Analysis of Machine Learning Methods for Heart Disease Prediction

André Igor Nóbrega, Douglas Yoshioka,
Matheus Silva, Renan Borges, Taylla Theodoro
Institute of Computing
State University of Campinas
Campinas, São Paulo
RA: 203758, 196163, 241882, 186447, 219596

***Abstract*—In Brazil, heart disease was responsible for 27% of deaths in 2019. There are several indicators for heart disease, such as gender, age, socioeconomic status, location. Considering that immense amounts of population health data are generated daily, such data can be valuable for the creation of tools that are able to help health professionals to diagnose diseases quickly and reliably. Machine Learning is a statistical inference technique in which a certain task is learned by the computer from training data. Therefore, based on this principle, this work aims to use machine learning methods to predict the probability of a person having heart disease given indicators. For this, the dataset *Key Indicators of Heart Disease* from the Kaggle website was chosen, where 400,000 adults were interviewed in 50 US states about possible indicators of heart disease. The dataset has 319,795 samples - with 18 variables, 17 of which are possible regressors and 1 target variable (existence or not of heart disease) - and was divided into 80% for training, 20% for validation. The dataset was explored and a feature selection was performed. The best dataset was chosen and model by two main approaches, traditional machine Leaning methods (Naive Classifier, Naive Bayes, Logistic Regression, Random Forest) and deep learning method (Fully Connected Neural Network). The models were trained and analysed in a set of metrics, considering that the dataset is imbalance, the most reliable metric is ROC-AUC. The model with the best ROC-AUC result was Random Forest, with 0.81, which we conclude that is the best model for this task and this dataset.**

1. Introduction

In Brazil, cardiovascular diseases were responsible for 27% of deaths, according to data from 2019 (OLIVEIRA et al., 2022). This mortality rate has increased in the last three decades, probably due to the increase and aging of the population. In the country, among the main cardiac comorbidities, there are heart failure, myocardial infarction, atrial fibrillation and arterial hypertension.

According to the 2019 *Global Burden of Disease* Study, there are several indicators for heart disease, such as gender, age, socioeconomic status, location. It could be noted that men are more prone to the disease compared to women. However, both genders have the age group of 50-69 as the one with the highest incidence, followed by the age group over 70 years. As for the location, higher incidences were recorded in the northeastern states, such as Maranhão and Alagoas.

From 2008 to 2019, 8,743,403 clinical and surgical cardiovascular procedures were performed, paid for by SUS (OLIVEIRA et al., 2022). Furthermore, in 2015 alone, cardiovascular diseases had a significant financial impact, leading to a cost of R$56.2 billion for the four main comorbidities (STEVENS et al., 2018).

Thus, it is notable that a good preventive policy against cardiovascular diseases would result in benefits for the health of the population and, consequently, for the country's economy. A possible alternative is to use a good dataset and carry out a preventive study capable of predicting the chances of a person having cardiovascular diseases.

Huge amounts of population health data are generated daily - considering the current scenario of the world and technology - which can be valuable for the creation of tools that are able to help health professionals to diagnose diseases more quickly and reliable, possibly even before they manifest.

Machine Learning is a growing area that develop techniques for a computer to learn to perform a task without explicit instructions. It can be used to recognize patterns in data and predict new data by modelling these patterns, which usually takes a great amount of data.

The learning can be supervised (with the model having access to the ground truth, e.g. linear and Logistic Regression, Support Vector Machines, Random Forest and Neural Network) or unsupervised (when the model extract characteristic all from the data itself, e.g. K-mean, Clustering).

We aim to explore supervised methods to perform the task of predicting the probability that a person has or does not have cardiovascular problems using patient-related data. In order to do so, it will be used the dataset *Key Indicators of Heart Disease* from the Kaggle website - where 400,000 adults were interviewed in 50 US states about possible indicators of heart disease.

A State of art approach, such as (DEEPIKA; BALAJI, 2022), proposes an optimized unsupervised technique for feature selection and novel Multi-Layer Perceptron for En-

hanced Brownian Motion based on Dragonfly Algorithm (MLP-EBMDA) for classification of heart disease.

Our analysis will be perform two main approaches, first using machine learning traditional methods, such as naive classifiers, naive bayes, logistic regression and random forest. Secondly using the a deep learning model, the fully connected neural network.

We also aims to study the data with the least possible bias in relation to the data of the people investigated. In order to avoid ethical injustices, with bias as to race and gender. This is a very important attitude, since raising ethical issues are often ignored when machine learning models are created.

2. Objective

The objective of this project is to create a machine learning model capable of determining the probability that a person has or does not have cardiovascular problems.

2.1. Specific

In order to perform the object, it is also aimed to achieve the specific objectives of this project:

- Learn to analyse and manipulate a dataset in order to perform correct analysis and predictions;
- Learn to choose the best machine learning approach to solve and model a problem;
- Compare different approaches to solve the same problem.

3. Methodology

Firstly the dataset was analysed and curated to perform the model with the best selection of features and to reduce bias.

Secondly, we model the problem with two approches, using machine learning traditional methods: naive classifier, naive Bayes, logistic regression, random forest. Secondly using the a deep learning model, the fully connected neural network.

The naive classifier was chosen as the baseline due to its simplicity.

3.1. Materials

The data is called *Personal key Indicators of Heart Disease* and was obtained for carrying out the prediction of heart diseases was made available by the *CDC* (*Center of Disease Control*). According to the center itself, heart disease represents one of the leading causes of death for individuals in the United States. This data was collected through a CDC behavioral risk monitoring system, and subsequently post-processed and made available for analysis and inferences in Kaggle through the following https://www.kaggle.com/ datasets/kamilpytlak/personal-key-indicators-of-heart-diseaselink (PYTLAK, 2022).

This monitoring system interviews more than 400,000 adults every year across 50 US states, making it the largest continuous health data collection system in the world (PYTLAK, 2022). The most recent data, which we used throughout this work, has 319,795 samples and 18 fields, with 17 possible regressors and 1 target variable. Before we can propose processing strategies and a *baseline* model, it is necessary to carry out a brief analysis of these data. This is shown in the following subsections, in which we present the general aspects of *dataset*, we show a univariate analysis of the categorical variables and, finally, a multivariate analysis using the *target*.

Variable	Type
HeartDisease	Categorical (binary)
BMI	Numerical
Smoking	Categorical (binary)
AlcoholDrinking	Categorical (binary)
Stroke	Categorical (binary)
PhysicalHealth	Numerical
MentalHealth	Numerical
DiffWalking	Categorical (binary)
Sex	Categorical (binary)
AgeCategory	Categorical
Race	Categorical
Diabetic	Categorical (binary)
PhysicalActivity	Categorical (binary)
GenHealth	Categorical
SleepTime	Numerical
Asthma	Categorical (binary)
KidneyDisease	Categorical (binary)
SkinCancer	Categorical (binary)

Table 1: Dataset Features Description.

3.1.1. Univariate Analysis of Categorical Variables. Intuitively, we can imagine that many of the categorical variables must have a strong correlation with the onset of heart disease. The class of smokers, for example, is one of these variables, as we can imagine that such people are more likely to have heart complications.

Thus, it becomes important to analyze the distribution of the main categorical variables through pie charts. We mainly want to identify possible imbalances between the classes - both in the target variable and in the regressors. Figure 1 below illustrates these distributions.

Important considerations:

- Most of the variables present unbalance between classes. This is expected due to the fact that the distribution of cardiac patients, for example, in society is unbalanced;
- It is necessary to deal with the imbalance of the target variable in the construction of the model. Among the techniques to deal with this we can mention:
  - Use of metrics like precision, recall and f1-score instead of just accuracy;
  - *Resampling* of the data in order to increase the frequency of the ”No” class;
  - Severe error penalty in the lowest frequency class;

- We can see that there is also a great imbalance between the race of the individuals who participated in the research. Thus, it is important to highlight that models built using this database inherently run the risk of presenting high bias in relation to this field. This needs to be discussed from an ethical point of view.

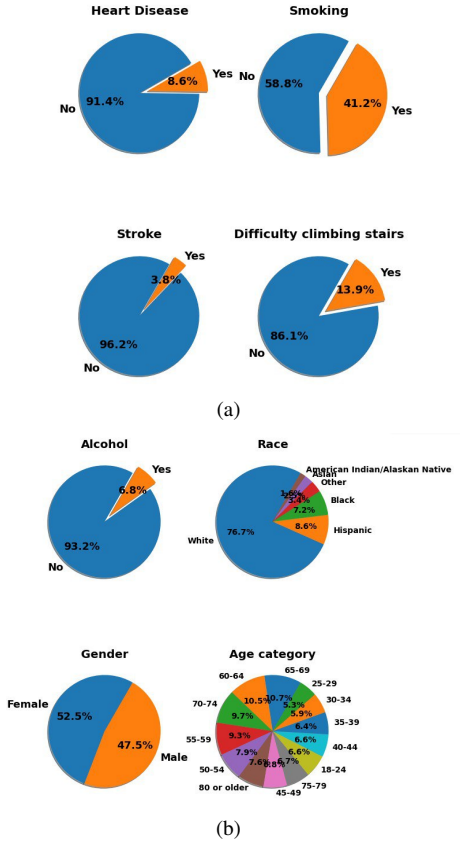


Figure 1: Distribution of categorical variables.

3.1.2. Multivariate analysis. After a brief analysis of the dataset, we can start looking for relationships between the fields and the target variable. At this point, the main point of view will be qualitative, that is, we seek to infer through graphs how the attributes of the dataset can affect whether or not someone has heart disease.

First, we present the relationship between age, sex and the probability of developing heart disease. Figure 2 below shows a bar graph grouped by gender and age group. The vertical red dotted line indicates the overall frequency in the *dataset* of people who have the disease, as shown in the upper left graph of Figure 1. We can see that there is a strong relationship between age and the onset of heart problems. Young people between the ages of 18 and 50 have a rare incidence of the disease. Around the age of 55-59, among men, there is an average incidence. At older ages, the probability of having the disease is almost 30 times greater. It is also possible to notice a strong influence of the genre. There is not a great gender imbalance, as seen in Figure 1, however, it is apparent that the disease is much more common in men. This discrepancy is even more pronounced in older age groups.

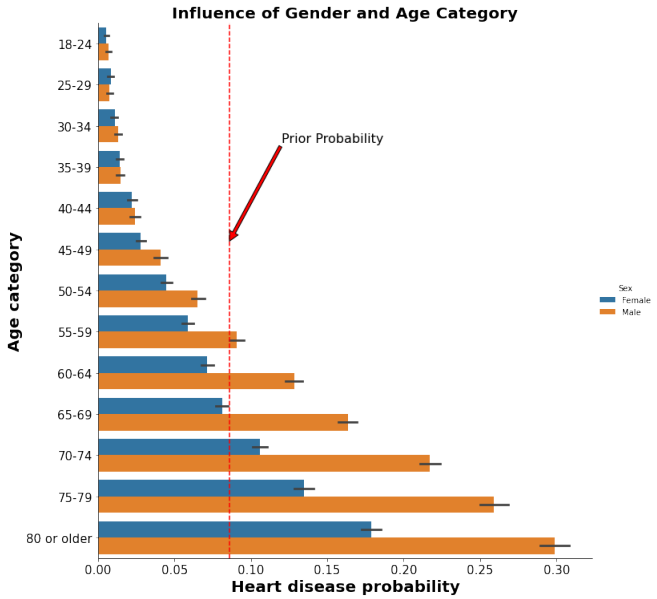


Figure 2: Influence of gender and age

We then analyzed the relationship between smoking, alcohol consumption and physical difficulty in climbing stairs with the target variable. Figure 3 below has grouped bar graphs that illustrate this.

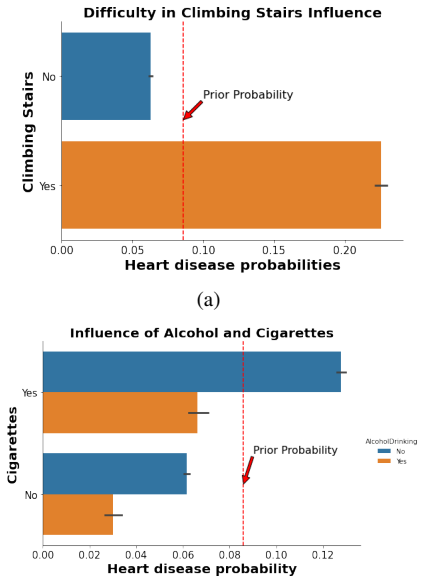


Figure 3: Influence of smoking, alcoholism and physical difficulties.

Again, we can see a great predictive potential in these

variables. In the first graph (on the left), for example, we observe a large difference in the distribution according to the difficulty that individuals have in climbing stairs. Despite this, we cannot indicate that there is causality of this *feature* with the target. Cigarettes and alcohol also negatively affect a person's likelihood of having heart disease. We can see that people who consume a lot of alcohol (whether smoker or not) were almost twice as likely to have heart disease.

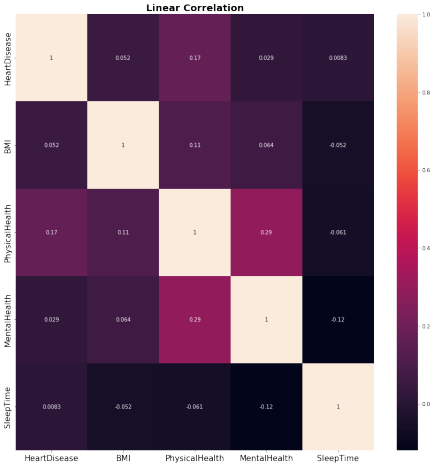
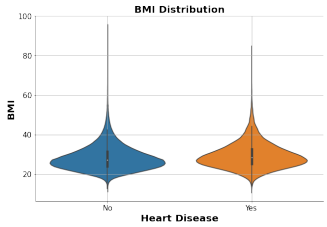
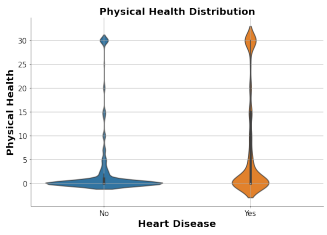


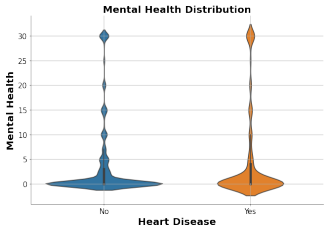
Figure 4: Correlation between numeric variables.



(a) BMI.



(b) Physical health.



(c) Mental health.

Figure 5: Distribution of variables according to heart disease.

In addition to the categorical variables, it is also important to analyze the relationship between the target and the numerical variables. For this, we can use linear correlation, whose calculation was performed from the transformation of the target variable into a binary: 0 indicates false and 1 true. Figure 4 shows the result. From the graph we can see that there is not a large linear correlation between the target and any numerical variable. The highest value was given with the field *PhysicalHealth*, which indicates how many days the individual felt physically ill in the last 30 days. While we might intuitively think that there is a relationship between these variables and having heart disease, we can conclude that this relationship is not linear in nature. This indicates that we would need to apply some kind of nonlinearity to the input of these variables if we wanted to use them as regressors.

Finally, in Figure 5, we analyze the distribution of numerical variables in relation to heart disease. We can see, as well as in the correlation, that there is not a big difference in the distributions for the two classes. This is particularly more noticeable in the IMC, where the distributions are almost similar.

3.2. Pipeline

The Figure 6 present the analysis pipeline proposed for this paper, which will be individually described below.

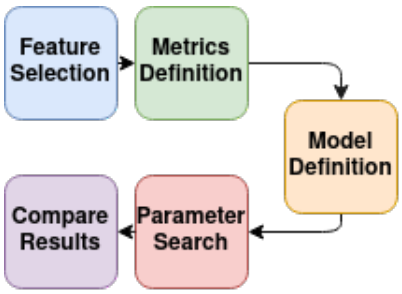


Figure 6: Analysis Pipeline.

3.2.1. Feature Selection. It was performed two different approaches to select the best set of features for the model and predictions.

Fist it was done a qualitative and intuitive approach, by analysing the dataset (Section 3.1), it was handly picked different sets of categorical features that presented greater impact on the label. As curious fact, the more uncorrelated they were, the better.

Secondly, it was performed K-Best-Features module from Sklearn, which is an automatic procedure that yields the subsets of K-Features that most correlate with the target.

3.2.2. Metrics Definition. The definition of the correct metric is an important step to avoid misleading results and a reliable report of analysis. Considering this, we select two metrics that can provide a significant analysis of our problem.

Fist we chose to work with the confusion matrix that provides insights insight into the model performance, as well as shows what kind of classes are being mislabeled. Also it is used a model from Sklean called classification_report from Sklearn to implement sensitive metrics, such as precision, recall and f1-score. Sensitivity refers to the true positive rate and summarizes how well the positive class was predicted. This is particularly important since there are few positive examples in our dataset.

Lastly, we defined the use of the ROC Curve, which is a Ranking Metric, that measures how effectively the model is separating the classes.

A ROC curve is a diagnostic plot for summarizing the behavior of a model by calculating the false positive rate and true positive rate for a set of predictions by the model under different thresholds.

The area under a ROC curve can evaluate the performance of the model. A no-skill classifier should give a ROC_AUC of 0.5, while a perfect classifier should give a result of 1.0.

3.2.3. Model Definition. The model definition was divided into two main approaches.

First we used traditional machine learning models and analysis, such as naive classifiers, naive bayes, logistic regression and random forest.

Secondly, it was performed a deep learning model, a fully connected layer neural network.

All will be detailed in Section 3.3.

3.2.4. Parameters Search. In order to define the which are the best parameters for our models, experiment were performed with the different datasets built, as well with different models with different parameters.

For each of these, with almost out-of-box parameters, we evaluate on the 7 different datasets, choosing the one that yields the best overall ROC AUC score.

For the best models, the hyperparameters were varied and the best one was chosen.

3.2.5. Compare Results. The results of each model will be presented and analysed for the chosen metrics.

3.3. Models

It is presented the models performed for the task. The first 5 are the traditional machine learning approach and the last one for the the deep learning approach.

3.3.1. Baseline. Two no-skill classifiers are defined as baseline model for future comparison. These are Naive Classifiers and take into consideration only the Prior Probability of the Heart Diseases distribution.

The first naive classifier predicts all zeros, since more than 90% of samples have this label in the dataset. The idea is to show how this can produce a a high accuracy, but fail in the better defined metrics, because it does not describe the task.

The second naive classifier predicts zeros with a probability of 91.4% and predicts ones with a probability of 8.6%. This is exactly the Prior Distribution. We expect this second approach to have a lower accuracy, but perform a little better in the remaining metrics.

3.3.2. Naive Bayes. It is a probabilistic model based on Bayes Theorem.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

Using Bayes theorem, we can find the probability of A happening, given that B has occurred. B is the evidence and A is the hypothesis and it is assumed that the features are independent.

There are several types of Naive Bayers models, based on different distributions of the data. We choose to perform a Gaussian Naive Bayes Classifier, considering that features take up a continuous value and are not discrete.

3.3.3. Logistic Regression. It was performed two Binary Logistic Regression, the standard one and with Class Weighting, considering our dataset has a imbalanced distributions of classes. A grid search was performed for the hyperparameter definition.

3.3.4. Random Forest. Random forest is an ensemble algorithm that fits multiple models on different randomly subsets of features used in each data sample of a training dataset, then combines the predictions from all models.

Considering that our dataset is imbalanced, it was 4 variations of the random forest model: Standard, with Class Weighting, with Bootstrap Class Weighting, with Random Undersampling.

3.3.5. Fully Connected Neural Network. is when all layers are Fully Connected, thus all inputs are connected to every activation unit of the next layer.

We will use a simple model, define with 1 hidden layer (96 activation units) with drop out and an output layer (16 activation). The model has 113 trainable parameters.

It was used binary cross entropy loss and Adam optimizer with a learning rate of $1e^{-4}$.

4. Results and Discussion

The results are presented considering all analysis and metrics used.

4.1. Feature Selection

It was selected 7 different dataset to be analysed searching for the best features for the models.

- **qualitative_features1** AgeCategory, Sex, Smoking, AlcoholDrinking, DiffWalking;
- **qualitative_features2** AgeCategory, SkinCancer, GenHealth, DiffWalking, AlcoholDrinking;

- **qualitative_features3** AgeCategory, KidneyDisease, Stroke, DiffWalking, Diabetic;
- **quantitative_features_K12** BMI, Smoking, Stroke, PhysicalHealth, DiffWalking, Sex, AgeCategory, Diabetic, PhysicalActivity, Asthma, KidneyDisease, SkinCancer;
- **quantitative_features_K3** AgeCategory, Stroke, DiffWalking;
- **quantitative_features_K5** Stroke, PhysicalHealth, DiffWalking, AgeCategory, Diabetic;
- **quantitative_features_K8** Smoking, Stroke, PhysicalHealth, DiffWalking, AgeCategory, Diabetic, PhysicalActivity, KidneyDisease.

These datasets were trained with the traditional machine learning models and the results were:

Table 2: Qualitative Dataset 1

qualitative_df1	Acc	Precision	Recall	AUC
LogisticRegression	0.91	0.42	0.02	0.51
LogisticRegression class_weight	0.69	0.18	0.75	0.72
RandomForest	0.91	0.0	0.0	0.5
RandomForest class_weight	0.68	0.18	0.77	0.72
RandomForest class_weight subsample	0.68	0.18	0.77	0.72
BalancedRandomForest	0.68	0.18	0.77	0.72
GaussianNB	0.86	0.25	0.31	0.61
MLPClassifier	0.91	0.0	0.0	0.5

Table 3: Qualitative Dataset 2

qualitative_df2	Acc	Precision	Recall	AUC
LogisticRegression	0.91	0.42	0.02	0.5
LogisticRegression class_weight	0.68	0.17	0.75	0.71
RandomForest	0.91	0.5	0.015	0.5
RandomForest class_weight	0.69	0.18	0.80	0.74
RandomForest class_weight subsample	0.68	0.18	0.80	0.74
BalancedRandomForest	0.68	0.18	0.80	0.74
GaussianNB	0.82	0.22	0.42	0.64
MLPClassifier	0.91	0.0	0.0	0.5

Table 4: Qualitative Dataset 3.

qualitative_df3	Acc	Precision	Recall	AUC
LogisticRegression	0.91	0.47	0.06	0.52
LogisticRegression class_weight	0.70	0.18	0.76	0.72
RandomForest	0.91	0.48	0.03	0.51
RandomForest class_weight	0.69	0.18	0.78	0.73
RandomForest class_weight subsample	0.69	0.18	0.77	0.73
BalancedRandomForest	0.69	0.18	0.78	0.73
GaussianNB	0.87	0.29	0.34	0.63
MLPClassifier	0.91	0.54	0.03	0.51

Table 5: Quantitative Dataset K 12.

quantitative_K_12	Acc	Precision	Recall	AUC
LogisticRegression	0.91	0.53	0.09	0.54
LogisticRegression class_weight	0.74	0.21	0.76	0.75
RandomForest	0.89	0.27	0.14	0.55
RandomForest class_weight	0.86	0.20	0.18	0.55
RandomForest class_weight subsample	0.86	0.20	0.18	0.55
BalancedRandomForest	0.70	0.18	0.71	0.70
GaussianNB	0.85	0.27	0.44	0.66
MLPClassifier	0.91	0.57	0.07	0.53

Table 6: Quantitative Dataset K 3.

quantitative_K_3	Acc	Precision	Recall	AUC
LogisticRegression	0.91	0.47	0.04	0.52
LogisticRegression class_weight	0.72	0.19	0.70	0.71
RandomForest	0.91	0.0	0.0	0.5
RandomForest class_weight	0.64	0.16	0.80	0.71
RandomForest class_weight subsample	0.64	0.16	0.80	0.71
BalancedRandomForest	0.64	0.16	0.80	0.71
GaussianNB	0.87	0.30	0.31	0.62
MLPClassifier	0.91	0.0	0.0	0.5

Table 7: Quantitative Dataset K 5.

quantitative_K_5	Acc	Precision	Recall	AUC
LogisticRegression	0.91	0.48	0.06	0.52
LogisticRegression class_weight	0.70	0.19	0.76	0.72
RandomForest	0.91	0.46	0.04	0.52
RandomForest class_weight	0.68	0.18	0.76	0.72
RandomForest class_weight subsample	0.69	0.18	0.76	0.72
BalancedRandomForest	0.68	0.18	0.78	0.72
GaussianNB	0.86	0.28	0.38	0.64
MLPClassifier	0.91	0.54	0.03	0.51

Table 8: Quantitative Dataset K 8.

quantitative_K_8	Acc	Precision	Recall	AUC
LogisticRegression	0.91	0.51	0.07	0.53
LogisticRegression class_weight	0.73	0.20	0.74	0.73
RandomForest	0.91	0.39	0.07	0.53
RandomForest class_weight	0.73	0.19	0.67	0.70
RandomForest class_weight subsample	0.72	0.19	0.67	0.70
BalancedRandomForest	0.70	0.19	0.77	0.73
GaussianNB	0.85	0.27	0.41	0.65
MLPClassifier	0.91	0.55	0.06	0.52

With the analysis, it has noted that models that deal with class imbalance showed a much more consistent result than those that don't, the Logistic Regressor and Random Forest

with class weight seem to have the better results among all of the models, so they will be chosen for next steps.

The Qualitative Dataset 2 showed the best results for the Random Forest models, therefore was chosen for the next experiments.

4.2. Confusion Matrix

It is presented in Table 9, the confusion matrix for all models performed. Its possible to see that the best model for true positive was random forest. For true negative was the baseline as expected since the naive classifier 1 (predicts zero for all) and 2 (predicts negative for most samples), however they do not that in account the class imbalance. The following best for true negative was random forest, which also has the few false positive. The neural network had the fewest false negative (worst case of error), however the difference from random forest is very little.

Based on the Confusion Matrix metric, it is possible to conclude that the random forest model presented the best result.

Table 9: Confusion matrix of all models.

Confusion Matrix	TP	TN	FP	FN
Baseline 1	0	58537	0	5422
Baseline 2	446	53482	5055	4976
RandomForest	4428	39835	18686	1010
Logistic Regression	4101	39463	19058	1337
Neural Network	4379	38354	20167	1059

4.3. Classification Report

It is presented in Table 10, the Classification Report for all models performed, considering the class 1, referent to the label "Yes" . For Accuracy metric, baseline 1 was better, however this is a misleading result, since it the models did not performed well for true labels, but mainly predicted right the false cases. From the other models, the neural network presented the highest value of 87%.

Precision is a good measure to determine when the costs of False Positive is high. The best case is Random Forest and secondly, we have the Logistic Regressions.

Recall calculates how many of the Actual Positives our model capture through labeling it as Positive. All values are low, and the best presented was from the Neural Network, with 27%.

F1 Score is a function of Precision and Recall, better metric to use if seeking a balance between Precision and Recall and there is an uneven class distribution with a large number of Actual Negatives, which is our case. The best case was Random Forest, with 31%, follow by Logistic Regression and Neural Network, both with 29%.

Based on the Classification Report metric, it is possible to conclude that Random Forest model presented the best result.

Table 10: Classification report of all models - Class 1 (Yes).

Classification Report	Accuracy	Precision	Recall	F1-score
Baseline 1	0.92	0.00	0.00	0.00
Baseline 2	0.84	0.08	0.08	0.08
RandomForest	0.69	0.81	0.19	0.31
Logistic Regression	0.68	0.75	0.18	0.29
Neural Network	0.87	0.33	0.27	0.29

4.4. ROC and AUC

It is presented in Table 11, the ROC-AUC for all models performed. For ROC-AUC the closest to 1, better is the model, so for this analysis the best result was random forest model, followed by neural network and logistic regression. This result can also be observed in Figure 7, which shows that the random forest model was the best model for this task.

The result for baseline 1 and 2, proves the misleading presented in previous metrics, since a no-skill classifier should give a ROC_AUC of 0.5 and 0.498, respectively, and our baseline models presented values closed to this.

Table 11: ROC-AUC of all models.

Models	ROC AUC
Baseline 1	0.500
Baseline 2	0.498
RandomForest	0.818
Logistic Regression	0.779
Neural Network	0.797

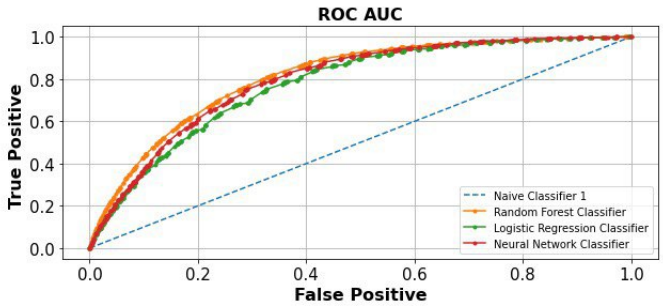


Figure 7: ROC Curve of all models.

5. Conclusion

It was presented the task of prediction of a heart disease in a person, based on key factor of her/his way of living, using the dataset *Key Indicators of Heart Disease*. First the dataset was explored and a feature selection was performed in a search for the best set of feature to better solve this task. The best dataset was chosen and model by two main machine learning approaches, traditional machine Leaning methods (Naive Classifier, Naive Bayes, Logistic Regression, Random Forest) and deep learning method (Fully Connected Neural Network). The models were trained and analysed in a set of metrics. As the dataset is imbalanced,

the most reliable metric is ROC-AUC. Considering this, it is concluded that the best model for this task and this dataset is random forest with a ROC-AUC of 0.818.

5.1. Next Step

A possible recommendation as next step, is to investigate unsupervided models as presented in the state-of-art (DEEPIKA; BALAJI, 2022).

Acknowledgments

The authors would like to thank the Institute of Computing, the State University of Campinas, CAPES and Griaule Company for opportunity and finance support.

Contributions

Authors individuals contributions:

- **Andre** Dataset analysis, data preparation, classical models analysis (Logistic Regression and Random Forests), jupyter notebook;
- **Douglas** Video, previous work with the dataset;
- **Matheus** Video, exploration of neural network architectures, trying to use downsampling and oversampling and modifying some parameters;
- **Renan** Neural Network analysis, train and evaluation, jupyter notebook;
- **Taylla** Dataset imbalance analysis, writing the Article and results discussion.
- All decided the dataset and the development of the paper.

Repository

All code used for this analysis of methods of machine learning and the *Key Indicators of Heart Disease* dataset can be found in the GitHub repository (⟨https://github.com/andreigor/MO444_HeartDiseasePrediction⟩).

References

DEEPIKA, D.; BALAJI, N. Effective heart disease prediction using novel mlp-ebmda approach. *Biomedical Signal Processing and Control*, v. 72, p. 103318, 2022. ISSN 1746-8094. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S1746809421009150⟩. page.11, page.88

OLIVEIRA, G. M. M. d. et al. Estatística cardiovascular – brasil 2021. *Arquivos Brasileiros de Cardiologia*, v. 118, n. 1, p. 115–373, jan. 2022. page.11

PYTLAK, K. *Personal Key Indicators of Heart Disease*. 2022. Disponível em: ⟨https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease⟩. page.22

STEVENS, B. et al. Os custos das doenças cardíacas no brasil. *Arquivos Brasileiros de Cardiologia*, v. 111, n. 1, p. 29–36, jul. 2018. page.11