Prediction of premature beats in the human hearts using machine learning techniques

Abu Emran Emon
Department of ECE
North South University
Dhaka, Bangladesh
abu.emon01@northsouth.edu

***Abstract*—The premature beat is a common cardiac disorder among humans. The arrhythmia can cause multiple dis-function in the human heart. Such as weakened heart muscle (cardiomyopathy), dangerous heart rhythms, Myocardial infarction, and possibly sudden cardiac death. This paper looks into a case of Premature beat in a care facility in Brazil. Machine learning techniques have been used to predict premature beats. Data has been prepossessed and various methods like Logistic Regression, decision tree classifier, SVM, KNN, naive Bayes and baseline classifiers have been used in the experiment. The paper reports that the model can predict with an accuracy of over 73.9%. furthermore, the data-set has 51 attributes regarding the health issues and the features can be useful to develop the prediction.**

***Index Terms*—premature beats, premature atrial and ventricular complexes, machine learning, classification**

I. Introduction

Premature beats or ventricular premature complexes [1] are abnormal heartbeats that originate in the ventricles or lower pumping chambers and disrupt the heart's normal rhythms [2]. The reason is, the stroke volume is low because left ventricular contraction occurs before the filling is completed [3]. Normally, older people are diagnosed with such problems. However, the disease can appeal to the hearts of normal healthy people or patients who are already diagnosed with heart problems. The premature beat is a symptom of a possible myocardial infarction. But the complication is, most patients with premature beats are asymptomatic. If premature beats in the hearts can be detected earlier, a huge chance of reducing the heart damage will arise. So, the objective of the paper is to generate a model by which one can detect whether premature beats are present in the heart.

In this paper, we are presenting predictive models which will be able to detect if a patient has a premature beat. We are using machine learning techniques to construct the model. The data-set [4] is available in Harvard Dataverse (https://dataverse.harvard.edu/) and was published on Aug 21, 2018, containing details about patients in a care facility in Brazil. The goal of the paper lies in the use of machine learning techniques to predict the presence or absence of premature beats. We tried to apply such a methodology that can predict premature beats in the human heart. So that, people can get the benefit and lead a healthy life.

The rest of the paper is organized as follows: In section 2, a critical literature review has been given. In section 3, different types of research methodologies that we have worked with, are presented. Experimental results are presented in section 4. Lastly, the conclusion of the paper with the highlights of the results obtained are in Section 5.

II. Background

The heart has four chambers. Two on the top and two on the bottom. Top chambers are called atria and bottom chambers are called ventricles. Blood enters into the heart through the right atrium and moves into the right ventricle. The left ventricle pumps them to the body through the left atrium [5]. When the left ventricle pumps the blood before the completion of filling, it is called premature beats. In other words, Premature beats are single ventricular beats that collapse before the next expected supra-ventricular beat, and usually those beats originate from different ventricular sites [6]. According to Harrison's Cardiovascular Medicine: premature beats or premature ventricular contractions are known as single ventricular beats which fall earlier than the next desired supra-ventricular beats [7]. The premature beat is a type of irregular rhythm that occurs earlier than anticipated, interrupting the natural heartbeat [8]. The beginning point of premature beats is those ventricles or atria. The disease is recognized as an abnormal heartbeat or irregular heart rhythm. The reason and exact causes of premature beats are not clear yet. Usually, the disease does not carry any rigid symptoms. That is one of the reasons, premature beats are hard to find out. But in some phases, the irregular heart rhythm can be strong enough to generate pain or discomfort in the chest and that can lead to a heart attack. Various ingredients can influence the process of premature beats. For instance: alcohol, caffeine, obesity, tobacco, high blood pressure, and anemia can influence the disease. Some symptoms like feeling dizziness, tiredness, anxiety, and weakness can be the reasons for irregular heart rhythm. According to the patients with premature beats: they feel like their hearts are skipping a beat but the case is fully opposite. Actually, in premature beats, the heart beats an extra beat [9]. Because of this, the muscles of the heart can be damaged and myocardial infarction and possibly sudden cardiac death may happen. The risk is

extensive. Various sources have described that because of premature beats or premature atrial and ventricular complexes, the irregular heart rhythm can be the cause of several heart problems and can damage the heart. A study found that, among the patients who were diagnosed with premature beats, 0.5% were under the age of 20 and 2.2 % were over 50 years old [2]. It is very essential to find out the arrhythmia earlier and get proper care to reduce the ratio of death causing heart diseases.

III. Research Work

The paper looks into the factor that can relate to premature beats in a care facility in Brazil. In the care facility, there was a record of patients and various test results related to their health and the presence or absence of premature beat. All the details are recorded in a data set. The data-set was first published by [4] in harvard dataverse. The source is open and available for all to use.

A. Dataset

The data-set has a total of 628 instances (after conversion to CSV file) (Atrial-270 & Ventricular- 358) and for each instance, there are 51 attributes present and it was published in 2018. Table I shows the details of each attribute that is recorded in the data set. One of the attributes represents the premature beat. The attribute is very crucial for our model to accomplish the expected prediction. The data of the data set is not large and the marge of the instances have been done. So, it is hard to get remarkable accuracy through using Machine learning techniques.

B. Research Methodology

Our objective is to develop a model that can predict premature beats in the human heart which can be helpful for both consultants and patients to get an idea about the situation of their heart at a very preliminary stage and take proper actions to prevent unwanted situations. As we noticed, various reasons and many factors are contributing to premature beats. Also, there is no one-stop solution for the declaration of premature beats. So, we use machine learning techniques to predict premature beats. scikit-learn [10], a python library for data science, has been used to get the data mining task. The research methodology has been given in fig1. The raw data-set (Premature atrial and ventricular complexes dataset) is prepossessed to a form that is compatible with using machine learning algorithms. The prepossessing phase includes data cleaning (removal of attributes, removal of missing instances, filling missing instances), data transformation, and data scaling. After prepossessing, we use machine learning algorithms to predict the performances.

C. Prepossessing of Attributes

Prepossessing is highly needed to transform the raw data into a form that can be suitable to apply machine learning algorithms. Our first approach towards prepossessing is to drop the duplicate data and eliminate the 'BNP' attribute because the missing values of the attribute were almost half of its total. Next, the values which were not available marked as 'NA' in the data-set are dropped. So, the instances are now free from null values. Afterward, we are going to work on our predicted attributes, so the data-set is being divided into 2 sections(predictor attributes and targeted attribute). The predictor attributes are named as 'X' and 'y' for the targeted attribute. Then, the variables of the predicted attributes ['sex', 'smoking', 'sedentary lifestyle', 'hypertension', 'diabetes', 'dyslipidemia', 'coronary disease', 'diuretic', 'ACEI / ARB', 'beta-blockers', 'calcium channel blockers', 'statin', 'palpitation', 'left atrium overload', 'left ventricular overload', 'intraventricular blocks', 'electrically inactive area', 'Chagas serology', 'TSH', 'magnesium', 'potassium', 'valvular disease', 'diastolic dysfunction'] are converted from categorical to numeric values. In this process, after applying the dummy variables, several new attributes are created and a total of 73 attributes are taking place in the newly formed dataset. After that, we split the predictor attributes into train set and test set where the train set is 85% and the test set is 15% of its total data and we use the normalization formula to scale the values as the values were not scaled. 25 features had to normalize and we used the normalization formula to handle the problem. Here, equation(1) shows the formula for normalization [11]. Here, the train set's max and min values are used for both the train and test set to normalize the values. After the scaling, all the values are formed between 0 to 1.

$$X'_{featurescaling} = \frac{X_i - Min}{Max - Min} \tag{1}$$

D. Feature Selection

The selection of relevant features is crucial in machine learning tasks. Irrelevant features behave as noise and influence decreasing the predictability. Hence, Removing the redundant features allow us to deal with fewer dimensions than before and it can improve the predictability power of the model. In this model, We use two techniques to select the features.

In the feature selection process, firstly, we implement the zero variance features selection technique or variance threshold method. In the variance threshold method, we investigate whether any feather has a variation of values and we fix the threshold up to 1. That means, if there is no variation or just one variation in the values of any attributes, the method immediately drops the feature as the feature is not providing any effect on the ongoing process. The feature selection method is only used on the train set. However, we execute the feature dropping process for both the train and test set since both sets have the same features. After implementing the method, the feature called 'magnesium_N' was dropped as it was a zero variance feature.

Afterward, the Pearson correlation coefficient is used for selecting features. The method is mechanized to select the features that are highly correlated. After selecting, it removes the first feature which is strongly correlated to another feature. The threshold applied to the method is 80%. That means, if any feature is 80% or more correlated to any other feature,

TABLE I
ATTRIBUTES AND DESCRIPTION

Attributes	***Data Type***	***Description***	***Min***	***Max***	***Mean***
Atrial	Boolean	0 For atrial off and 1 for open	0	1	x
sex	Object	Gender of the patients	x	x	x
Age	Integer	Age of the Patients	18	87	55.56
BMI	Float	Body Mass Index	14.25	53.15	27.35
Smoking	Object	'N' for no smoker otherwise 'S'	x	x	x
Alcohol	Float	The amount of Alcohol in Blood	0	52	1.25
Caffeine	Float	The amount of caffeine in Blood	0	2000	267.619
sedentary life style	Object	Mostly Leading a inactive life	x	x	x
Hypertension	object	the pressure in blood vessels is too high	x	x	x
Diabetes	Object	'S' for having diabetes otherwise 'N'	x	x	x
Dyslipidemia	Object	Abnormal level of cholesterol and other lipids	x	x	x
coronary disease	Object	Impedance or blockage of one or more arteries	x	x	x
Diuretic	Object	Medications designed to increase the amount of water('N' for no otherwise 'S')	x	x	x
ACEI / ARB	Object	In a medication for hing blood pressure('N' for no otherwise 'S')	X	x	x
beta-blockers	Object	drugs that can lower stress on the heart and blood vessels('S' for yes otherwise 'N')	X	x	x
Calcium channel blockers	Object	medications that relax blood vessels and increase the supply of blood and oxygen to the heart('N' for no otherwise 'S')	X	x	x
Statin	Object	Drugs that can lower cholesterol('N' for no othwewise 'S')	x	x	x
palpitation	Object	abnormally rapid or irregular beating of the heart (0 for no othwewise 1)	X	x	x
Systolic blood pressure	Float	the pressure in arteries when heart beats	90	220	129.62
Diastolic blood pressure	Float	The pressure in the arteries when the heart rests between beats	60	140	82.98
Heart Rate	Integer	Number of times the heart beats per minute	42	150	73.519
left atrium overload	Object	Left atrial enlargement due to pressure	x	x	x
left ventricular overload	Object	Heart's left pumping chamber's pressure and volume overload	x	x	x
intraventricular blocks	Object	Heart block of the ventricles of the heart	x	x	x
electrically inactive area	Object	The ventricular activation does not occur as expected	x	x	x
Chagas serology	Object	Diagnosis of chronic Chagas disease is made by serologic tests for antibody	x	x	x
TSH	Object	Thyroid stimulating hormone level OK or not	x	x	x
glucose	Float	Level of blood sugar	51	433	111.006
glycosylated hemoglobin	Float	Average level of blood sugar over the past few months	4.7	12.6	6.007
total cholesterol	Float	The total amount of cholesterol in blood	101	341	199.136
HDL	Float	High-density lipoprotein(good cholesterol) in blood	11	131	47.368
LDL	Float	Low-density lipoprotein (bad cholesterol)in blood	21	210	122.216
triglycerides	Float	The amount of triglycerides(fat or lipid) in blood	32	641	146.148
magnesium	Object	The level of magnesium in the blood	x	x	x
potassium	Object	Detect abnormal potassium levels, including high potassium (HYPER) and low potassium (HIPO)in blood	x	x	x
creatinine	Float	Blood test is used to assess kidney function	0.50	2.50	0.929
c_creat	Float	test report of blood	25.82	193.99	86.282
uric acid	Float	A waste product found in blood	2	10.8	5.371
BNP	Float	Measure the levels of a protein in blood	2	424.7	23.958
diastolic diameter	Float	Heart's Ventricle's relaxation contraction measurement	38	74	49.643
systolic diameter	Float	Heart's Ventricle's relaxation measurement	20	62	30.565
ventricular mass index	Float	A parameter used in echocardiography and cardiac MRI	54.24	256.40	102.433
ventricular ejection fraction	Float	The amount left ventricle (or right ventricle) pumps blood with each heart beat	33	82	68.159
left atrium size	Float	The size of left atrium	22	55	35.587
posterior wall thickness	Float	The thickness of heart's posterior wall	6	13	9.566
septal thickness	Float	Thickness of heart's septal wall	7	14	9.732
valvular disease	Object	Cardiovascular disease	x	x	x
diastolic dysfunction	Object	A cardiac condition caused by the heart's ventricles	x	x	x
mean heart rate	Integer	The number of times the heart beats per minute	41	123	78.646
QT interval	Float	The measurement of electrocardiogram used to assess some of the electrical properties of the heart	123	652	434.747
record of symptoms	Object	Premature beats present/absent record	x	x	x

the first one will be redundant by the method. We use the selecting method only for the train set and execute the dropping algorithm for both the train and test set as both sets have the same features. In the process, the method removes 26 features as they are highly correlated to another feature.

After executing both methods, 46 features are remaining from 73 and 27 features are redundant by the methods. In table II the name of the methods and the number of redundant features are given.

TABLE II
FEATURE SELECTION METHOD

Method name	***Number of redundant features***
Zero Variance	1
Pearson Correlation	26

IV. EXPERIMENTS AND RESULTS

Finding other works related to premature beats using machine learning or data mining algorithms is difficult. So, getting information from previous works is inconvenient.

In this part, the experimental methodologies and their related results are going to be discussed. Before conducting the experiments, we prepossessed the entire data-set and used the Zero variance and Pearson correlation methods to the select features. After conducting the processes, We have proceeded to the experiments.

The experiments are conducted in train test split strategy and several classifiers such as Baseline Classifier, tree-based classifiers(Logistic regression, Decision Tree classifier), k-nearest neighbors (kNN), support vector machine(SVM), and Naive Bayes classifiers are applied. The methodology has been outlined in figure 1. And table III denotes the performance results of each classifier.

Our first algorithm is the baseline classifier. The classifier is a basic model to generate results and is used to create a baseline to compare with complex solutions in order to get better performance. And as expected, the model performs average in the experiment.

The decision tree is a classifier where the data is splitting continuously according to certain parameters. The tree can be explained in two parts: nodes and leaves. The data is splitting in nodes and the leaves are the final results [12]. In our experiment, we execute the method by using several depths from 2 to 50 and observe that the tree produces better results when the depth is 2 and the test accuracy for the depth is

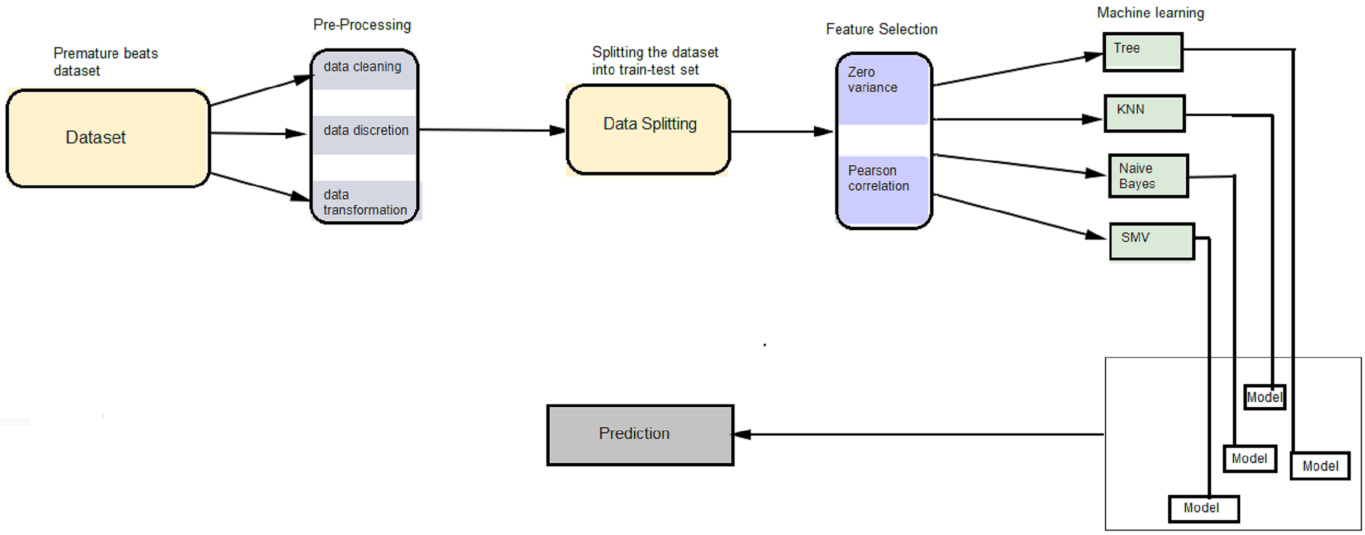

Fig. 1. Research Methodology

69.6%.

k-nearest neighbors algorithm (kNN) is an algorithm where it tries to find the possible similar data, if not the same, in a common category. The algorithm is also known as a lazy and non-parametric algorithm as it does not establish assumptions on underlying data. The algorithm is used in the experiment with different metrics, including Euclidean and Manhattan to measure the distances. The measure of true straight line distance between two points in euclidean space is known as Euclidean distance. And the distance between two points is the sum of the differences of their Cartesian coordinates is known as Manhattan distance [13]. The distance is also known as city block distance. Equations (2) and (3) provide the formula of euclidean distance and Manhattan distance accordingly. In the experiment, we conduct several tests using different values for the nearest neighbor for both Euclidean and Manhattan distances. And observe, when 25 is the value of K, the model provides higher test accuracy for euclidean with 71.7% and Manhattan distance provides 73.9% accuracy when K is 20.

Our next algorithm is the Logistic regression classifier. It works in an environment where the target value belongs to one class or another and the algorithm behaves in the situation of a categorical environment. We use logistic regression without regularization and with L2 regularization in our experiment. L2 regularization means, it will not construct sparse models and all coefficients are reduced by the same factor [14]. The test accuracy of the model with no regularization is 73.9% and with L2 regularization 71.7%. Both the regularization are providing a tiny difference in their performance.

Naive Bayes classifier is a probabilistic classifier, which means it predicts on the basis of the probability of an object [15]. As we implement the algorithm in our model, the test accuracy of the method is 71.7%.

The final algorithm used in the experiment is the Support Vector Machine(SVM). The algorithm's goal in the model is to create an environment where it can break down the decision boundary into possible dimensional spaces into classes so that it can predict possible results. The algorithm provides an accuracy of 71.7% in this experiment.

In figure 2, shows the accuracy comparison of the models.

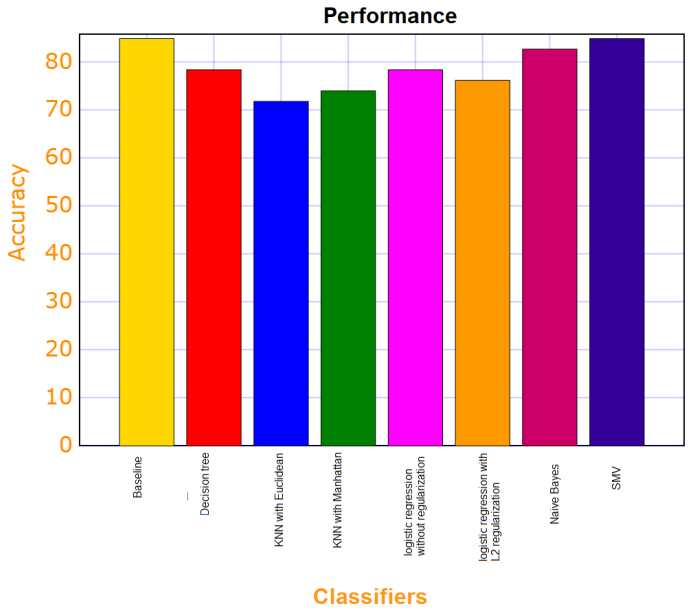

Fig. 2. Performance of the models

$$D_{Euclidean}(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{2}$$

$$D_{Manhattan}(p,q) = \sum_{i=1}^{n}(|p_i - q_i|) \tag{3}$$

TABLE III
AVERAGE OUTPUTS OF EXPERIMENTS

Experiment name	***Precision(no/yes)***	***Recall(no/yes)***	***F-measure(no/yes)***	***ROC area***	***Accuracy***
Baseline Classifier	71.7 / 0.0	100.0 / 0.0	83.5 / 0.0	57.0	71.7 %
Decision Tree Classifier	72.1 / 33.3	93.9 / 7.7	81.6 / 12.5	57.0	69.6 %
KNN with Euclidean Distance	71.7 / 0.0	100.0 / 0.0	83.5 / 0.0	65.0	71.7 %
KNN with Manhattan Distance	73.3 / 100.0	100.0 / 7.7	84.6 / 14.3	65.0	**73.9** %
Logistic Regression with no regularization	76.9 / 57.1	90.9 / 30.8	83.3 / 40.0	60.0	**73.9** %
Logistic Regression with L2 regularization	73.8 / 50.0	93.9 / 15.4	82.7 / 23.5	60.0	71.7 %
Multinomial Naive Bayes classifier	71.7 / 0.0	100.0 / 0.0	83.5 / 0.0	52.0	71.7 %
SMV	71.7 / 0.0	100 / 0.0	83.5 / 0.0	62.0	71.1 %

V. Discussion

The data-set has gone through different algorithms and the experiments provide variation in their performances. As we observe in table III Experimental results are fluctuating from one method to another. In addition, the scores of precision, recall, and f-measure are provided with the percentages of 'no' and 'yes' individually and the accuracy row indicates the test accuracy of the models. The table also indicates an accuracy of 71.7% obtained using SMV. Surprisingly baseline classifiers also acquire the same accuracy which is 71.7% and the Decision tree classifier provides the best performance when the depth is 2. For KNN algorithms, different metrics are being used to determine the distances. KNN with euclidean while K=25, provides 71.7% accuracy, and KNN with Manhattan marks an accuracy of 73.9% while K=20 and it is the highest performance gained among these algorithms. On the other hand, logistic regression without regularization and with L2 regularization both achieves 71.7% accuracy. Naive Bayes classifier also obtained 71.7% accuracy. After observing all the results, we can say, the experiments yield a model that can predict the premature beat.

VI. Conclusion

The paper investigates the premature beat's data-set where the detailed records of health conditions of patients from a primary care facility which is situated in Brazil, were stored. Initially, 628 instances and 51 attributes were recorded. The data-set had to go through multiple prepossesses and machine learning algorithms were applied after the completion of the prepossessing and feature selection method. Different algorithms were applied to the data-set. Each algorithm provides variation in their performances. After obtaining all the results, finally, the model can predict premature beats in the human heart with over 73.9% accuracy.

References

[1] C. Koester, A. M. Ibrahim, M. Cancel, and M. R. Labedi, "The ubiquitous premature ventricular complex," *Cureus*, vol. 12, no. 1, 2020.

[2] "Premature ventricular contractions," https://my.clevelandclinic.org/health/diseases/17381-premature-ventricular-contractions, online; accessed 8 August 2021.

[3] S. H. Ralston, I. D. Penman, M. W. Strachan, and R. Hobson, *Davidson's Principles and Practice of Medicine E-Book*. Elsevier Health Sciences, 2018.

[4] W. N. Ribeiro, A. T. Yamada, C. J. Grupi, G. T. d. Silva, and A. J. Mansur, "Premature atrial and ventricular complexes in outpatients referred from a primary care facility," *Plos one*, vol. 13, no. 9, p. e0204246, 2018.

[5] A. K. Datta, *Essentials of human anatomy: Superior and Inferior Extremities*. Current books international, 2009.

[6] J. L. Jameson, *Harrison's principles of internal medicine.* McGraw-Hill Education,, 2018.

[7] J. Loscalzo, *Harrison's cardiovascular medicine 2/E*. McGraw-Hill Education, 2013.

[8] Wexner Medical Center, The Ohio State University, "Premature Heartbeats," https://wexnermedical.osu.edu/heart-vascular/heart-rhythm/premature-heart-beats, online; accessed 3 August 2021.

[9] Hope Cristol, "What Are Premature Ventricular Contractions?" https://www.webmd.com/heart-disease/premature-ventricular-contractions-facts, 2020, online; accessed 3 August 2021.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[11] Aniruddha Bhandari, "Feature scaling for machine learning: Understanding the difference between normalization vs. standardization," https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/, 2020, online; accessed 29 August 2021.

[12] "Decision trees for classification: A machine learning algorithm," https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html, 2017, online; accessed 29 August 2021.

[13] Sarang Anil Gokte, "Most popular distance metrics used in knn and when to use them," https://www.kdnuggets.com/2020/11/most-popular-distance-metrics-knn.html, 2020, online; accessed 29 August 2021.

[14] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[15] "Naïve bayes classifier algorithm," https://www.javatpoint.com/machine-learning-naive-bayes-classifier, online; accessed 29 August 2021.