DOI 10.29042/2019-5009-5023

Helix Vol. 9 (3): 5009- 5023

Evaluation of SVM Kernels for Health Risks Assessment

[1]Amrik Singh, [2]K. R. Ramkumar

[1]Department of Computer Engineering, M.B.S. College of Engineering & Technology, Jammu, India

[2]Chitkara University Institute of Engineering and Technology, Chitkara University Punjab, India

Email: amriksingh07@gmail.com, k.ramkumar@chitkara.edu.in

Received: 27th December 2018, Accepted: 13th February 2019, Published: 30th June 2019

Abstract

As per the statistics of the National Family Health Survey – 4 (2015-16) only 28.7% of families in India have been covered under Health Insurance. There are many categories of the population that are not covered by the insurance companies as they are reluctant to insure them. The main reason is that they are not able to compute fitness level of the people who wants to get insured properly. In this paper, the preliminary analysis of data set is given equal importance as in constructing and fine-tuning machine learning model. The selection of features for machine learning models was done based on correlation as well as on the medical significance of the attributes. The features that are medically significant and has a minimum correlation among themselves were selected for constructing SVM kernel models. The selection of the most appropriate SVM Kernel was done by multiple evaluations of six SVM kernels. It was found that the Medium Radial kernel performs best in term of accuracy but linear kernel training time is least among all kernels. The values of C-statistics are consistent with the accuracy values in almost all cases and it shows that medium radial kernel is the best choice to automate the health risk assessment as it is faster as well as most accurate in prediction.

Keywords

Support Vector Machine, Kernels, Health Risk, Classification.

Introduction

Support Vector Machine [1] [2] [3] [4][5] (SVM) is an algorithm that distinguishes boundaries between the classes within a dataset [6]. The basis of the SVM is a regression, but it is more than just finding and separating borderlines between the classes [7]. The algorithm uses a 'kernel trick' [8] to transform the dataset into separate planes [9]. The word support vector represents the coordinates of each individual observations mapped on the vector space model [10]. For example, if (x_1,y_1) corresponds to Class A and (x_2,y_2) corresponds to the Class B, and if these data points are far apart, then they can be considered as 'support vectors' to differentiate between the Class A and Class B. In other words, higher the distance/difference, easier it is to identify particular class of 'support vector'. Hence by definition, the aim of the SVM is to identify the separating hyperplane (optimal) which maximizes the margin of the training dataset between respective classes [11].

Secondly, the logic of the SVM algorithm(s) is to maximize the distance between the two decision boundaries of the classes that it deals with. This means increasing the separation distance by choosing the best hyperplanes [12][13]. If the boundary is non-linear due to overlapping data points then the inner product method is used to transform the data into higher feature space. This is primarily called a kernelized machine learning models but Kernelization can only happen if the kernel function(s) satisfy the condition of symmetry (k(x,y) = k(y,x)) and positive semidefiniteness [14].

Checking symmetry is a straightforward method but checking semi-definiteness can be done by random simulation or experimentation with the feature data. Hence, selecting the right kernel function(s) is required and this can be done using an empirical approach and evaluations on the datasets. The current narrative in every field including insurance [15] [16] [17] [18] is to apply machine learning algorithms to automate the workflows [19] [20]. For, this there is always a need to evaluate the nature of the dataset as well.

Organization of the Paper

Section 3 gives a Review of the work done whereas Section 4 explains the Scope of work and objectives. In Section 5 implementation is done and in Section 6 data collection and dataset, characteristics are discussed. Section 7 shows performance evaluation and comparison and finally, Section 8 is a conclusion and Section 9 future scope.

Review

Many researchers have used support vector machine learning algorithms to check [16][15] [18] [17], insurance claims [21] [22] and classifying the nature subjects data for Health/disease risks [23] [24] [25] . Pieces of evidence can be found about the use of support vector machine algorithms to check the level of risks associated with the person. Study of all these research reveals that there is always the need to evaluate existing algorithms for new datasets before designing new algorithms or reinventing the wheel. There are many cases where fine-tuning of parameters [26] , appropriate selection of training data [27][28] [29] and optimization of the kernel functions [30] [31] [32] of support vector machines have yielded to highly accurate results.

 Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

Combinational approaches [33] [34] [35] have also been found to improve the accuracy of the Support vector machine algorithms .

Support Vector Machine learning models have been applied to a multitude of problems due to the effectiveness of Kernel functions. In the paper [36] the researchers have processed the structured as well as unstructured data patterns. The processing of both types of data is done to construct a machine learning model that can classify and predict cerebral infraction. The authors have evaluated multiple models (unit and multi models) to arrive at an optimal learning model that produces a high degree of accuracy. The algorithms employed methods that are variants of convolution models. Elaborate experimentation of text-based data and tabular form have been shown in the results section. The result claimed in the paper show that the approach in pre-processing and selection of the algorithm for the said data set works well to produce accuracy above 94%.

This research work [37] investigates the application of machine learning about predicting heart-related issues using clinical data, that shows machine learning models such as random forest, logistic regression, neural network and gradient boosting algorithm do help in improving the predictability of the data for classification of cardio-vascular issues. Out of all these models neural networks has the maximum ability to do so as it is most accurate in solving such problems. According to the authors, the result was validated using multiple performance parameters such as AUC.

In the works of [30], the Support Vector Machine (Correlation Kernel) has been revised to supports positive semi-definite values. The outcome of the modified classifier shows better performance achieved, as compared to the KNN and Decision Tree algorithm. These comparative conclusions were made on the basis of cancer data. The generic correlation kernel may produce negative values, as certain parameters may have negative casual relationships. In order to avoid the problem of non-positive values the kernel matrix, and diagonal matrix 'P' is transformed into P: where all values in the matrix are always positive in nature.

Wellness health insurance [38] [39] [40] is a method of rewarding and giving a bonus to the people, who maintain good health. It is a way by which the health insurance companies and improve their operating margins as a number of claims may go down. At the same time, people can maintain their fitness level to get rewarded. This idea has been applied to do health risk analysis [41] [25] [42] [43] applicable to insurance by many researchers. The adoption is a form of health insurance depends on the advances in medical sensor [41] technologies [44]. Many pieces of evidence can be found in the contemporary literature about the use of medical sensors , cloud and IoT technologies employed in the collection of biological data [44] [45] [38] [46] . Such advances in collection and monitoring [47] of biophysical data methods can help the health insurance industry immensely according to many researchers [48] [39] [40]. But, it also evident that use such technologies need careful evaluation/assessment of the algorithms, legal bindings, and platforms for it to be successful.

The Scope of Work and Objectives

It is abundantly clear that a preliminary analysis of the dataset can help in improving the classification tasks, especially when the datasets are not benchmarked. Computing fitness levels [49], health risks [50] and proneness to the disease are requisite in "wellness health insurance". Therefore, based on a systematic review on the problems and issues of optimization of the support vector machine [13] [32] based functions for new datasets the scope and objectives of the work can be formulated as follows:

i. Data Collection for Health Risk Assessment
ii. Analysis of health risk attributes values for gaining apriori knowledge for selection of kernels of SVMs.
iii. Application and evaluation of Kernels of SVM (six) for automating the health risk assessment.

Implementation

This section explains the implementation steps taken to achieve the aforementioned objectives. Care has been taken to follow the well-established procedures for evaluating the six kernels of the SVM for automating the Health Risk Assessment. For maintaining the quality of the dataset, all procedures related to missing values treatment, data cleaning, data grouping, and data validation have been followed during data collection and analysis. The next section explains the characteristics of Health Risk Data.

Data Collection and Dataset Characteristics

The Health Risk dataset as shown in Table 1 consists of sixteen attributes or the indicators of the health risk. The values were collected using Omron body composition machine [51]. Demographically the dataset consists of Asians located in North India, it is available at [52]. The numbers of instances are 500 and dataset does not contain any missing or any duplicate values [53]. The key objective of data collection is to leverage the use of machine learning algorithm for conducting an automated health risk assessment. The description of the dataset is as follows

Attributes	Description	Notation
Birth Age	Age of a person	BA
Height	The height of a person in centimeters	H
Gender	Male or Female	G
Weight	The weight of a person in Kg	WT
Body fat	It is body fat mass w.r.t total body weight.	BF
Visceral Fat	Fat around the internal organs such as the liver, pancreas, intestines etc.	VF

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

Skeleton Muscle	Muscles attached to organs and bones.	SM
Body Age	Biological age of a person	BDA
Resting Metabolism	It is calories needed to ingest to provide energy for body functioning.	RM
Body Mass Index	It is the ratio of weight and height of a person.	BMI
Blood Pressure Systolic	Maximum arterial pressure during contraction of the left ventricular contraction occurs is called systolic	BPsys
Blood Pressure Diastolic	Minimum arterial pressure during relaxation	BPdia
Pulse	Heartbeat Rate	P
Sugar F	Blood Sugar measured on fasting	SF
Sugar PP	Blood sugar measured After taking Meal	SPP
Waist	External Measure of the area below the ribs and above the hips in centimeters	W

Table 1: Health Risk Attributes and Notations

Attribute	**Range value(s)**		**Fitness Level class**
BMI	18.5 – 24.9 [54] > 25		FL1 FL2
BP	**Systolic**	**Diastolic** [55]	
	100 – 120 > 120	65 - 80 > 80	FL1 FL2
P	<60 - 100 [56] >100		FL1 FL2
BF (%age)	**Gender**	**Range**	
	Male	10.0 – 19.9 > 20.0	FL1 FL2
	Female	20.9 – 29.9 > 30.0	FL1 FL2
VF(%age)	1 – 9 % >10 %		FL1 FL2
SM(%age)	**Gender**	**Range**	
	Male	32.9 – 35.7 > 35.8	FL1 FL2
	Female	25.9 – 27.9 > 28	FL1 FL2
BDA(Yrs.)	± 3 > 3		FL1 FL2
Blood Sugar (mg/dL)	**Fasting**	**After Meal (PP)**	
	80 – 120 > 120	80 – 140 > 140	FL1 FL2
Waist (cm)	(Height (cm) / 2) + (0.03 * Height) (Height (cm) / 2) > (0.03 * Height)		FL1 FL2

Table 2: Medical Classification of Fitness Level of the Subject Under Observation

FL1 – Fitness Level 1 (Healthy), FL2 – Fitness Level 2 (Unhealthy)

Table 2 gives information on the medical conditions and the logic based on which the fitness data is grouped. Care has been taken to use universal medical standards, definitions, and ranges for each fitness factor recorded.

6.1 Preliminary Analysis of Dataset

The accuracy of the machine learning algorithm depends largely on the nature of the dataset [57]. If the dataset has large numbers of dimensions, it may lead to the "curse of dimensionality" [58] and may render the performance of the machine learning model in a reduced form. Secondly, some non-significant factors also play role in reducing the overall performance where missing values and correlation between the factors also contribute [57] . The selection of SVM Kernel also depends on the shape of the distribution and linearity between the various factors [58] [32] [59] [31]. Hence, this section 6.1 gives the initial statistical observations of the dataset that would help in constructing the machine learning model.

6.2 Non-linearity and Erratic value Analysis

It is statistically and visually clear from the graph Fig.1 (a) – Fig. (l) that the data is non–linear data. But, a minimal degree of linear trend can also be seen, if a trend line is created between the data points. The dataset shape shows that linear, as well as non-linear kernels, may work up to a certain level of accuracy. But, for a full evaluation, the speed at which the algorithm learns (training time), the empirical values of prediction speed also need to be evaluated for final selection. In general sense, most of the health parameters seem to be non-linear in nature. This fact can be validated by drawing mathematical relation equations [Table 3] from the dataset using curve fitting methods. It can be seen that most of the attributes have a polynomial curve (degree 3) equation in relation to the response variable 'y' ($y \in$ fitness Level 1 | Fitness Level 2). If the dataset is inappropriately labeled or has erratic values, the accuracy of the learning model will suffer. Hence, it appears that there is a need to choose or construct those kernels that can handle non-linear data at first glance.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

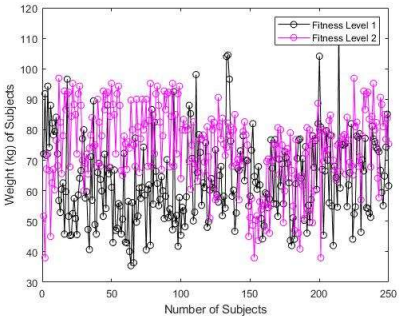

Fig. 1(a) Fitness level for Weight

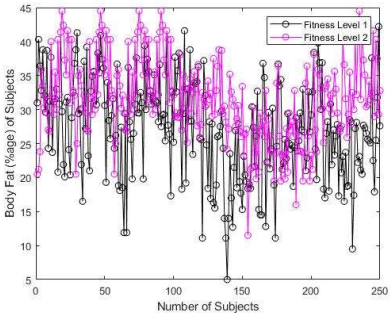

Fig. 1(b) Fitness Level for Body Fat

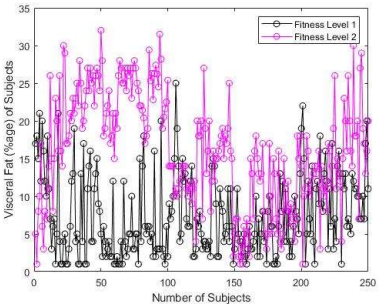

Fig. 1(c) Fitness Level for visceral Fat

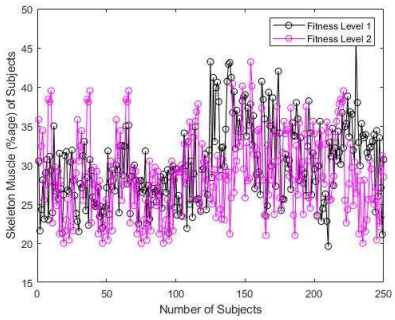

Fig. 1(d) Fitness Level for Skeleton Muscle

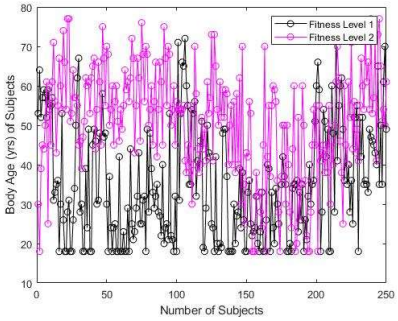

Fig. 1(e) Fitness Level for Body Age

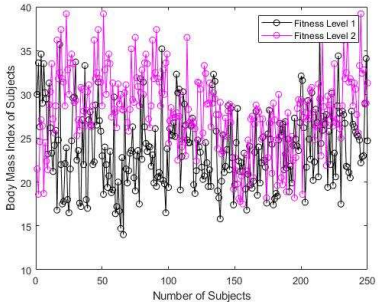

Fig. 1(f) Fitness Level for BMI

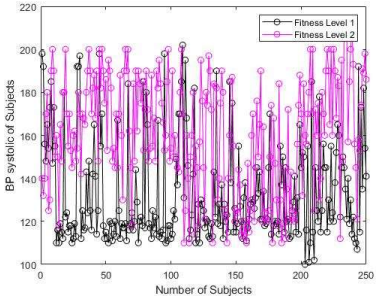

Fig. 1(g) Fitness Level for BP Systolic

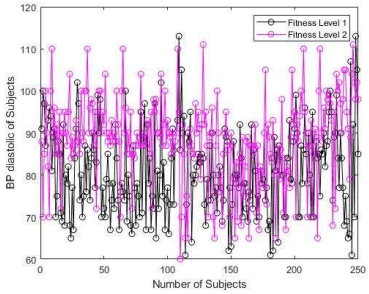

Fig. 1(h) Fitness Level for BP Diastolic

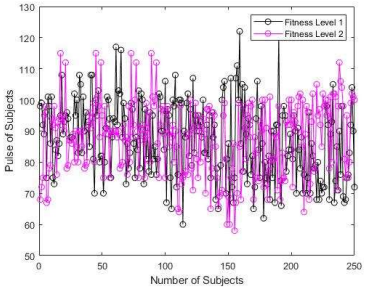

Fig. 1(I) Fitness Level for Pulse

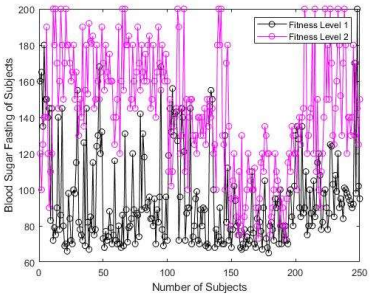

Fig. 1(j) Fitness Level for Blood Sugar Fasting

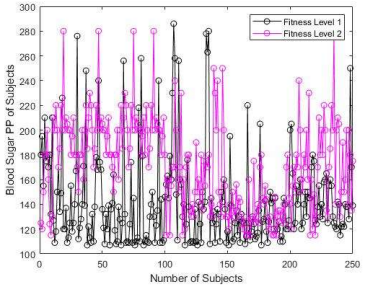

Fig. 1(k) Fitness Level for Blood Sugar PP

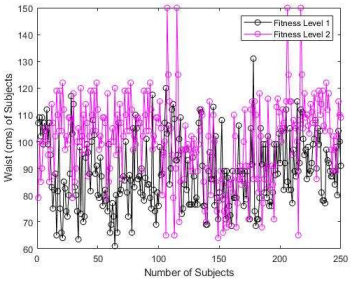

Fig. 1(l) Fitness Level for Waist

Fig. 1(a) – Fig. (l) shows the Line Graph for Different Fitness Attributes

Attributes	Fitness Level 1 Dataset		Fitness Level 2 Dataset	
	f(x) = p1*x^3 + p2*x^2 + p3*x + p4	Goodness of fit	f(x) = p1*x^3 + p2*x^2 + p3*x + p4	Goodness of fit
BA	p1 = 2.664 (0.775, 4.553) p2 = 6.76 (5.102, 8.417) p3 = -7.235 (-10.94, -3.528) p4 = 22.46 (20.24, 24.68)	SSE: 3.479e+04 R-square: 0.2521 Adjusted R-square: 0.24 RMSE: 11.89	p1 = 1.204 (-0.5497, 2.957) p2 = -0.02769 (-1.564, 1.509) p3 = -3.453 (-6.883, -0.02384) p4 = 38 (35.95, 40.06)	SSE: 2.971e+04 R-square: 0.02105 Adjusted R-square: 0.009114 RMSE: 10.99
H	p1 = -1.77 (-3.167, -0.3737) p2 = 0.1287 (-1.097, 1.354) p3 = 4.766 (2.026, 7.506) p4 = 162.3 (160.7, 164)	SSE: 1.901e+04 R-square: 0.0551 Adjusted R-square: 0.04357 RMSE: 8.792	p1 = 0.7202 (-0.4954, 1.936) p2 = -0.04148 (-1.107, 1.024) p3 = -1.001 (-3.378, 1.377) p4 = 162.6 (161.2, 164.1)	SSE: 1.428e+04 R-square: 0.006982 Adjusted R-square: -0.005128 RMSE: 7.619

WT	p1 = -1.339 (-2.999, 0.3203) p2 = 2.585 (1.129, 4.041) p3 = 2.946 (-0.3104, 6.203) p4 = 56.09 (54.14, 58.04)	SSE: 2.685e+04 R-square: 0.05894 Adjusted R-square: 0.04746 RMSE: 10.45	p1 = 0.2904 (-1.245, 1.826) p2 = 0.5168 (-0.8292, 1.863) p3 = -1.515 (-4.518, 1.489) p4 = 78.58 (76.78, 80.38)	SSE: 2.279e+04 R-square: 0.0135 Adjusted R-square: 0.001465 RMSE: 9.625
BF	p1 = 0.4425 (-0.5414, 1.426) p2 = 0.498 (-0.3653, 1.361) p3 = -2.731 (-4.662, -0.8001) p4 = 25.16 (24, 26.31)	SSE: 9439 R-square: 0.097 Adjusted R-square: 0.08599 RMSE: 6.194	p1 = -1.106 (-2.039, -0.1739) p2 = 0.121 (-0.6963, 0.9384) p3 = 1.012 (-0.8121, 2.836) p4 = 32.75 (31.66, 33.84)	SSE: 8404 R-square: 0.04809 Adjusted R-square: 0.03648 RMSE: 5.845
VF	p1 = -0.3588 (-0.9858, 0.2683) p2 = 2.269 (1.719, 2.819) p3 = -0.3886 (-1.619, 0.8419) p4 = 3.656 (2.92, 4.393)	SSE: 3833 R-square: 0.2561 Adjusted R-square: 0.247 RMSE: 3.947	p1 = 1.345 (0.5959, 2.095) p2 = 0.2437 (-0.413, 0.9005) p3 = -5.229 (-6.694, -3.763) p4 = 17.53 (16.65, 18.4)	SSE: 5426 R-square: 0.2943 Adjusted R-square: 0.2857 RMSE: 4.697
SM	p1 = -1.074 (-1.821, -0.3263) p2 = 0.5024 (-0.1533, 1.158) p3 = 2.966 (1.499, 4.433) p4 = 30.37 (29.49, 31.24)	SSE: 5446 R-square: 0.08316 Adjusted R-square: 0.07198 RMSE: 4.705	p1 = 0.7127 (-0.1073, 1.533) p2 = 0.02622 (-0.6925, 0.7449) p3 = -0.9831 (-2.587, 0.6208) p4 = 27.87 (26.91, 28.83)	SSE: 6498 R-square: 0.01503 Adjusted R-square: 0.003018 RMSE: 5.14
BDA	p1 = 0.2683 (-1.601, 2.137) p2 = 6.47 (4.83, 8.11) p3 = -3.121 (-6.788, 0.5467) p4 = 25.71 (23.52, 27.91)	SSE: 3.406e+04 R-square: 0.2294 Adjusted R-square: 0.22 RMSE: 11.77	p1 = 0.423 (-1.146, 1.992) p2 = 0.8068 (-0.5681, 2.182) p3 = -2.956 (-6.024, 0.1125) p4 = 53.38 (51.54, 55.22)	SSE: 2.378e+04 R-square: 0.05405 Adjusted R-square: 0.04252 RMSE: 9.832
RM	p1 = -47.32 (-79.18, -15.46) p2 = 75.43 (47.48, 103.4) p3 = 96.11 (33.59, 158.6) p4 = 1284 (1247, 1321)	SSE: 9.897e+06 R-square: 0.1325 Adjusted R-square: 0.1219 RMSE: 200.6	p1 = 29.58 (-1.086, 60.25) p2 = 23.64 (-3.241, 50.52) p3 = -59.4 (-119.4, 0.5952) p4 = 1579 (1543, 1615)	SSE: 9.091e+06 R-square: 0.02722 Adjusted R-square: 0.01535 RMSE: 192.2
BMI	p1 = -0.06208 (-0.5373, 0.4131) p2 = 0.9464 (0.5295, 1.363) p3 = 0.09969 (-1.032, 0.8328) p4 = 21.22 (20.66, 21.78)	SSE: 2202 R-square: 0.07985 Adjusted R-square: 0.06863 RMSE: 2.992	p1 = -0.1174 (-0.5875, 0.3527) p2 = 0.229 (-0.183, 0.641) p3 = -0.2751 (-1.195, 0.6444) p4 = 29.69 (29.14, 30.24)	SSE: 2136 R-square: 0.03228 Adjusted R-square: 0.02047 RMSE: 2.946
BPsys	p1 = -1.697 (-3.998, 0.6036) p2 = 4.882 (2.863, 6.901) p3 = 1.013 (-3.503, 5.528) p4 = 118.6 (115.9, 121.3)	SSE: 5.162e+04 R-square: 0.1082 Adjusted R-square: 0.09736 RMSE: 14.49	p1 = -3.266 (-6.051, -0.4809) p2 = -4.609 (-7.05, -2.168) p3 = 3.8 (-1.647, 9.248) p4 = 175.4 (172.1, 178.7)	SSE: 7.495e+04 R-square: 0.08419 Adjusted R-square: 0.07303 RMSE: 17.46
BPdia	p1 = -0.5416 (-1.885, 0.8014) p2 = 1.605 (0.4267, 2.783) p3 = 0.5087 (-2.127, 3.144) p4 = 75.62 (74.05, 77.2)	SSE: 1.758e+04 R-square: 0.03383 Adjusted R-square: 0.02204 RMSE: 8.455	p1 = -0.7762 (-1.937, 0.3843) p2 = -0.04891 (-1.066, 0.9683) p3 = 2.132 (-0.1384, 4.402) p4 = 92.31 (90.95, 93.67)	SSE: 1.302e+04 R-square: 0.01727 Adjusted R-square: 0.005282 RMSE: 7.274
P	p1 = -0.1082 (-1.867, 1.65) p2 = -5.21 (-6.753, -3.667) p3 = 1.457 (-1.994, 4.908) p4 = 86.8 (84.74, 88.87)	SSE: 3.015e+04 R-square: 0.1621 Adjusted R-square: 0.1519 RMSE: 11.07	p1 = -2.221 (-3.595, -0.8473) p2 = -0.7112 (-1.915, 0.493) p3 = 5.636 (2.949, 8.323) p4 = 94 (92.4, 95.61)	SSE: 1.824e+04 R-square: 0.07754 Adjusted R-square: 0.06629 RMSE: 8.611
SF	p1 = -1.584 (-4.957, 1.789) p2 = 12.94 (9.978, 15.9) p3 = -9.565 (-16.18, -2.946) p4 = 80.26 (76.3, 84.22)	SSE: 1.109e+05 R-square: 0.3954 Adjusted R-square: 0.388 RMSE: 21.23	p1 = -1.873 (-6.044, 2.298) p2 = -0.07454 (-3.73, 3.581) p3 = -11.52 (-19.68, -3.362) p4 = 144.8 (139.9, 149.7)	SSE: 1.681e+05 R-square: 0.2487 Adjusted R-square: 0.2395 RMSE: 26.14
SPP	p1 = -2.188 (-5.624, 1.248) p2 = 7.973 (4.958, 10.99) p3 = -4.127 (-10.87, 2.616) p4 = 124.6 (120.6, 128.7)	SSE: 1.151e+05 R-square: 0.2054 Adjusted R-square: 0.1957 RMSE: 21.63	p1 = -0.4375 (-6.284, 5.409) p2 = 7.064 (1.94, 12.19) p3 = -5.259 (-16.69, 6.177) p4 = 175.4 (168.6, 182.3)	SSE: 3.303e+05 R-square: 0.05446 Adjusted R-square: 0.04293 RMSE: 36.64
W	p1 = -1.996 (-3.68, -0.3112) p2 = 4.176 (2.698, 5.654) p3 = 2.442 (-0.8638, 5.747) p4 = 80.25 (78.27, 82.23)	SSE: 2.767e+04 R-square: 0.1382 Adjusted R-square: 0.1277 RMSE: 10.6	p1 = -1.491 (-3.346, 0.3637) p2 = -0.6328 (-2.259, 0.993) p3 = 3.313 (-0.3149, 6.941) p4 = 105.8 (103.7, 108)	SSE: 3.325e+04 R-square: 0.01541 Adjusted R-square: 0.0034 RMSE: 11.63

Table 3: Mathematical Equations for Fitness Level 1 and Fitness Level 2 Curve Fitting

Fitness Level 1: where x is normalized by mean 125.1 and std 71.9 Coefficients (with 95% confidence bounds)
Fitness Level 2: where x is normalized by mean 125.5 and std 72.31Coefficients (with 95% confidence bounds)

It can be seen from Table 3 that in both the cases polynomial degree 3 equation can be used to represent the behavior of each factor and in all cases, the values of x are normalized. The second thing that can be observed from these equations is that the value of R-square remains between 0.05 to 0.33 which shows that the data is quite high in degree of non-linearity, as it is unable to do data fitting. This clearly shows that for constructing a trend in this dataset, there is a need for doing some kind of data

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

smoothing and interpolation. This can also be reaffirmed from the values of the R-squared test and SSE (Sum of Square Error) metric. From all these statistical tests it appears that the machine learning algorithm will have difficulty in identifying instances of each fitness class. Those kernels or function that transform the data into linear form as a preprocessing step may be able to identify data points of these two classes.

6.3 Separation and Overlapping Analysis

Scatter plots [Fig. 2(a) – Fig. 2 (l)] were constructed initially to understand the nature of attributes values distribution and it was found that the Fitness Level 1 and Fitness Level 2 class instance(s) values overlap and it shall be difficult to find hard margins and boundaries between the classes for most of the support vector machine kernel functions.

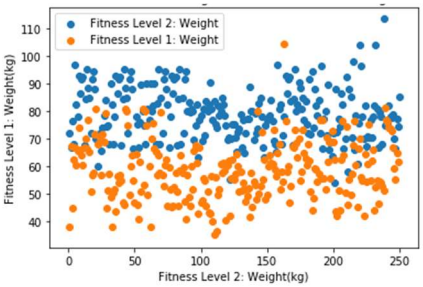

Fig. 2(a) Fitness Level for Weight

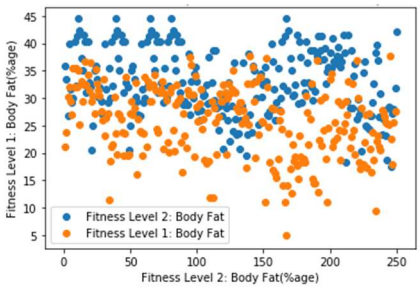

Fig. 2(b) Fitness Level for Body Fat

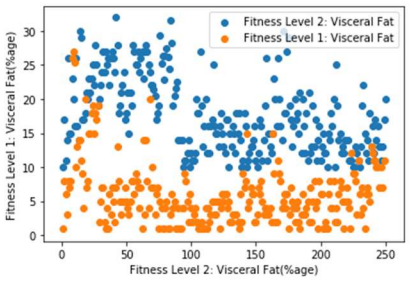

Fig. 2(c) Fitness Level for Visceral Fat

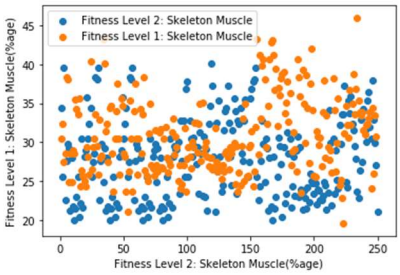

Fig. 2(d) Fitness Level for Skeleton Muscle

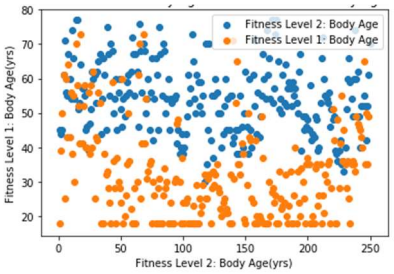

Fig. 2(e) Fitness Level for Body Age

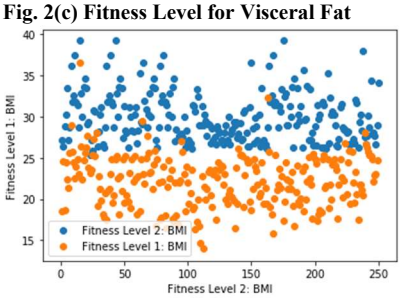

Fig. 2(f) Fitness Level for BMI

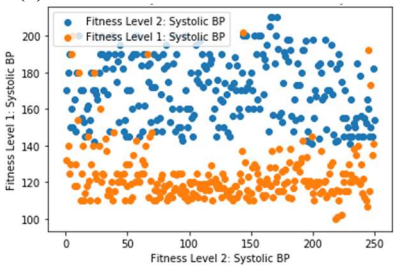

Fig. 2(g) Fitness Level for Systolic BP

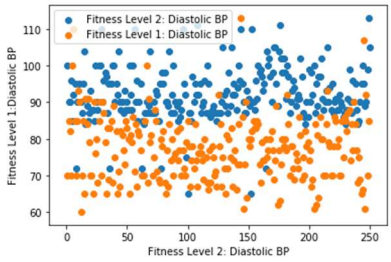

Fig. 2(h) Fitness Level for Diastolic BP

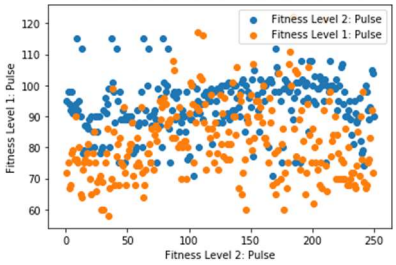

Fig. 2(i) Fitness Level for Pulse

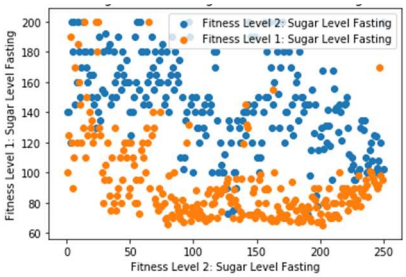

Fig. 2(k) Fitness Level for Blood Sugar PP

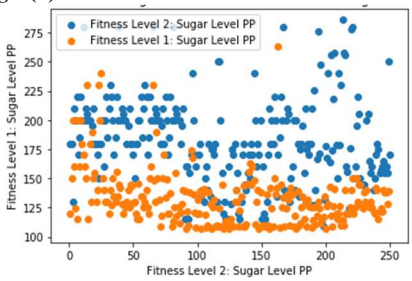

Fig. 2(l) Fitness Level for Waist

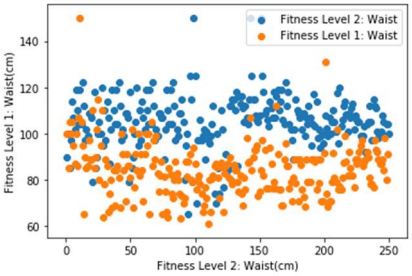

Fig. 2(j) Fitness Level for Blood Sugar Fasting

Fig. 2 [a – l]: Scattered Plot of Different Fitness Attributes

It can be seen that in case of body fat Fig. 2(b), skeleton muscle Fig. 2(d) and pulse Fig. 2(i) attributes the degree of overlapping is too high and in case of weight Fig. 2(a), visceral fat Fig. 2(c) body age Fig. 2(e), BMI Fig. 2(f), BP Fig. 2(g) - Fig. 2 (h), sugar level Fig. 2(j) – Fig. 2(k) and waist Fig. 2(l) the overlapping is moderate. The scatter plot of the feature dataset also showed that some of the data points are forming non-linear trend lines. And apparently, the SVM algorithm might have to deal with many outliers and employ way to ignore these outliers. In Fig. 2 (a) to Fig (l), it can be seen that the SVM will not have a linear hyperplane, to resolve such a condition. the SVM might have to introduce a virtual or additional attribute such that the separation between the classes increases to some extent. All these conditions force us to run more evaluations with many combinations of parameters of each kernel.

6.4 Correlation between the Attributes

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

Theoretically, adding more features should help in improving the discrimination power of the dataset. But this is not always true, especially when the features have some degree of causal relationship computed with the help of correlation. Hence, there is a need to eliminate irrelevant features if required. Correlation is one of the reliable methods to eliminate irrelevant features, especially when the correlation among the attributes is complex and high in nature. However, in the context of the current problem, it is not surprising that the Birth Age and Skeleton Muscle are highly correlated with each other. In fact, the analysis shows that they have a strong relationship among themselves as shown in Table 4.

Pearson correlation coefficient [60]

$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}} \quad (1)$$

where

N = number of pairs of scores.
$\sum xy$ = sum of the products of paired scores.
$\sum x$ = sum of x scores.
$\sum y$ = sum of y scores.
$\sum x^2$ = sum of squared of x scores.
$\sum y^2$ = sum of squared of y scores.

The relationship among the response y= Fitness Level 1| Fitness Level 2 and predictor Birth Age and Skeleton Muscle (92%) are quite high as per the Pearson Correlation method matrix [61]. Hence, depending on which predictor is included in the learning model, the results obtained will lead to different estimation of slope when the SVM kernel will compute the boundaries. After all, the SVM is basically a regression modeling. However, for supervised machine learning tasks, the accomplishment is faulty if there is a correlation between the indicators (Weight, Blood Sugar etc.) A valid feature matrix is the one, which contains a subset of features that are highly correlated with (predictive of) the class, yet has uncorrelated with (not predictive of) each other

	Birth Age	Height	Gender	Weight	Body Fat	Visceral Fat	Skeleton Muscle	Body Age	RM	BMI	BPsis	BPdia	Pulse	Sugar F	Sugar PP	Waist	Group
Birth Age	1	-0.06	-0.26	0.36	0.29	0.53	0.92	-0.78	0.34	0.43	0.39	0.35	-0.04	0.43	0.32	0.35	-0.33
Height	-0.06	1	-0.48	0.45	-0.24	0.06	0.37	0	0.51	-0.03	0.07	0	0	0	0.04	0.12	-0.01
Gender	-0.26	-0.48	1	-0.29	0.44	-0.16	-0.55	-0.15	-0.53	-0.06	-0.14	-0.13	0.15	-0.03	0	-0.09	0.07
Weight	0.36	0.45	-0.29	1	0.48	0.76	-0.2	0.73	0.85	0.87	0.65	0.51	0.27	0.63	0.62	0.71	-0.71
Body Fat	0.29	-0.24	0.44	0.48	1	0.56	-0.79	0.62	0.2	0.67	0.45	0.39	0.25	0.58	0.52	0.51	-0.5
Visceral Fat	0.53	0.06	-0.16	0.76	0.56	1	-0.32	0.81	0.62	0.82	0.71	0.58	0.25	0.77	0.7	0.63	-0.76
Skeleton Muscle	0.92	0.37	-0.55	-0.2	-0.79	-0.32	1	-0.4	0.02	-0.42	-0.22	-0.22	-0.24	-0.37	-0.37	-0.23	0.28
Body Age	0.78	0	-0.15	0.73	0.62	0.81	-0.4	1	0.57	0.82	0.67	0.58	0.2	0.71	0.63	0.65	-0.68
RM	0.34	0.51	-0.53	0.85	0.2	0.62	0.02	0.57	1	0.67	0.47	0.38	0.11	0.52	0.5	0.58	-0.51
BMI	0.43	-0.03	-0.06	0.87	0.67	0.82	-0.42	0.82	0.67	1	0.67	0.57	0.3	0.7	0.66	0.72	-0.79
BPsis	0.39	0.07	-0.14	0.65	0.45	0.71	-0.22	0.67	0.47	0.69	1	0.79	0.47	0.67	0.62	0.67	-0.82
BPdia	0.35	0	-0.13	0.51	0.39	0.58	-0.22	0.58	0.38	0.57	0.79	1	0.39	0.56	0.53	0.51	-0.69
Pulse	-0.04	0	0.15	0.27	0.25	0.25	-0.24	0.2	0.11	0.3	0.47	0.39	1	0.25	0.29	0.28	-0.48
Sugar F	0.43	0	-0.03	0.63	0.58	0.77	-0.37	0.71	0.52	0.7	0.67	0.56	0.25	1	0.81	0.59	-0.67
Sugar PP	0.32	0.03	0	0.62	0.52	0.7	-0.37	0.63	0.5	0.66	0.62	0.53	0.29	0.81	1	0.55	-0.62
Waist	0.35	0.12	-0.09	0.71	0.51	0.63	-0.23	0.65	0.58	0.72	0.67	0.51	0.28	0.6	0.55	1	-0.67
Group	-0.34	-0.01	0.07	-0.71	-0.5	-0.76	0.28	-0.68	-0.51	-0.79	-0.82	-0.69	-0.48	-0.67	-0.62	-0.67	1

Table 4: Correlation between Health Attributes

6.5 Further Inferences and Interpretations

It is always desired that all the variables must have a direct correlation with the response variable (f(y) = Fitness Level 1 | Fitness Level 2), and must have an insignificant correlation among themselves. The observations show in table 4 that SVM kernels will face difficulty due to this to identify the data points of each class. The computation of the Pearson correlation allows the screening of irrelevant noisy attributes. This would ultimately helps in improving the classification accuracy with highly relevant features. According to this correlation matrix attributes that have a correlation of more than 90 % may be eliminated from the current feature set, but at the same time, it raises the question of the elimination process. It may happen such that:

iv. Certain parameters that are significantly relevant may not be medically relevant or dominant in declaring a person/subject category as healthy or unhealthy (*Fitness Level 1 and Fitness Level 2*) [Table 4]. Hence, this can be stated as
P→ "All medically relevant health indicators are statistically not correlated to response variable y = "Fitness Level 1 | Fitness Level 2 "

v. The converse is true i.e certain parameters may be critical in making a medical decision, but statistically, in this specific dataset, they are not relevant.
Q→ "All not correlated health indicators are medically relevant to response variable y = " Fitness Level 1 | Fitness Level 2 "

Therefore, by considering these two statements (P,Q) or arguments, truth table [Table 5] can be constructed to show that , P→Q (P implies Q) and Q→P (Q implies P) need to be evaluated as the selection of attributes for machine learning models based on some mathematical formula with a "fallacy", if medical facts are not considered .

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

P	Q	P→Q	Q→P (converse)
T	T	T	T
T	F	F	T
F	T	T	F
F	F	T	T

Table 5: Truth Table

Let 'STM" be an argument for selecting an attribute of the form P implies Q (P → Q). Then the converse of 'STM' is the statement Q implies P (Q → P). Table 5 shows that the statements (P, Q) and converse of the statements (P, Q) are not logically equivalent unless both terms imply each other. Therefore, for selection of attributes both must be considered. The next section gives information on the selection process based on medical criteria using Delphi method [62] .

6.6 Medical Relevance

A questionnaire [63] was prepared for ranking all the indicators of health on a scale of 10 from three medical experts as per their role and influence in classifying a person/subject to fall under Fitness Level 1 or Fitness Level 2. Table 6 gives the ranking given by each respondent. The modus operandi of collected responses for this questionnaire was based on the Delphi method.

Attributes	**Rater 1**	**Rater 2**	**Rater 3**
Body Age	5	6	8
Height	5	5	5
Gender	10	9	9
Weight	9	9	9
Body Fat	-	-	-
Visceral Fat	9	9	8
Resting Metabolism	0	0	0
Body Mass Index	8	9	10
BP systolic	8	8	9
BP diastolic	9	9	10
Pulse	9	8	8
Sugar Fast	10	10	10
Sugar PP	9	9	9
Waist	7	6	7

Table 6: Rater Score Based on Questionnaire

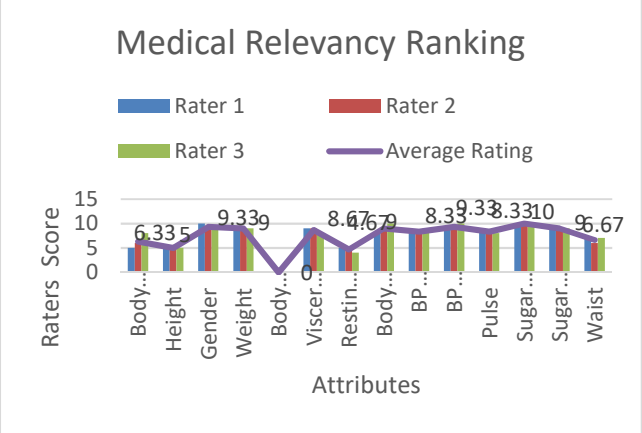

Fig. 3: Medical Relevancy Ranking

The average ranking matrix shows that attributes body age, gender, weight, visceral fat, BMI, BP, pulse, sugar and waist plays a major role in classifying a person/subject under Fitness Level 1 or Fitness Level 2 medically as their score is about 60% of threshold.

Therefore, based on the correlation matrix and medical relevancy test of attributes the final dataset that can be subjected to SVM can be selected using set theory as follows

 Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

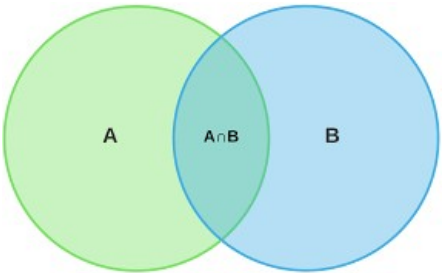

Fig. 4: Venn Diagram showing Common Attributes based on Correlation and Medical Relevancy

A= { *H, G, WT, BF, VF, BDA, RM, BMI, BPsys, BPdia, P, SF, SPP, W* }, where A ϵ feature set selected on the basis of correlation.

B = {*G, WT, VF, BDA, BMI, BPsys, BPdia, P, SF, SPP, W* }, where B ϵ the feature set selected on the basis of medical relevancy test.

A ∩ B = { *G, WT, VF, BDA, BMI, BPsys, BPdia, P, SF, SPP, W* }

6.7 Fine Tuning Learning Models

For classification and prediction tasks the dataset is divided into three parts for ensuring the quality of results. The work begins with a right balance of proportion (50:50) of the features rows of the classes (Fitness Level 1 | Fitness Level 2). This is done to give the machine algorithm a balanced dataset for learning. If a proportion of one class set of data is more, it is likely that the machine learning algorithm may develop a bias in the learning process, and this is done in all other phases (Testing and Validation) of the algorithm. Typically, there is a need to start with default values for each SVM kernel, especially in the context of the Radial/Gaussian Kernel. The values of gamma 'g' parameter play a major role in maintaining a good accuracy at all stages. There is normally a need to scale and standardized the dataset values before they become suitable for SVM algorithms. The default values of the parameters are given in Table 7.

S. No	Parameters	Medium Radial	Polynomial Degree 3	Linear	Polynomial Degree 2	Coarse Radial	Fine Gaussian
1	Kernel offset (Scale)	3.9	1	1	1	15	0.97
2	Standardized Data	Yes					
3	Multiclass method	One-vs-One					

Table 7: Parameters of SVM Algorithms

2. Performance Evaluation and Comparison

This section gives details of six kernel algorithms evaluations and their performances in the context of the problem of health risk classification and prediction.

7.1 Time and speed Analysis of Kernels

I. Training Time

It is time taken by the algorithm to learn from the existing dataset and groups/labels given to each data point. Lower training times is desired:

$$\text{Training Time} = \frac{\sum tt_i}{n} \quad (2)$$

where 'tt' is time taken by SVM kernel to learn from the given data set in seconds and n are the number of observation.

S. No	SVM Classifier Kernel	Training Time (sec)
1	Linear	7.0458
2	Polynomial Degree 2	7.0847
3	Polynomial Degree 3	7.5249
4	Fine Gaussian	7.9644
5	Medium Radial	8.5103
6	Coarse Radial	8.4147

Table 8: Time Analysis of SVM Classifier Kernels

 Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

Fig. 5: Training Time for SVM Classifier Kernel

It is clear from the tabular data [Table 8] and bars graph that Linear kernel is taking minimum time to learn and construct hyperplanes/decision boundaries. This can be attributed to the fact that SVM transforms the data into higher dimension linearly and finding the distance between the said classes is not that hard and time-consuming as compared to other kernels. The Polynomial degree 2 algorithm is performing second and Polynomial degree 3 algorithm is in the third position.

II. Prediction Speed

It is time taken by the algorithm to predict the given dataset with respect to the actual class

S. No	SVM Classifier Kernel	Prediction speed (obs/sec)
1	Linear	6600
2	Polynomial Degree 2	9500
3	Polynomial Degree 3	14000
4	Fine Radial/Gaussian	11000
5	Medium Radial	12000
6	Coarse Radial / Gaussian	12000

Table 9: Kernel Prediction Speed Analysis

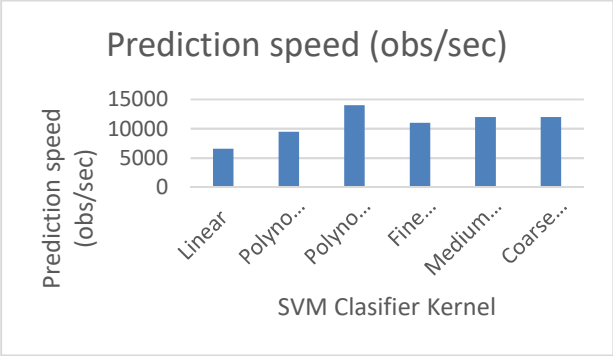

Fig. 6: Prediction Sped for SVM classifier Kernel

$$\text{Prediction time} = \frac{\sum pt_i}{n} \quad (3)$$

where 'pt' is time taken to predict per observation in seconds and n are the number of observation.

Higher prediction time is always desired especially when the prediction is to be done in real time. The tabular data [Table 9] and a bar graph shows that Linear kernel takes 7.0458 seconds to perform 6600 number of observations while polynomial degree 2 takes 7.0847 seconds to perform 9500 numbers of observations and Medium Radial takes 8.5103 seconds to perform 12000 of observations

From all these outcomes, it can be concluded that in terms of training time and prediction speed Medium Radial SVM algorithm performance best.

7.2 Classification Accuracy

It is measured as the fraction of sum of the true positive instances and true negatives instance by total instances in the dataset. Table 10 gives the accuracy of each classifier kernel

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

S. No	SVM Classifier Kernel	Accuracy Percentage
1	Linear	96.0
2	Polynomial Degree 2	97.8
3	Polynomial Degree 3	97.2
4	Fine Gaussian	90.4
5	Medium Radial	98.2
6	Coarse Radial	96.0

Table 10: SVM Classifier Kernel Accuracy

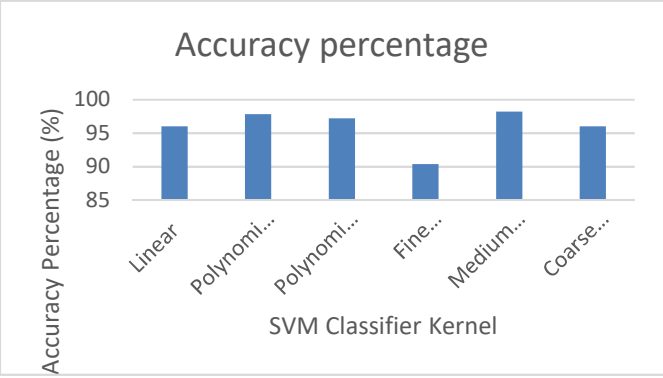

Fig.7: Accuracy Percentage for SVM Classifier Kernel

Interpretation

The Medium Radial SVM is giving the maximum accuracy i.e. 98.2 %, which shows that this kernel is able to identify the boundary between the two classes more distinctively and has the lowest false alarm rate as compared to the other kernels. It further shows that the Medium Radial SVM is able to select and reject the instances as per the actual class more accurately as compared to other kernels. Hence, it can be safely said that the Medium Radial SVM algorithm has the maximum predictive power among Linear, Polynomial degree 2, Polynomial degree 3, Fine Gaussian and Coarse Radial.

7.3 Confusion Matrix

This matrix is also called the contingency matrix. It helps to understand the abilities of the classifier in terms of selecting as well as rejecting instances with respect to each class. If the classifier is able to match the fitness level 1 instance with fitness level 2 instances and predict healthy instances of healthy with healthy class, then the logic of implementing the SVM algorithm will be successful. The results are as follows:

Linear SVM	**Polynomial Degree 2**	**Polynomial Degree 3**	**Fine Gaussian SVM**	**Medium Radial SVM**	**Coarse Radial SVM**
TP=245 FP=5 FN=15 TN=235	TP=247 FP=3 FN=8 TN=242	TP=246 FP=4 FN=10 TN=240	TP=249 FP=1 FN=47 TN=203	TP=249 FP=1 FN=8 TN=242	TP=243 FP=7 FN=13 TN=237
TPR = 0.98	TPR = 0.988	TPR = 0.984	TPR = 0.996	TPR = 0.996	TPR = 0.972
FPR = 0.06	FPR = 0.032	FPR = 0.04	FPR = 0.188	FPR = 0.032	FPR = 0.052
ACC = 0.96	ACC = 0.978	ACC = 0.972	ACC = 0.904	ACC = 0.982	ACC = 0.96
Formulas	$TPR = \frac{TP}{TP + FN}$ TPR – True Positive Rate	$FPR = 1 - \frac{TN}{TN + FP}$ FPR – False Positive Rate	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$ ACC - Accuracy		

Table 11: Performance Evaluation of Six SVM Kernel

Note: TP – True Positive, TN – True Negative, FP – False Positive, FN – False Negative

Interpretation

It is clear from Table 11 that Medium Radial kernel and Fine Gaussian Kernel is producing the maximum number of true positives (>99%) and lowest false alarms. From this value, it can be safely interpreted that there are <1% chances that this classifier might go wrong, which is the lowest among all the other classifiers.

7.4 Area Under the Curve

 Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

The "Area under the curve" is also known as C-statistics. It gives the likelihood that the classifier in question will rank (or assign score) a randomly selected positive instance, higher than the randomly selected negative one. Higher the AUC value, higher is the probability of the positive instances and consequently higher is the chance of accuracy. [Table 12]

S.No	SVM Classifier Kernel	AUC
1	Linear	0.99
2	Polynomial Degree 2	1.0
3	Polynomial Degree 3	0.99
4	Fine Gaussian	0.99
5	Medium Radial	1.0
6	Coarse Radial	0.99

Table 12: Area Under the Curve for SVM Classifier Kernel

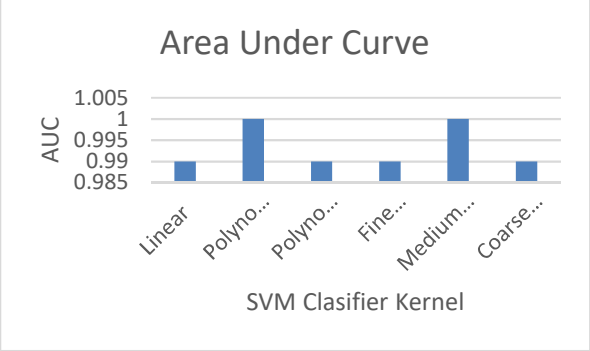

Fig.8: Area Under Curve Graph for SVM Classifier Kernel

Interpretation

It is clear from the AUC value of Polynomial Degree 2 SVM and Medium Radial SVM has the highest probability of getting positives and the lowest possibility of getting false alarm rate. The outcome of this metric is consistent with respect to the accuracy and values of the true positive rates. In this case, also, the Polynomial Degree 2 SVM and Medium Radial SVM is performing well in terms of Area Under the curve (1.0).

Conclusion

This paper work validates the latest finding in the current literature on the quality of dataset, feature sets and process of selection of machine learning algorithms i.e all aspects of the process are critical for the successful implementation and application of the machine learning algorithms for problems of classification and prediction. Since machine learning algorithm takes its inspiration from statistics as well as from the pattern finding algorithm. This paper focused on both the aspects. The evaluation strategy followed in feature selection in this paper can be operated independently of any learning algorithm as it helps to remove insignificant features before the learning process begins. The feature selection process goes through multiple revisions to arrive at the final feature set that leads to the high accuracy of the six algorithms.

The classification and the prediction process goes through a series of evaluations and reruns of the SVM with six kernels. The Medium Radial kernel of the SVM gave the maximum prediction power on the dataset. It performed well in almost all evaluations, whether it was the values of true positive rate or Area Under a Curve (AUC) is 0.1. It was empirically found that the main advantage of using SVM kernel functions is that it allows us to change Euclidean geometry so that it fits into the context of the problem and the dataset becomes workable for classification.

SVM classification algorithms are useful in cases where the regression method works well to identify the trend lines between the datasets. And it may fail in cases where there are noise and a lot of overlapping data points in the dataset. The initial statistical analysis of the feature set showed some degree of overlapping but the trend lines show a good degree of separation between the classes. Secondly, it was found that, for finding "best" separation hyper-plane, quantity (number of Instances of features) did not always convert into quality, the advantage of SVM is that it does not require a large number of instances to produce a fair percentage of accuracy as it is clear from the outcomes.

Future Scope

This paper gives computing model of Health Risk Assessment using learning models that are based on six kernel tricks. The extension of this work can be done by converting these dataset group into a multiclass risk assessment problem. The health risk classification may further be divided into three class problem (Fitness Level 1, Fitness Level 2 and Fitness Level 3). Second, for further direction, this dataset and its features may use for formulating health insurance policies. Insurance companies can leverage the use of machine learning algorithms for computing health plus insurance risks.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

References

[1] A. Suresh and R. K. R. Varatharajan, "Health care data analysis using evolutionary algorithm," *J. Supercomput.*, 2018.

[2] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017.

[3] E. Avci and A. D. Extraction, "Performance Comparison of Some Classifiers on Chronic Kidney Disease Data," 2018.

[4] V. Nagendra, H. Gude, D. Sampath, S. Corns, and S. Long, "Evaluation of support vector machines and random forest classifiers in a real-time fetal monitoring system based on cardiotocography data," *2017 IEEE Conf. Comput. Intell. Bioinforma. Comput. Biol. CIBCB 2017*, 2017.

[5] C. Y. Hung, W. C. Chen, P. T. Lai, C. H. Lin, and C. C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 3110–3113, 2017.

[6] A. Ben-Hur and J. Weston, "A User's Guide to Support Vector Machines," Humana Press, 2010, pp. 223–239.

[7] D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare," *Int. J. Bio-Science Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013.

[8] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the Circulant Structure of Tracking-by-Detection with Kernels," Springer, Berlin, Heidelberg, 2012, pp. 702–715.

[9] C. Chang and C. Lin, "LIBSVM : A Library for Support Vector Machines," *ACM Trans. Intell. Syst. Technol.*, 2011.

[10] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua, "Bridging the Vocabulary Gap between Health Seekers and Healthcare Knowledge," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 396–409, Feb. 2015.

[11] C. K. I. Williams, "Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond," *J. Am. Stat. Assoc.*, 2003.

[12] M. Bloodgood, "Support Vector Machine Active Learning Algorithms with Query-by-Committee Versus Closest-to-Hyperplane Selection," in *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, 2018, pp. 148–155.

[13] S. Maldonado and J. López, "Synchronized feature selection for Support Vector Machines with twin hyperplanes," *Knowledge-Based Syst.*, vol. 132, pp. 119–128, Sep. 2017.

[14] D. Conforti and R. Guido, "Kernel based support vector machine via semidefinite programming: Application to medical diagnosis," *Comput. Oper. Res.*, vol. 37, no. 8, pp. 1389–1394, Aug. 2010.

[15] F. S. G, O. Olumide, J. Sadiku, and G. J. A, "Application of Data Mining Technique for Fraud Detection in Health Insurance Scheme Using Knee-Point K-Means Algorithm Redeemer ' s University Mowe Ogun State Nigeria University of Ilorin Nigeria," vol. 7, no. 8, pp. 140–144, 2013.

[16] C. Francis, N. Pepper, and H. Strong, "Using support vector machines to detect medical fraud and abuse," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 8291–8294.

[17] G. G. Sundarkumar, V. Ravi, and V. Siddeshwar, "One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection," in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2015, pp. 1–7.

[18] Han Tao, Liu Zhixin, and Song Xiaodong, "Insurance fraud identification research based on fuzzy support vector machine with dual membership," in *2012 International Conference on Information Management, Innovation Management and Industrial Engineering*, 2012, pp. 457–460.

[19] A. Dubey, T. Parida, A. Birajdar, A. K. Prajapati, and S. Rane, "Smart Underwriting System: An Intelligent Decision Support System for Insurance Approval & Risk Assessment," in *2018 3rd International Conference for Convergence in Technology (I2CT)*, 2018, pp. 1–6.

[20] Yi Tan and Guo-Ji Zhang, "The application of machine learning algorithm in underwriting process," in *2005 International Conference on Machine Learning and Cybernetics*, 2005, p. 3523–3527 Vol. 6.

[21] S. Aftab, W. Abbas, M. M. Bilal, T. Hussain, M. Shoaib, and S. H. Mehmood, "Data mining in insurance claims (DMICS) two-way mining for extreme values," in *Eighth International Conference on Digital Information Management (ICDIM 2013)*, 2013, pp. 1–6.

[22] P. Saripalli, V. Tirumala, and A. Chimmad, "Assessment of healthcare claims rejection risk using machine learning," in *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 2017, pp. 1–6.

[23] Y. H. Chong *et al.*, "Type 2 Diabetes Genetic Variants and Risk of Diabetic Retinopathy," *Ophthalmology*, vol. 124, no. 3, pp. 336–342, Mar. 2017.

[24] B. Zheng, J. Zhang, S. W. Yoon, S. S. Lam, M. Khasawneh, and S. Poranki, "Predictive modeling of hospital readmissions using metaheuristics and data mining," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 7110–7120, 2015.

[25] A. Georgia, "Health Risks , Nutrition Assessments and Disease Prevalence Among African Immigrant Groups in," 2013.

[26] F. Friedrichs and C. Igel, "Evolutionary tuning of multiple SVM parameters," *Neurocomputing*, vol. 64, pp. 107–117, Mar. 2005.

[27] M. Kawulok and J. Nalepa, "Towards robust SVM training from weakly labeled large data sets," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 464–468.

[28] M. Kawulok and J. Nalepa, "Support Vector Machines Training Data Selection Using a Genetic Algorithm," 2012, pp. 557–565.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

[29] J. Balcázar, Y. Dai, and O. Watanabe, "A Random Sampling Technique for Training Support Vector Machines," 2001, pp. 119–134.
[30] H. Jiang and W. K. Ching, "Correlation kernels for support vector machines classification with applications in cancer data," *Comput. Math. Methods Med.*, vol. 2012, 2012.
[31] S. Ali and K. A. Smith-Miles, "A meta-learning approach to automatic kernel selection for support vector machines," *Neurocomputing*, vol. 70, no. 1–3, pp. 173–186, Dec. 2006.
[32] N. Cristianini and B. Scholkopf, "Support vector machines and kernel methods - The new generation of learning machines," *AI Mag.*, vol. 23, no. 3, pp. 31–41, 2002.
[33] F. Angiulli and A. Astorino, "Scaling Up Support Vector Machines Using Nearest Neighbor Condensation," *IEEE Trans. Neural Networks*, vol. 21, no. 2, pp. 351–357, Feb. 2010.
[34] M. Barros de Almeida, A. de Padua Braga, and J. P. Braga, "SVM-KM: speeding SVMs learning with a priori cluster selection and k-means," in *Proceedings. Vol.1. Sixth Brazilian Symposium on Neural Networks*, pp. 162–167.
[35] R. Li, B. Bhanu, and K. Krawiec, "Hybrid coevolutionary algorithms vs. SVM algorithms," in *Proceedings of the 9th annual conference on Genetic and evolutionary computation - GECCO '07*, 2007, p. 456.
[36] M. Wistuba and A. Rawat, "Scalable Multi-Class Bayesian Support Vector Machines for Structured and Unstructured Data."
[37] S. M. Alzahani, A. Althopity, A. Alghamdi, B. Alshehri, and S. Aljuaid, "An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction," *Lect. Notes Inf. Theory*, vol. 2, no. 4, pp. 310–315, 2015.
[38] "Connected Digital Health and Wellbeing Platform and System," Aug. 2016.
[39] S. Danagoulian, "Taking the hassle out of wellness: Do peers and health matter?," *Int. J. Heal. Econ. Manag.*, vol. 18, no. 1, pp. 1–23, Mar. 2018.
[40] P. Zane Pilzer, *The New Wellness Revolution Second Edition How to Make A Fortune in the Next Trillion Dollar Industry*. 2002.
[41] D. Lakdawalla, A. Malani, and J. Reif, "The insurance value of medical innovation," *J. Public Econ.*, vol. 145, pp. 94–102, 2017.
[42] B. A. Goldstein, A. M. Navar, and R. E. Carter, "Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges," *Eur. Heart J.*, vol. 38, no. 23, pp. 1805–1814, 2017.
[43] R. P. Ellis and T. J. Layton, "Risk Selection and Risk Adjustment," *Encycl. Heal. Econ.*, pp. 289–297, 2014.
[44] M. S. Hossain and G. Muhammad, "Cloud-assisted Industrial Internet of Things (IIoT) – Enabled framework for health monitoring," *Comput. Networks*, vol. 101, pp. 192–202, Jun. 2016.
[45] M. Hassanalieragh *et al.*, "Health Monitoring and Management Using Internet-of-Things (IoT) Sensing with Cloud-Based Processing: Opportunities and Challenges," in *2015 IEEE International Conference on Services Computing*, 2015, pp. 285–292.
[46] N. Sultan, "Reflective thoughts on the potential and challenges of wearable technology for healthcare provision and medical education," *Int. J. Inf. Manage.*, vol. 35, no. 5, pp. 521–526, Oct. 2015.
[47] M. McCarthy and P. Spachos, "Using mobile environment sensors for wellness monitoring," in *2016 IEEE 21st International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (CAMAD)*, 2016, pp. 135–139.
[48] M. McCarthy and P. Spachos, "Wellness assessment through environmental sensors and smartphones," in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.
[49] M. Acikkar, M. F. Akay, K. T. Ozgunen, K. Aydin, and S. S. Kurdak, "Support vector machines for aerobic fitness prediction of athletes," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3596–3602, Mar. 2009.
[50] Y.-J. Son, H.-G. Kim, E.-H. Kim, S. Choi, and S.-K. Lee, "Application of support vector machine for prediction of medication adherence in heart failure patients.," *Healthc. Inform. Res.*, vol. 16, no. 4, pp. 253–9, Dec. 2010.
[51] omron, "Omron Healthcare Wellness & Healthcare Products," *Private*, 2018. [Online]. Available: https://omronhealthcare.com/. [Accessed: 24-Dec-2018].
[52] A. Singh and K. Ramkumar, "Health Risk Dataset." mendelay, 2018.
[53] T. Razzaghi, O. Roderick, I. Safro, and N. Marko, "Fast Imbalanced Classification of Healthcare Data with Missing Values."
[54] "BMI." [Online]. Available: https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/. [Accessed: 17-Dec-2018].
[55] "BP." [Online]. Available: https://www.nhs.uk/conditions/high-blood-pressure-hypertension/. [Accessed: 17-Dec-2018].
[56] "Vital Signs (Body Temperature, Pulse Rate, Respiration Rate, Blood Pressure) | Johns Hopkins Medicine Health Library." [Online]. Available: https://www.hopkinsmedicine.org/healthlibrary/conditions/cardiovascular_diseases/vital_signs_body_temperature_pulse_rate_respiration_rate_blood_pressure_85,p00866. [Accessed: 24-Dec-2018].
[57] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.
[58] A. Salimi, M. Ziaii, A. Amiri, M. Hosseinjani Zadeh, S. Karimpouli, and M. Moradkhani, "Using a Feature Subset

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

Selection method and Support Vector Machine to address curse of dimensionality and redundancy in Hyperion hyperspectral data classification," *Egypt. J. Remote Sens. Sp. Sci.*, vol. 21, no. 1, pp. 27–36, Apr. 2018.
[59] Bor-Chen Kuo, Hsin-Hua Ho, Cheng-Hsuan Li, Chih-Cheng Hung, and Jin-Shiuh Taur, "A Kernel-Based Feature Selection Method for SVM With RBF Kernel for Hyperspectral Image Classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 7, no. 1, pp. 317–326, Jan. 2014.
[60] M. M. Mukaka, "81576-194640-1-Pb," vol. 24, no. September, pp. 69–71, 2012.
[61] "The Pearson Correlation Coefficient - Statistics in a Nutshell, 2nd Edition [Book]." [Online]. Available: https://www.oreilly.com/library/view/statistics-in-a/9781449361129/ch07.html. [Accessed: 24-Dec-2018].
[62] E. Rosowsky, A. S. Young, M. C. Malloy, S. P. J. van Alphen, and J. M. Ellison, "A cross-validation Delphi method approach to the diagnosis and treatment of personality disorders in older adults," *Aging Ment. Health*, vol. 22, no. 3, pp. 371–378, Mar. 2018.
[63] A. Singh and K. . Ramkumar, "Medical Health Status Parameter Survey Questioner." mendelay, 2018.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495