Application of Supervised Machine Learning in Detection of Cardiovascular Diseases in Patients

Christina Morgenstern
christinamorgenstern@lewisu.edu
DATA-51000-002, 20
Data Mining and Analytics
Lewis University

I. Introduction

Cardiovascular diseases (CVDs) are a group of human diseases that affect the heart and vascular system such as coronary heart disease with heart attacks and strokes being the consequences. According to the World Health Organization (WHO), CVDs account for the number 1 cause of death globally with approx. 18 million deaths a year [1]. Risk factors contributing to the build-up of CVD are manifold and include increased levels of blood sugar (i.e. glucose) and blood lipids, raised blood pressure as well as overweight and obesity [1]. In order to prevent premature deaths, individuals at heightened risk must be identified for preventive measures. Supervised Machine Learning (ML) models can be used to classify patients into low and high-risk categories of suffering from CVD. Thus, the burden on patients as well as the medical system could be decreased because of early diagnosis and tailored treatment for those identified at risk.

The aim of this work is to investigate an open CVD dataset for its suitability to develop and train a ML model to be used by clinicians for future predictions on the presence or absence of CVD in patients.

This report describes the dataset, the methodology, the results along with a discussion and ends with a conclusion. In section II a description of the dataset is given with a discussion on the important features and their impact as well as selected visualizations. Subsequent sections discuss the methodology used to approach the problem with the algorithms discussed and the measures to validate stated (section III). The results from the individual ML models are presented and discussed in section IV. The last section, section V, summarizes the key points of this paper and provides possible future directions.

II. Data Description

The dataset used in this work was obtained from Kaggle [2] and contains three different types of input features related to the outcome of CVD. Objective information such as age, height, weight and gender describing the physical appearance of a patient are given. Furthermore, the patients have been asked to provide information on habits such as smoking, alcohol intake and physical activity. These features were of binary data type, denoted by 0 and 1 for the absence and presence of that habit, respectively. The dataset is complemented with clinical data that resulted from medical examination such as measuring blood pressure and blood testing. Table 1 lists the attributes present in the dataset along with the data type, an example value and a description.

In total, there were 70,000 samples and 13 feature columns present in the dataset. For the exploratory data analysis the data were loaded into the Jupyter notebook [3] environment and analyzed with the help of Python 3 programming software [4] using NumPy [5], pandas [6], Matplotlib [7] and seaborn [8] libraries. The visual programming software Orange [9] was used for creating appealing visualizations. The dataset did not contain any missing values. Nor were there duplicate patient entries as assessed by the uniqueness of the patient identification number (ID).

The distribution of the target label “cardio” is approx. balanced, with 35,021 samples categorized as 0, i.e. without a CVD, and 34,979 diagnosed with a CVD and hence labeled with 1. Thus, the dataset is balanced in terms of the two label classes which is beneficial for the performance of any ML algorithm. Interestingly, there are almost twice as many samples from women as compared to men (45,530 female and 24,470 male observations). However, the two CVD classes comprise of approximately equal numbers of female and male samples each (see Fig.1). Since CVD accounts for the leading cause of death in women it becomes increasingly important to account for these sex-differences in prevention, diagnosis and treatment of CVD [10].

TABLE 1. DESCRIPTION OF CVD DATASET

Attribute	Type	Example Value	Description
AGE	Numeric (integer)	18393	Patient´s age given in days
GENDER	Numeric (integer)	2	Patient´s gender (1: women, 2: men)
HEIGHT	Numeric (integer)	168	Patient´s height measured in cm
WEIGHT	Numeric (float)	62.0	Patient´s weight measured in kg
SYSTOLIC BLOOD PRESSURE (AP_HI)	Numeric (integer)	110	Patient´s systolic blood pressure in mmHg
DIASTOLIC BLOOD PRESSURE (AP_LO)	Numeric (integer)	80	Patient´s diastolic blood pressure in mmHg
CHOLESTEROL	Categorical (integer)	1	Patient's cholesterol level (1: normal, 2: above normal, 3: well above normal)
GLUCOSE	Categorical (integer)	1	Patient's blood glucose level (1: normal, 2: above normal, 3: well above normal)
SMOKING	Categorical (integer)	0	Whether patient smokes (0: no, 1: yes)
ALCOHOL INTAKE	Categorical (integer)	0	Whether patient drinks alcohol (0: no, 1: yes)
PHYSICAL ACTIVITY	Categorical (integer)	1	Whether patient exercises (0: no, 1: yes)
PRESENCE OR ABSENCE OF CARDIOVASCULAR DISEASE	Categorical (integer)	0	Target variable indicating healthy state (0) or cardiovascular disease (1)

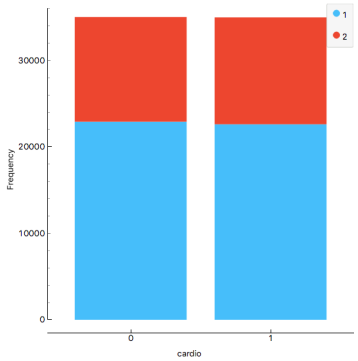

Fig. 1. Distribution of label types in CVD dataset demonstrating the absence (0) and presence (1) of CVD. The blue colour denotes female cases (1) while the red bars represent male samples (2).

Using descriptive statistics and Python´s pandas command `DataFrame.describe()`[11], the feature columns were explored for their distribution in mean, standard deviation, minimum and maximum values and different percentile measures (Table 2). This statistical analysis revealed outliers in the dataset (see Table 2 min and max values for AP_HI and AP_LO). It is not possible to have negative values for systolic and diastolic blood pressure. In this case it could be that the negative sign was introduced accidentally. The 150 mmHg to 70 mmHg could be interpreted as normal blood pressure values. On the other end of the scale, the maximum values of blood pressure are exceeding the normal range. 16020 and 11000 do not represent valid systolic and diastolic blood pressure values. The range for blood pressure is between 90 mmHg and 250 mmHg for the top (systolic) and 60 mmHg to 140 mmHg for the bottom (diastolic) values. In general, normal blood pressure ranges less than 120 mmHg for systolic and less than 80 mmHg for diastolic blood pressure [12]. Thus, hypertension is divided into three severity categories starting from elevated blood pressure levels of above 130 mmHg for systolic and higher than 80 mmHg for diastolic blood pressure [12]. The highest blood pressure measured in an individual thus far was 370/360 [13]. For that reason, systolic pressure values of higher than 300 mmHg and lower than 70 mmHg were removed from the dataset. Some of the 40 high values recorded were ridiculously high with several thousand mmHg. 189 samples were removed from the lower spectrum of the systolic pressure values. From the diastolic pressure entries, 953 values exceeding 200 mmHg and 53 values with diastolic pressure values below 30 mmHg were deleted. That left the dataset with 68,728 entries for further analysis. The new distribution of class labels is 34,730 with the absence (label 0) and 33,998 with the presence (label 1) of a CVD.

TABLE 2. DESCRIPTIVE STATISTICS ON FEATURE COLUMNS OF CVD DATASET

	Age [days]	Height [cm]	Weight [kg]	AP_HI [mmHg]	AP_LO [mmHg]
count	70000	70000	70000	70000	70000
mean	19468.87	164.36	74.21	128.82	96.63
std	2467.25	8.21	14.39	154.01	188.47
min	10798	55	10	-150	-70
25%	17664	17664	65	120	80
50%	19703	165	72	120	80
75%	21327	170	82	140	90
max	23713	250	200	16020	11000

The cleaned data was exported as .csv file and loaded into Orange for visual exploration of the feature distribution. Fig. 2 shows the distribution of recorded blood pressure values for systolic (Fig. 2a) and diastolic (Fig. 2b) pressures. From these visualizations it is obvious, that high blood pressure values are prevalent in cases with CVD as jugged by the high number of red labels present in the upper ranges of both blood pressure values. High blood pressure or hypertension has been described as one of the major risk factors for CVD [14].

The factor age has also been implicated in contributing to CVD with increased age being a risk factor [15]. The distribution of patients according to their age confirms this aspect with more cases with CVD in the older age group (Fig. 3).

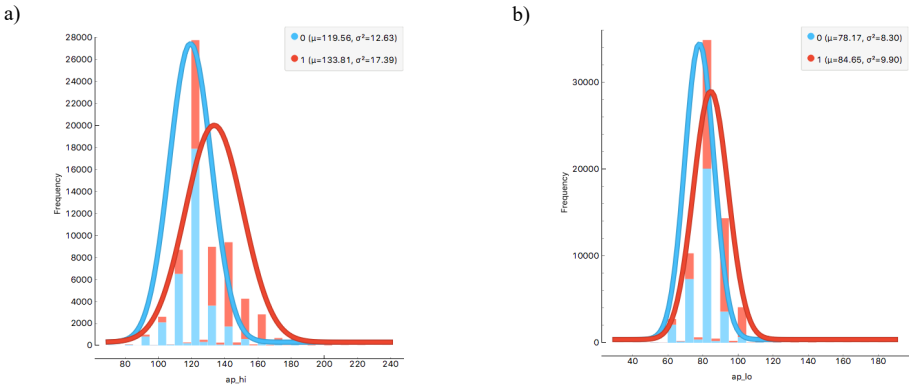

Fig. 2. Frequency distribution of systolic blood pressure (ap_hi) (a) and diastolic blood pressure (ap_lo) (b) in CVD dataset. Blue and red bars denote cases with and without CVD, respectively.

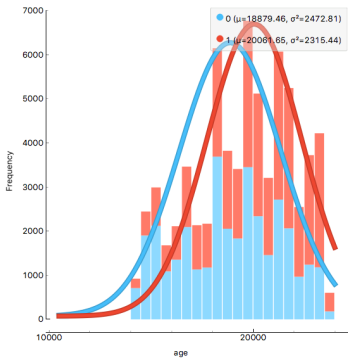

Fig. 3. Frequency distribution of the age attribute labeled with the absence (blue, 0) or presence (red, 1) of CVD.

In summary, the features present in the dataset are among the most relevant in assessing the future development of a CVD. In the next section the methodology of developing a prediction model based on these features is described.

III. METHODOLOGY

Classification is a supervised learning problem with the goal of predicting the categorical class label of unseen instances, based on previous observations [16]. The goal of this study is to learn a model from the labeled CVD dataset introduced in section II in order to make future predictions on patients whether they are susceptible to CVD or not. This section describes the process taken of developing an accurate model capable of predicting future patient's cardiovascular health. The individual steps taken in this approach are highlighted in the flowchart in Fig. 4. Upon retrieval of the dataset from Kaggle, the data was explored and cleaned using IPython and pandas as explained in section II. Subsequent feature engineering process led to the extraction of labels (CVD present or absent) and features. In order to bring the features onto the same scale, the numeric attributes (age, height, weight, api_lo and api_hi) were standardized at mean 0 and standard deviation 1. This makes the data comparable on the same scale and improves model performance.

The data were further normalized (i.e. rescaled to a range of [0,1]) and subsequently split into training and test set using an 80% to 20% split ratio. The training data were used for building different machine learning models, the test data were used to validate the model´s predictive performance. In order to generate the most accurate model, several algorithms were deployed: Logistic Regression [17], Support Vector Machine (SVM) [18], k-Nearest Neighbor (KNN) [19] and Random Forest [20]. These machine learning models are available via the Python scikit-learn [21] library and are adequate for the binary classification problem at hand.

The start made Logistic Regression, one of the most widely used algorithms for classification which predicts the probability that a specific sample is part of a particular class [17]. Support Vector Machine, a powerful classification algorithm, that tries to maximize the margin between the decision boundary and the training examples, was applied next [18]. The KNN algorithm finds k-samples in the training set that are most similar to the point to be classified and was also chosen as further supervised learning algorithm for this problem [19]. Lastly, a Random Forest model was applied to the data because this type of ensemble technique of decision trees are known for their good classification performance [20]. Initially all models were trained using the default parameters of the scikit-learn package. Parameter tuning and feature selection was performed in additional steps to increase the performance of the individual classifiers.

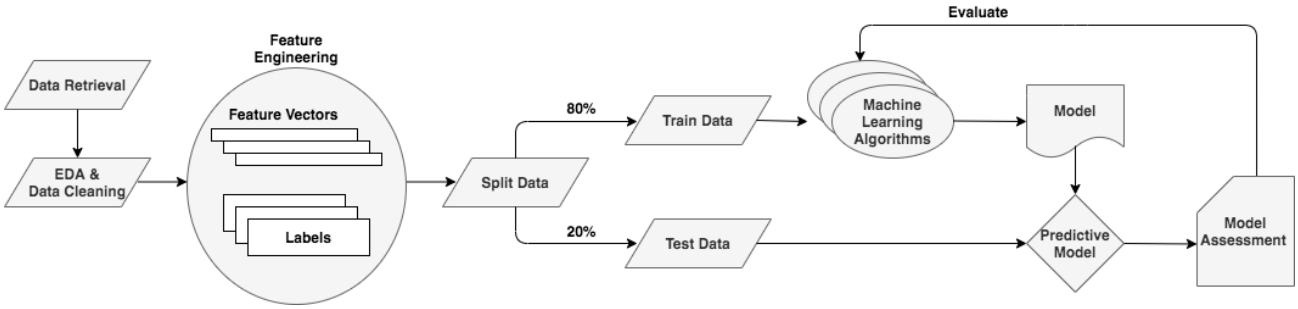

Fig. 4. Flowchart describing the process of applying supervised machine learning to cardiovascular dataset.

IV. RESULTS AND DISCUSSION

In this section, the results from the predictive assessment of the individual supervised machine learning algorithms are presented and compared based on their accuracy score (A) This metric compares the predicted labels with the actual labels and calculates the proportion that were correct. Section B and C discuss two approaches on feature selection. The results part ends with section D and a discussion on the best performing classifier and its suitability in clinical decision making.

A. Comparison of Supervised Machine Learning Algorithms

Table 3 gives an overview of the performance of the selected classifiers based on their training and test accuracy scores using the default parameters.

TABLE 3. PERFORMANCE COMPARISON OF SELECTED MACHINE LEARNING ALGORITHMS

Machine Learning Algorithm	Training Accuracy	Test Accuracy
Logistic Regression	0.73	0.73
Support Vector Machine	0.73	0.73
k-Nearest Neighbor (k=5)	0.87	0.69
Random Forest	0.99	0.71

The highest accuracy score that could be achieved with the initial settings was 73% for the Logistic Regression model and the SVM (Table 3).The KNN and the Random Forest model were overfitting i.e. the predictions on the training data were doing much better than on the test set. Since overlearning is not desirable, hyperparameter tuning was performed for the Random Forest Model. Setting the criterion to 'gini', n_estimators to 50, max_depth to 10, min_samples_leaf to 10 and max_features to 0.7 solved the issue of overfitting but didn´t increase the accuracy. The modified Random Forest algorithm had 76% accuracy on the train set and 73% accuracy on the test set.

In order to see whether feature selection could have an impact on the model accuracy, two different feature selection methods were applied. Section B describes the sequential feature selection algorithm with KNN and section C describes a Random Forest model for assessing the feature importance. Both algorithms were taken from Raschka and Mirjalili, 2017 [22].

B. Sequential Feature Selection and KNN

The aim of feature selection is to find the subset of features that best represents the whole dataset and improves the performance of the ML algorithm [23]. The Sequential Backward Selection (SBS) was used with the framework of the KNN algorithm to reduce the dimensionality of the dataset for better performance [22]. Fig. 5 shows the plot of the accuracy of the KNN algorithm versus an increased number of features. Form this analysis it doesn´t make any difference for the KNN algorithm whether only two or all features of the dataset are used. The performance of the KNN model stays just above 60%.

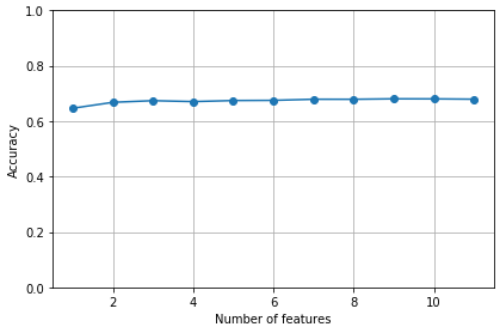

Fig. 5. Sequential feature selection and KNN performance.

C. Assessing of Feature Importance using Random Forests

An approach to select relevant features is to use a Random Forest which allows to measure the average impurity decrease from all decision trees in a forest using the `feature_importances_` attribute after fitting a Random Forest classifier [22]. Fig. 6 shows the features according to their importance. As assumed, the feature representing the systolic blood pressure value (ap_hi) has greatest importance, followed by the diastolic blood pressure (ap_lo), age and weight. From this analysis, alcohol and smoking don´t seem to be relevant for building the model. Subsequent models were built without taking into account the features alcohol use and smoking. Unfortunately, the accuracy of the Random Forest classifier and the other models could not be improved. Reducing the feature subset to ap_hi, ap_lo and age only, did not make any difference in accuracy either.

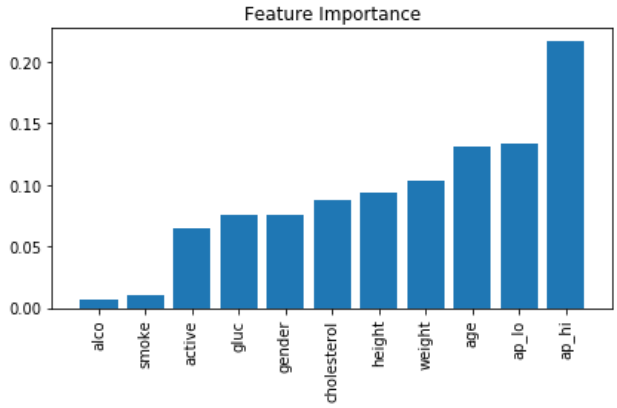

Fig. 6. Importance of individual features in CVD dataset as assed by using a random forest.

D. Application of Logistic Regression

Since Logistic Regression is one of the most used classification algorithms and the one model that performed among the best on that dataset, the results of this predictive modeling should be discussed in more detail in this section. It should be assessed whether this model has the potential of being used to classify patients as having a CVD or not. Table 4 shows the results from the classification presented as confusion matrix. The model classified 5,291 instances correctly with having the label "presence of CVD" i.e. as being true positives. Furthermore, 4,722 samples were correctly classified as true negatives with label "absence of CVD". 1,584 instances were wrongly classified as having no CVD although the actual label was 1 (presence of CVD). Further 2,159 samples were incorrectly classified as positives although the label actually was 0 (absence of CVD). Thus, this algorithm has a precision of 71% and a recall value of 77%. The F1-value can be derived from these two measures and was calculated to be 74%. Altogether these evaluation metrics are not promising for a classifier to be used to make clinical decisions. In almost a third of cases the algorithm would make a mistake and wrongly classify patients.

TABLE 4. CONFUSION MATRIX OF LOGISTIC REGRESSION MODEL

	1	**0**
1	5291	1584
0	2159	4722

V. CONCLUSIONS

Cardiovascular diseases are the global leading cause of death with almost 18 million people dying every year [1]. Predicting the presence or absence of CVD can help practitioners to inform decisions and impact patient´s health. Machine learning can be used in this decision process in classifying whether a patient is susceptible to CVD based on his/her physical appearance and clinical parameters. In this study a dataset with different types of patient attributes was used in order to build and evaluate a ML model capable of assessing the CVD status of a patient. In this effort, several supervised ML models have been used and their performances compared. However, none of them proved adequate to classify patients at this stage. The maximum accuracy of the tested models was 73% which is not suitable for a clinical decision system. While feature importance has been addressed in this study, no improvement could be made in the performance of the algorithm.

To improve this work, further possibilities in hyperparameter tuning could be assessed. Furthermore, other ML models could be deployed, such as neural networks. The quality of the dataset might also need to be investigated and probably more suitable features included.

REFERENCES

[1] "Cardiovascular diseases." https://www.who.int/westernpacific/health-topics/cardiovascular-diseases (accessed Jun. 18, 2020).

[2] "Cardiovascular Disease dataset." https://kaggle.com/sulianova/cardiovascular-disease-dataset (accessed Jun. 18, 2020).

[3] F. Perez and B. E. Granger, "IPython: A System for Interactive Scientific Computing," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 21–29, 2007, doi: 10.1109/MCSE.2007.53.

[4] Pilgrim, M., & Willison, S. (, "Dive Into Python 3," vol. Springer, .

[5] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy Array: A Structure for Efficient Numerical Computation," *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, Mar. 2011, doi: 10.1109/MCSE.2011.37.

[6] W. McKinney, "Data Structures for Statistical Computing in Python," Austin, Texas, 2010, pp. 56–61, doi: 10.25080/Majora-92bf1922-00a.

[7] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, May 2007, doi: 10.1109/MCSE.2007.55.

[8] Michael Waskom *et al.*, *mwaskom/seaborn: v0.8.1 (September 2017)*. Zenodo, 2017.

[9] J. Demšar and B. Zupan, "ORANGE: DATA MINING FRUITFUL AND FUN," p. 4.

[10] M. Garcia, S. L. Mulvagh, C. N. B. Merz, J. E. Buring, and J. E. Manson, "Cardiovascular Disease in Women: Clinical Perspectives," *Circ. Res.*, vol. 118, no. 8, pp. 1273–1293, Apr. 2016, doi: 10.1161/CIRCRESAHA.116.307547.

[11] "pandas.DataFrame.describe — pandas 1.0.5 documentation." https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html (accessed Jun. 21, 2020).

[12] H. H. Publishing, "Reading the new blood pressure guidelines," *Harvard Health*. https://www.health.harvard.edu/heart-health/reading-the-new-blood-pressure-guidelines (accessed Jun. 19, 2020).

[13] J. A. Narloch and M. E. Brandstater, "Influence of breathing technique on arterial blood pressure during heavy weight lifting," *Arch. Phys. Med. Rehabil.*, vol. 76, no. 5, pp. 457–462, May 1995, doi: 10.1016/s0003-9993(95)80578-8.

[14] P. K. Whelton, "Epidemiology of hypertension," *Lancet Lond. Engl.*, vol. 344, no. 8915, pp. 101–106, Jul. 1994, doi: 10.1016/s0140-6736(94)91285-8.

[15] S. Costantino, F. Paneni, and F. Cosentino, "Ageing, metabolism and cardiovascular disease," *J. Physiol.*, vol. 594, no. 8, pp. 2061–2073, Apr. 2016, doi: 10.1113/JP270538.

[16] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, Mar. 2016, pp. 1310–1315.

[17] S. Sperandei, "Understanding logistic regression analysis," *Biochem. Medica*, vol. 24, no. 1, pp. 12–18, Feb. 2014, doi: 10.11613/BM.2014.003.

[18] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
[19] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, “k-Nearest Neighbor Classification,” in *Data Mining in Agriculture*, A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, Eds. New York, NY: Springer, 2009, pp. 83–106.
[20] A. Liaw and M. C. Wiener, “Classification and Regression by randomForest,” *undefined*, 2007. /paper/Classification-and-Regression-by-randomForest-Liaw-Wiener/6e633b41d93051375ef9135102d54fa097dc8cf8 (accessed Jun. 20, 2020).
[21] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Mach. Learn. PYTHON*, p. 6.
[22] “Python Machine Learning - Third Edition.” https://www.packtpub.com/eu/data/python-machine-learning-third-edition (accessed Jun. 20, 2020).
[23] H. Liu, “Feature Selection,” in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 402–406.