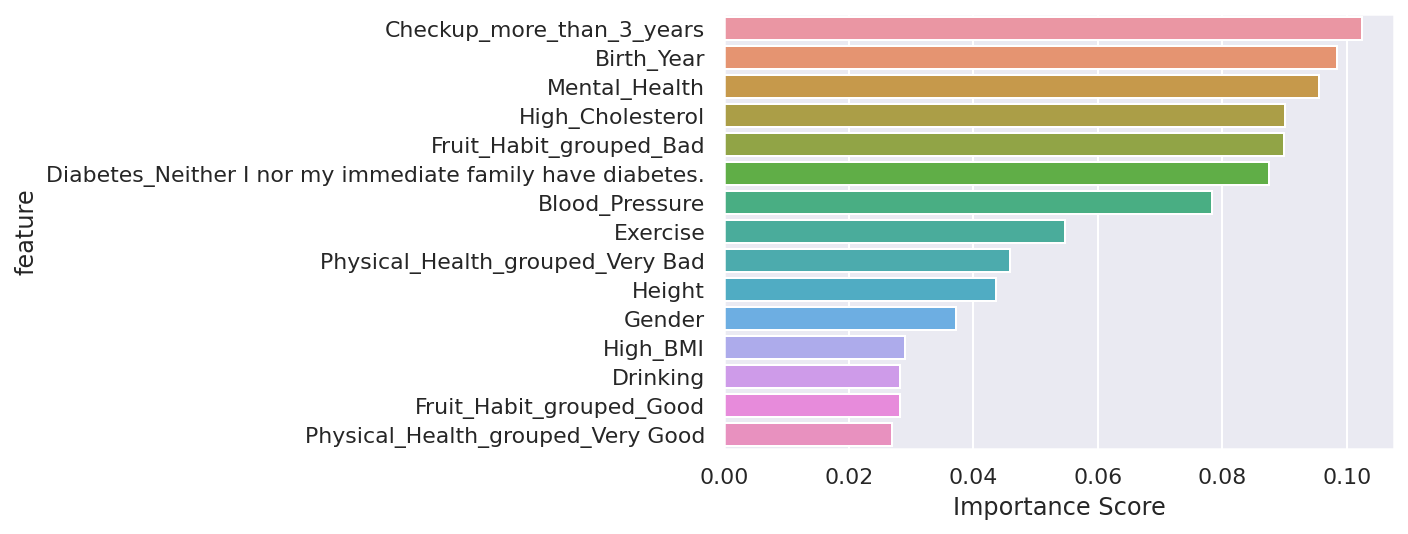
*Annex 10.* **Feature Selection Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Spearman** | **Logistic RFE** | **Chi-square** | **Lasso** |
| Birth\_Year | 1 | 0 | N.a. | 1 |
| Height | 1 | 0 | N.a. | 1 |
| *Weight* | 1 | 0 | N.a. | 1 |
| *High\_Cholesterol* | 1 | 1 | N.a. | 1 |
| Blood\_Pressure | 1 | 1 | N.a. | 1 |
| Mental\_Health | 1 | 1 | N.a. | 1 |
| Exercise | N.a. | 1 | 1 | 1 |
| Gender | N.a. | 1 | 1 | 1 |
| Region\_bad | N.a. | 0 | 1 | 1 |
| Checkup\_more\_than\_3\_years | N.a. | 1 | 1 | 1 |
| Drinking | N.a. | 0 | 1 | 1 |
| Smoking | N.a. | 0 | 1 | 1 |
| *Water\_Enough* | N.a. | 0 | 0 | 0 |
| High\_BMI | N.a. | 1 | 1 | 1 |
| *Diabetes\_I don't have diabetes, but I have direct family members who have diabetes.* | N.a. | 0 | 0 | 1 |
| Diabetes\_I have/had pregnancy diabetes or borderline diabetes. | N.a. | 0 | 1 | 1 |
| Diabetes\_Neither I nor my immediate family have diabetes. | N.a. | 1 | 1 | 1 |
| Education\_grouped\_Bad | N.a. | 0 | 1 | 0 |
| Education\_grouped\_Good | N.a. | 0 | 1 | 1 |
| Fruit\_Habit\_grouped\_Bad | N.a. | 1 | 1 | 1 |
| Fruit\_Habit\_grouped\_Good | N.a. | 0 | 1 | 1 |
| Physical\_Health\_grouped\_Good | N.a. | 0 | 1 | 1 |
| Physical\_Health\_grouped\_Very Bad | N.a. | 0 | 1 | 1 |
| Physical\_Health\_grouped\_Very Good | N.a. | 1 | 1 | 1 |

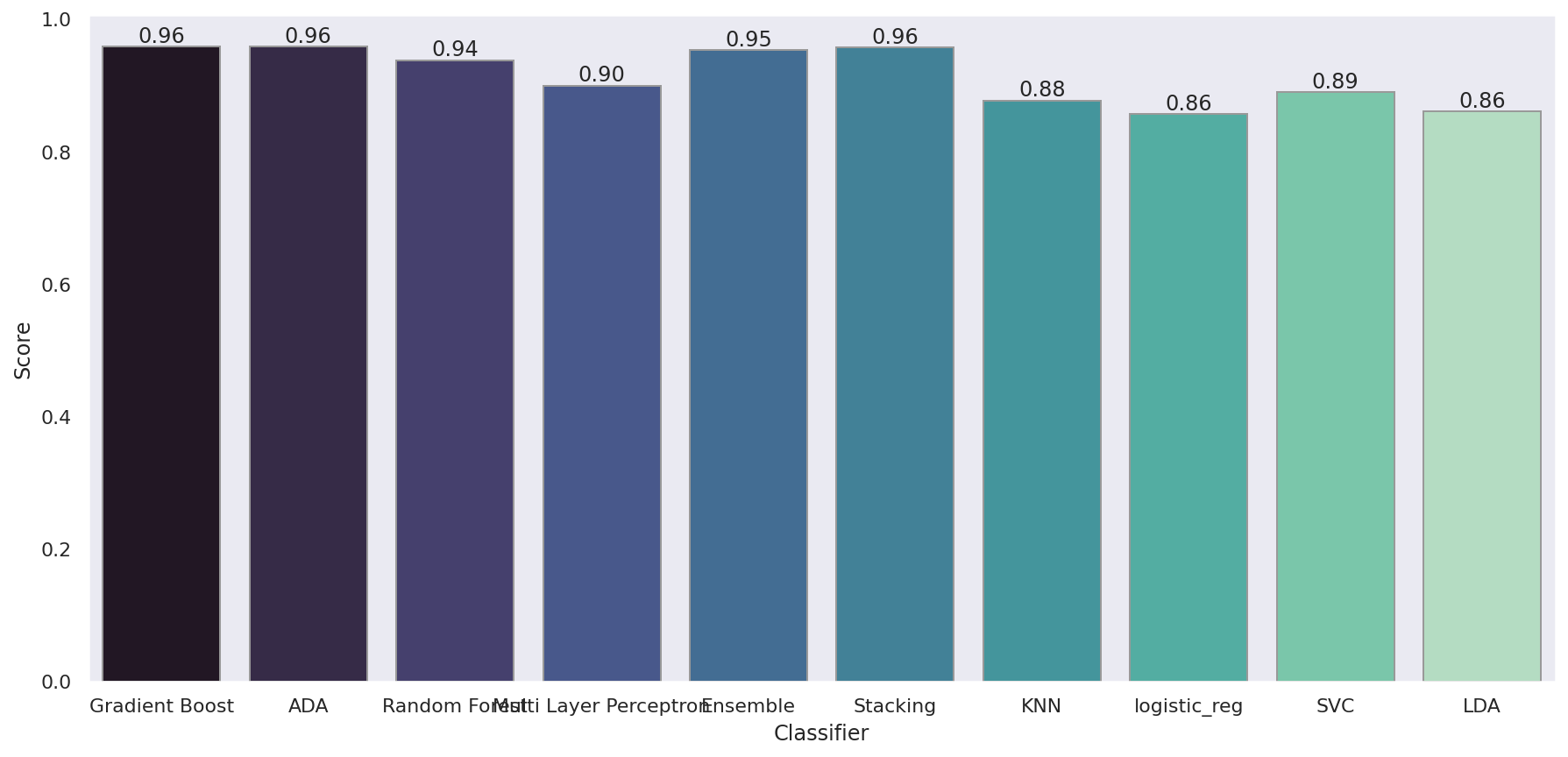
*Annex 11.* **Correlation Matrix**

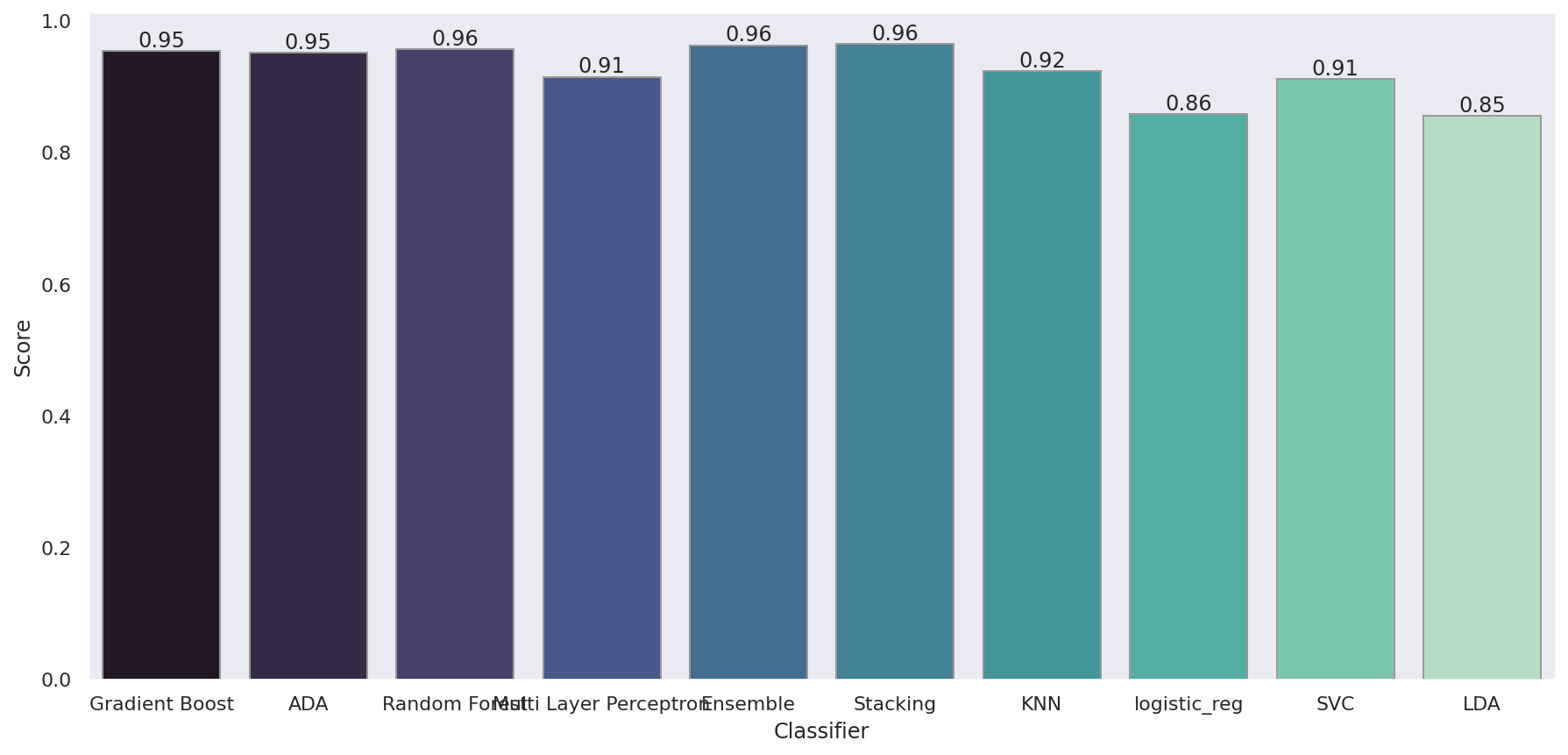


*Annex 12.* **Random Forest Feature Importance**

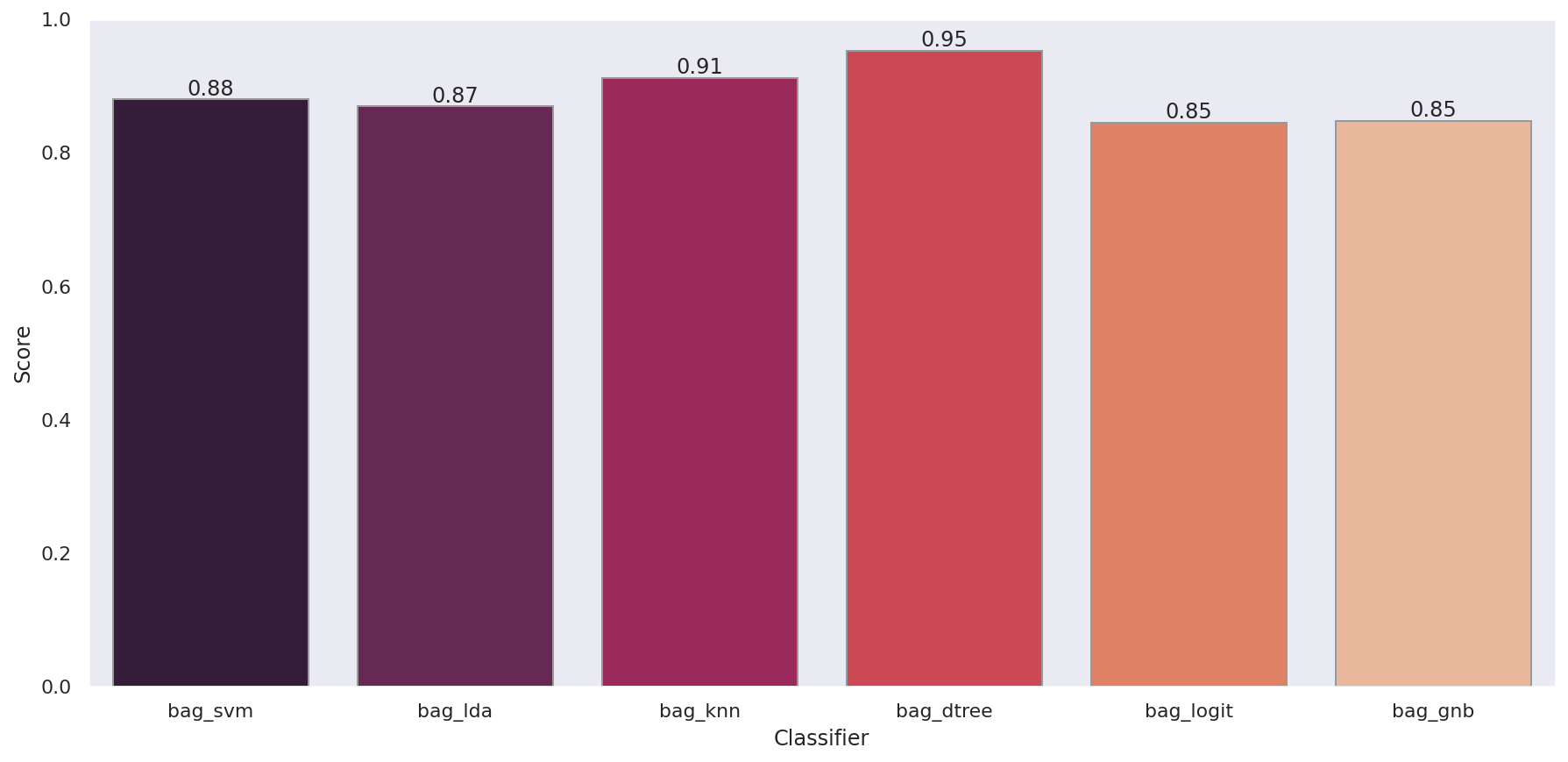


*Annex 13****. Modeling Performance for different datasets***

*Fig 1: Model performance for dataset before feature selection*

**

*Fig 2. Model (with Bagging Classifier) performance for dataset before feature selection*



*Fig 3. Model performance for dataset after feature selection*

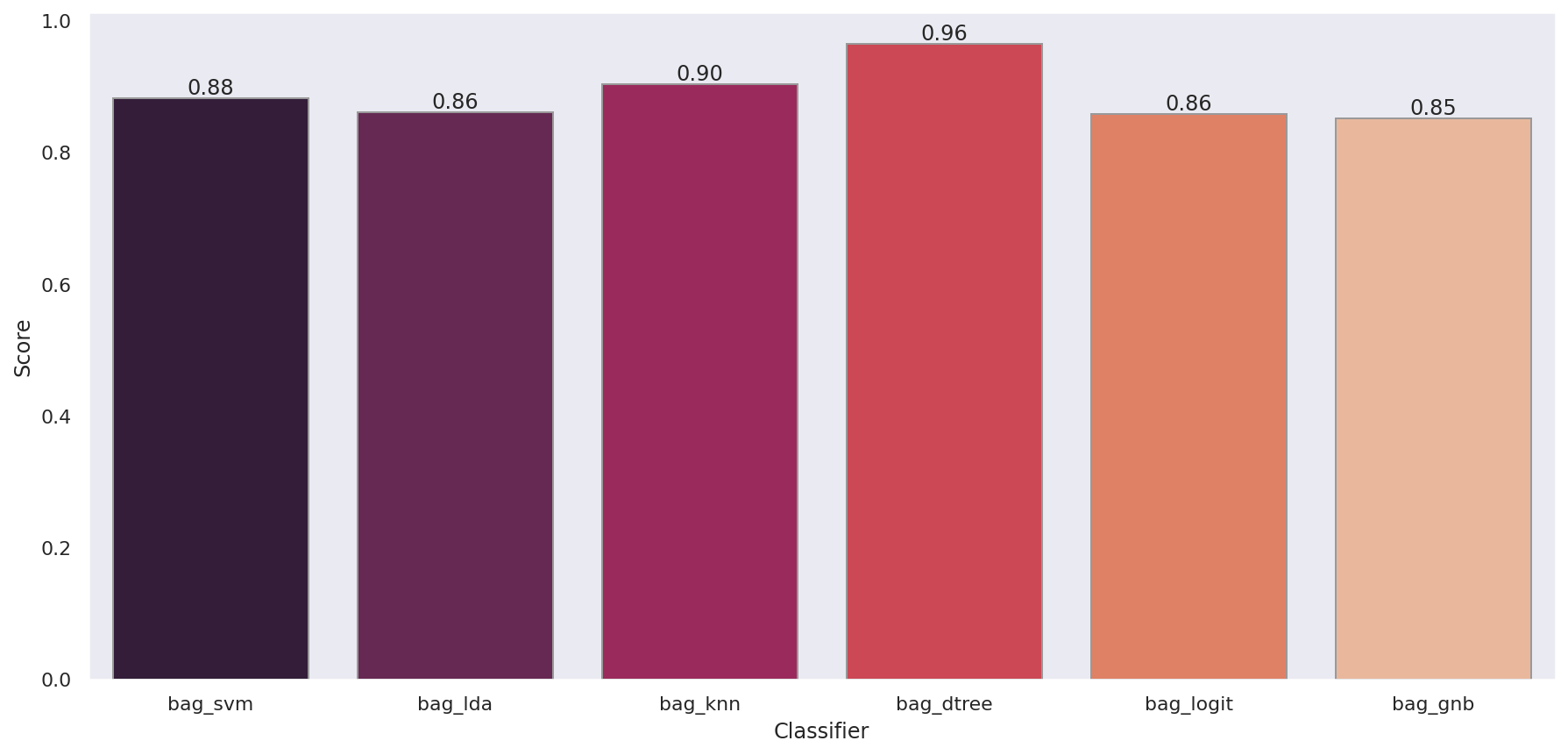
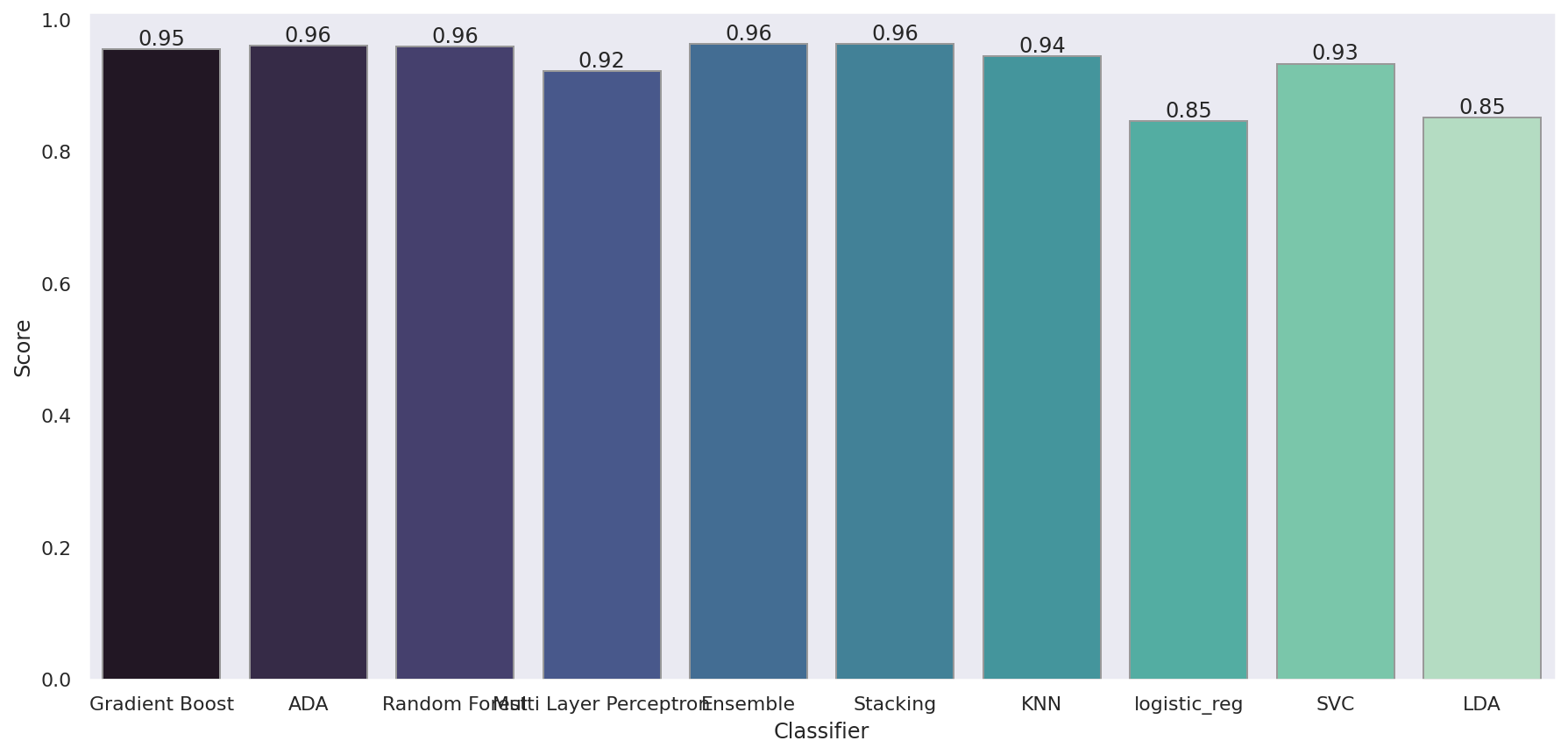
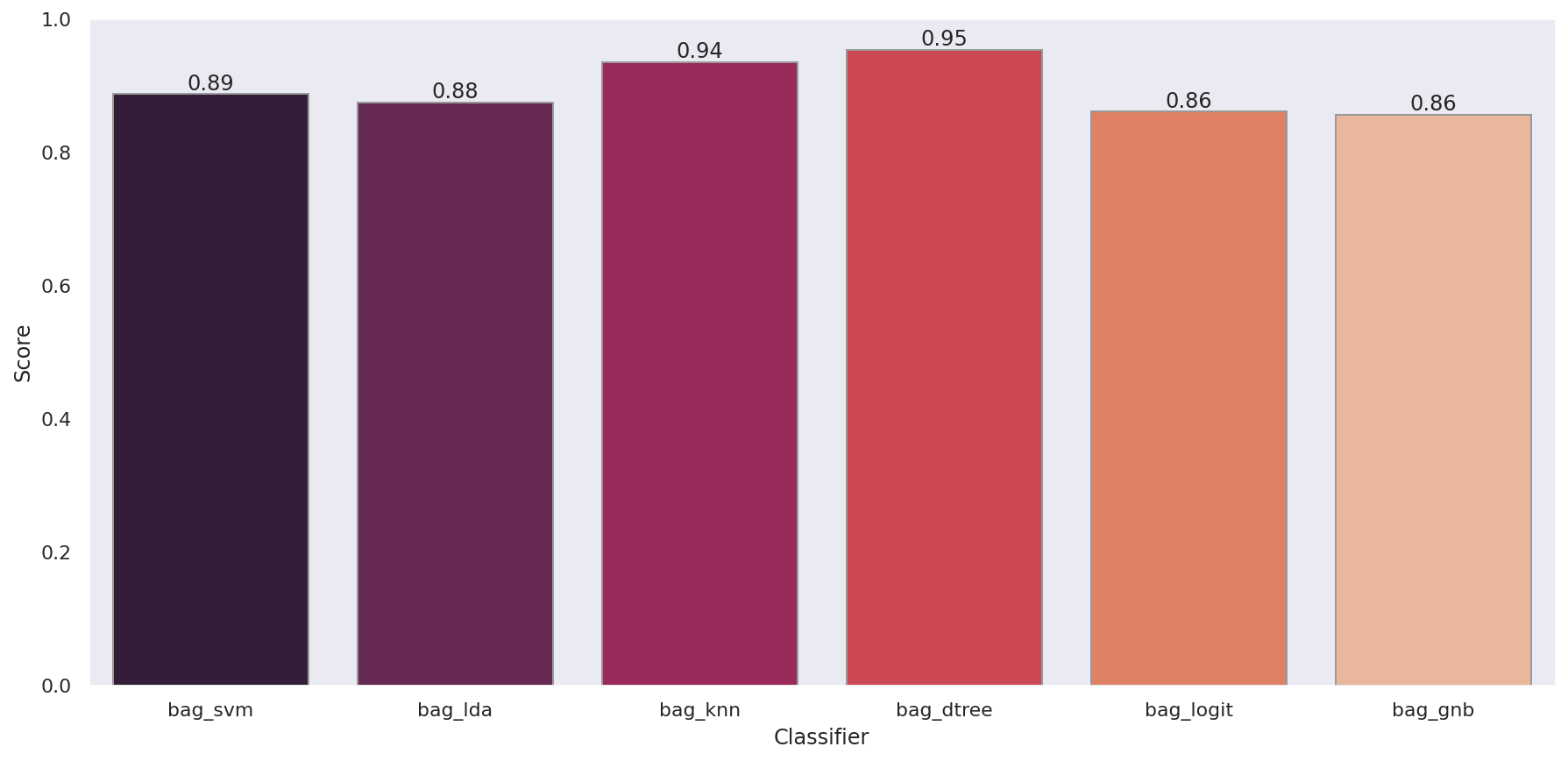


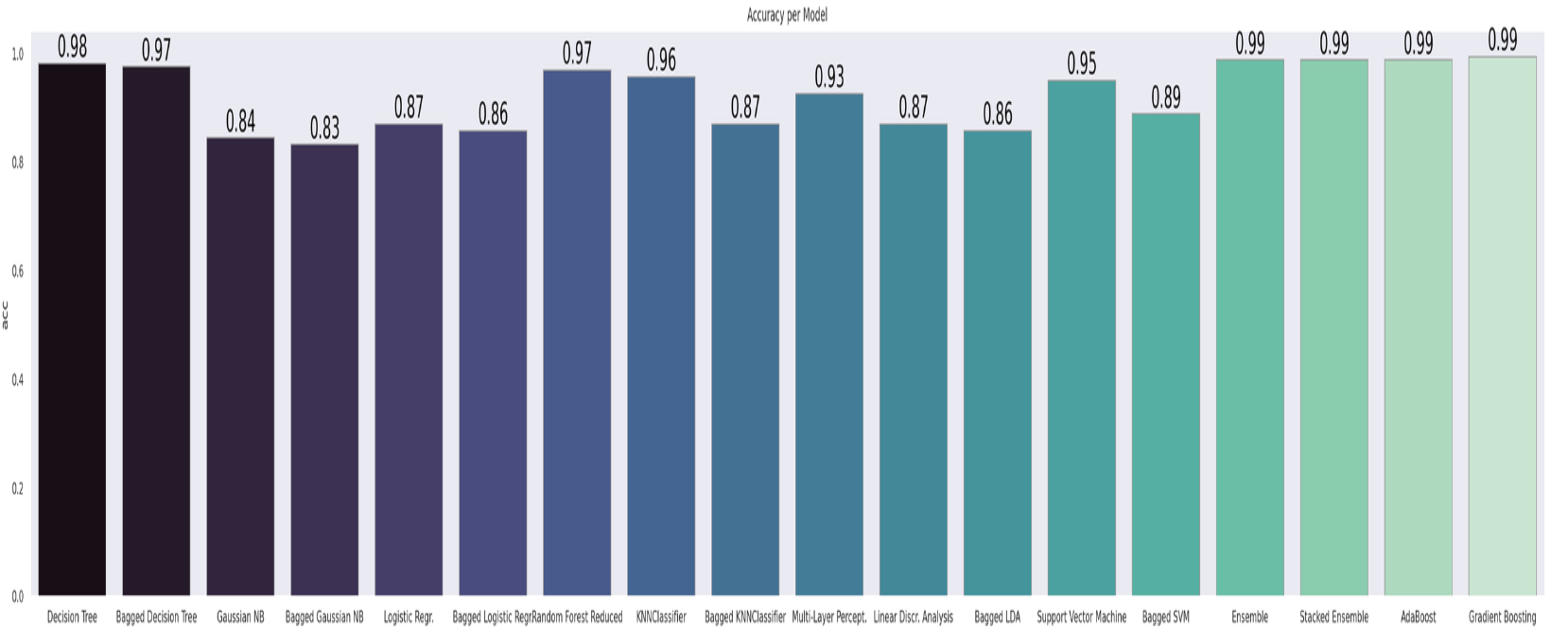
Fig 4. *Model ( with Bagging Classifier) performance for dataset after feature selection*

**

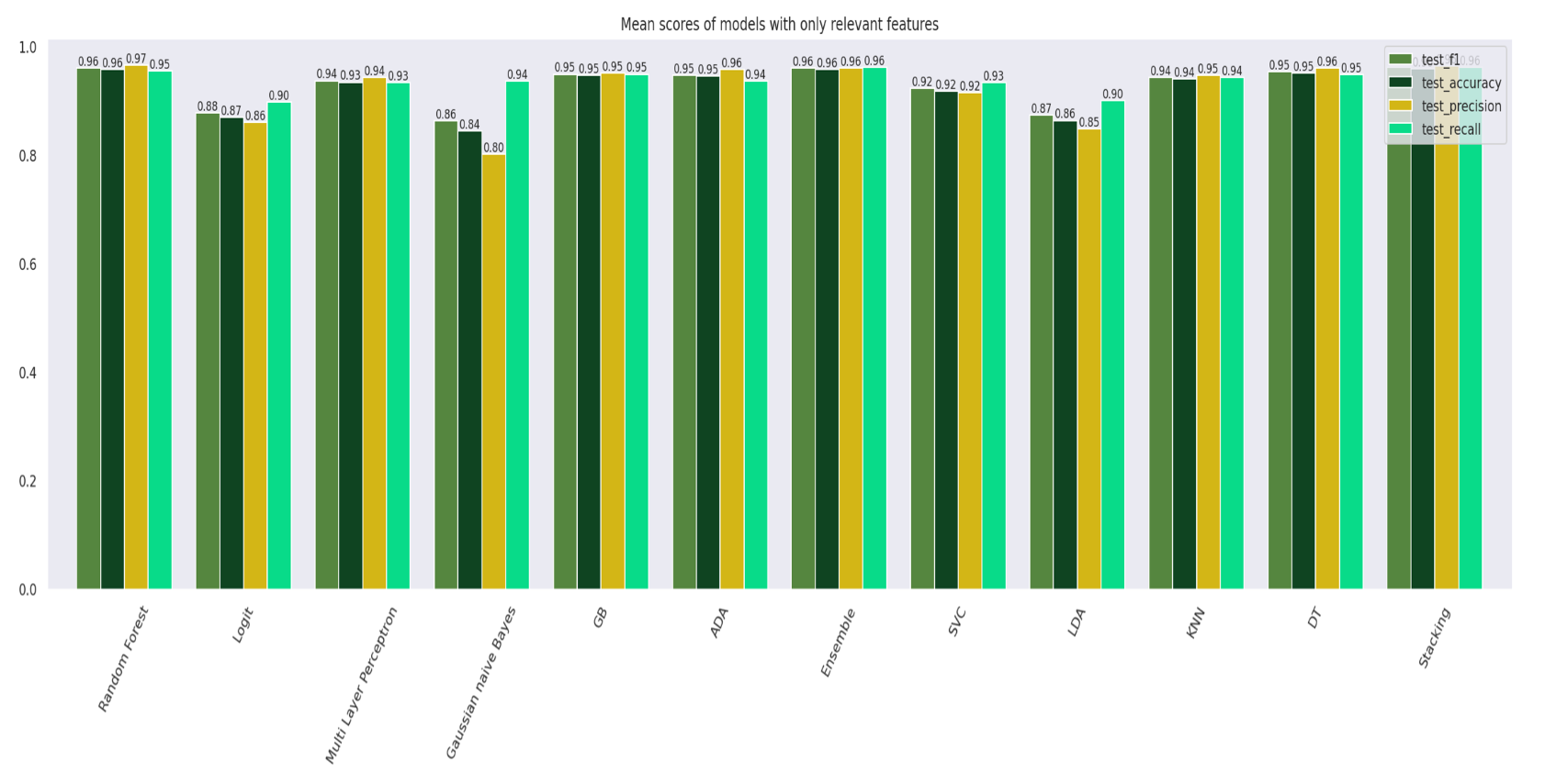
*Fig 5. Model performance for dataset after feature selection and with 15 most important features*

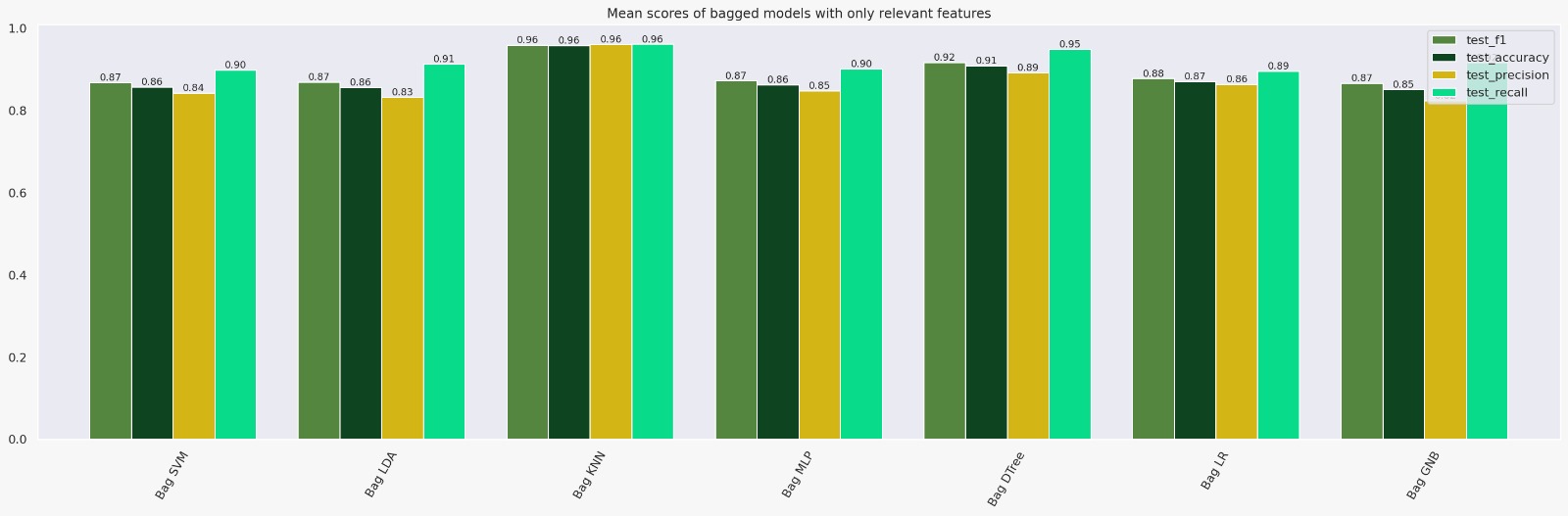
*Fig 6. Model (with Bagging Classifier) performance for dataset after feature selection and with 15 most important features*

**Annex 14** *- First Training Run*

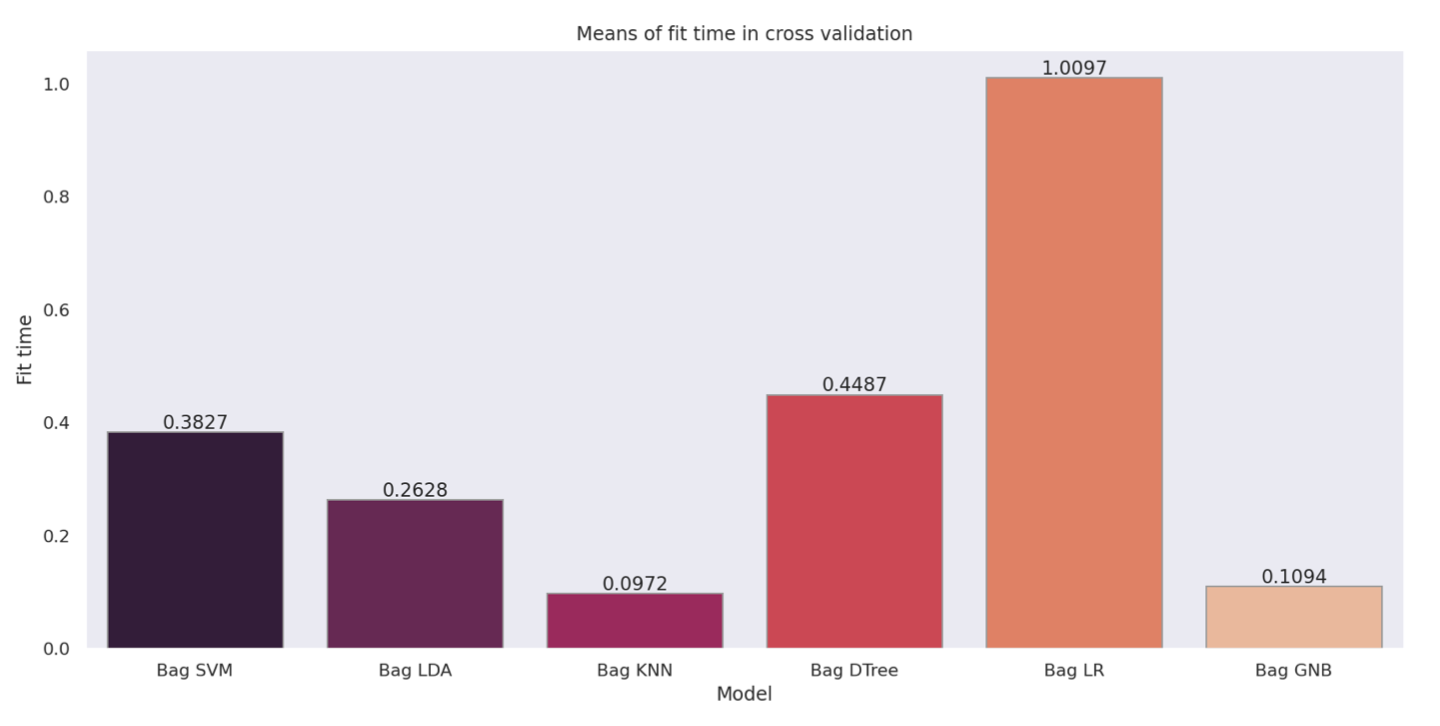
****

**Annex 15** *- Final Assessment on 15 feature*

****

****

*******Figure A*

*******Figure B*

**Annex 16: Hyper Parameters Table**

|  |  |  |
| --- | --- | --- |
| index | Model Name | Hyperparameters |
| 0 | Gradient Boosting | {'learning\_rate': 0.5, 'loss': 'exponential', 'max\_depth': 3, 'n\_estimators': 200} |
| 1 | Ada Boosting | {'base\_estimator': DecisionTreeClassifier(max\_depth=2), 'learning\_rate': 0.5, 'n\_estimators': 300} |
| 2 | Random Forest | {'criterion': 'entropy', 'max\_depth': None, 'max\_features': 'sqrt', 'n\_estimators': 100} |
| 3 | Logistic Regression | {'C': 1, 'penalty': 'l2'} |
| 4 | MLP | {'solver': 'sgd', 'max\_iter': 500, 'learning\_rate\_init': 0.05, 'learning\_rate': 'adaptive', 'hidden\_layer\_sizes': (50, 100, 50), 'alpha': 0.0001, 'activation': 'relu'} |
| 5 | Gaussian NB | {'var\_smoothing': 1.0} |
| 6 | LDA | {'solver': 'svd'} |
| 7 | KNN | {'algorithm': 'ball\_tree', 'n\_neighbors': 1, 'weights': 'uniform'} |
| 8 | Randomized SVM | {'oob\_score': True, 'n\_estimators': 50, 'max\_samples': 0.6, 'max\_features': 0.8, 'bootstrap\_features': False, 'bootstrap': True, 'base\_estimator\_\_kernel': 'rbf', 'base\_estimator\_\_C': 1} |
| 9 | Randomized LDA | {'oob\_score': True, 'n\_estimators': 80, 'max\_samples': 0.7, 'max\_features': 0.4, 'bootstrap\_features': False, 'bootstrap': True, 'base\_estimator\_\_solver': 'svd'} |
| 10 | Randomized KNN | {'oob\_score': False, 'n\_estimators': 150, 'max\_samples': 0.9, 'max\_features': 0.3, 'bootstrap\_features': False, 'bootstrap': True, 'base\_estimator\_\_weights': 'distance', 'base\_estimator\_\_n\_neighbors': 4, 'base\_estimator\_\_algorithm': 'kd\_tree'} |
| 11 | Randomized MLP | {'oob\_score': True, 'n\_estimators': 120, 'max\_samples': 0.1, 'max\_features': 0.7, 'bootstrap\_features': False, 'bootstrap': True, 'base\_estimator\_\_solver': 'adam', 'base\_estimator\_\_max\_iter': 400, 'base\_estimator\_\_learning\_rate\_init': 0.01, 'base\_estimator\_\_learning\_rate': 'constant', 'base\_estimator\_\_hidden\_layer\_sizes': (50, 50, 50), 'base\_estimator\_\_alpha': 0.0001, 'base\_estimator\_\_activation': 'relu'} |
| 12 | Randomized Tree | {'oob\_score': False, 'n\_estimators': 80, 'max\_samples': 0.3, 'max\_features': 0.7, 'bootstrap\_features': False, 'bootstrap': True, 'base\_estimator\_\_max\_depth': 10, 'base\_estimator\_\_criterion': 'gini'} |
| 13 | Randomized Logit | {'oob\_score': False, 'n\_estimators': 100, 'max\_samples': 1.0, 'max\_features': 0.6, 'bootstrap\_features': False, 'bootstrap': True, 'base\_estimator\_\_penalty': 'l2', 'base\_estimator\_\_C': 1} |
| 14 | Randomized Gaussian NB | {'oob\_score': False, 'n\_estimators': 80, 'max\_samples': 0.8, 'max\_features': 0.9, 'bootstrap\_features': True, 'bootstrap': True, 'base\_estimator\_\_var\_smoothing': 4.3287612810830526e-07} |