International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Validation of Machine Learning Models for Health Insurance Risks Assessment

Amrik Singh, K R Ramkumar

Abstract*: A universal healthcare policy success is impossible without the use of insurance instruments. The healthcare and insurance industries are on the verge of integrating seamlessly with the help of sensors and algorithms. This research work focuses on validating an algorithm that can help to model and classify health insurance risk data. Six algorithms Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Naive Bayes (NB) and Support Vector Machine (SVM) were evaluated and objective validation of these algorithms has been demonstrated. To maintain the replicability of the study the data and code are available in public repository. From the study, it is clear that the KNN algorithm is best suited as a risk classifier. This is evidence from the values of R^2, error metrics, completeness score, explained variance, normalized mutual score v measure score, precision, recall, f1 score, and accuracy metrics. Secondly, the algorithms have been validated using 10 k-fold method using five types of performance metrics. In almost all cases, it was found that the KNN algorithm performs consistently and is the most suitable numerically. This can be attributed that the standard deviation remains tight of performance metrics in evaluation. From all the validation test, it can be claimed that on the current dataset, the KNN algorithm with Accuracy, Homogeneity Score Explained variance and Normalized mutual score hyper-parameter configuration is the best performer.

Keywords* : Subjective Validation, Objective Validation, Health Insurance Risks.

I. INTRODUCTION

Many governments are finding hard to deliver advanced medical technological innovation [1] to masses. This is due to the fact, that building hospitals and medical facilities require a lot of time, investment and politized will. One of the easy solutions followed by many governments across the world to give a 'health insurance' policy or scheme to the individual [2][3][4][5][6]. So that moral and financial obligations regarding taking care of its citizen is off-loaded to an 'individual' himself / herself. But in a country, such as India, navigating efforts through the administrative procedures to acquire government benefits takes its toll and defeats the purpose of off-loaded responsibility of maintaining healthcare with individuals. The affordable health in India is mainly provided by the Indian governments (88%), and private player only serves the wealth section of Indian's population. The number of people who are covered under the health insurance scheme in India is close to 20 % [7] [8] [9] [10]. An average 86 % rural and 82% urban population are not covered. The cost of insurance is computed on the basis of healthcare cost mainly[11]. The healthcare cost has been steadily increasing all stakeholders are finding hard to justify the cost. To innovate to reduce the cost and move ahead with better 'health insurance' facilities remain a challenge. This is due to the fact, the dynamic and anatomy of 'risks' keeps changing rapidly with time. The way to identify and classify new risk vectors is becoming a challenging task day by day.

Insurance organization are now collaborating with diagnostics centers, fitness centers and even with an organization that is promoting healthy food and dancing etc. simply because they want to leverage digital performs and workflows to integrate with them to find lead to improve their business. But, the issue of computing risk becomes more complex as they need to cater to a wider audience. An audience, that might be suffering from new kinds of disease vectors such as Cardiovascular, breast cancer etc. [12][13]. A new base of customer that are highly prone to 'arsenic' posing as they live in an area where water is highly contaminated with lead and arsenic.

The insurance companies need to now retrieve and redesign new categories of policies as the growth of medical / fitness sensor grow [14] and availability of data is not an issue. More and more people are willing to share their health / fitness data so that preventive measures can be suggested to them. Due to this paradigm shift, the relevance of fitness trackers, portable sleep analyzers, blood pressures instruments, sugar / glucometers, oximeter, skin analyzer, urine analyzer becomes significant for the insurance companies also. An insurance company can use mobile and cloud network to build networks and data repository for developing new kinds of insurance scheme that can use real-time data and algorithm to compute risk and design the premium components of the health insurance. The analysis of the current trends in the context of health insurance shows that private players shall keep on playing a major role in providing health insurance coverage to people, but now with the use of remote services [15] [16] and cloud networks [17] they can now cater to the weaker section of the society also. This is possible by using a new class of algorithms that can learn and discover new insights and patterns from the data. Data Mining and knowledge discovery algorithms [18] can help them to analyze the risk pattern and vectors at individual levels. The machine learning algorithm such as Neural Network, SVM, [19][20][21][22] etc. can help to automate the process of classifying new categories of risk. The section discusses the initiative taken by the researches in this context.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Amrik Singh*, Department of Computer Science Engineering, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India. Email: amriksingh07@gmail.com.
K R Ramkumar, Department of Computer Science Engineering, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India. Email: k.ramkumar@chitkara.edu.in.

II. LITERATURE REVIEW

Validation is a full-length expression for a sequence of algorithmic steps taken in order to demonstrate and accord specific algorithm that can be reliably used for the said objectives of a system under construction. The outcome and quality of performance cannot adequately assured, just by doing limited numbers of trials, inspections and verifications, there is always a need to pre-determine limitation as well as the outcome of the algorithm. The validation process [23][24] gives the chance to evaluate data and algorithm right from the design stage till production, which establishes scientific evidence that a process is capable of consistently delivering quality in terms of performance. The validating process helps to determine appropriate process factors and also the necessary factors that need to be constructed before they are released in the market. It helps to confirm repeatability and reproducibility of the system and at the same time helps to identify the worst-case scenarios. The process helps to determine variability within and between the various approaches and methods. Higher level of scrutiny at each stage of system building helps to safeguard deviation from the standardized product or system. A quick survey of the tools and literature using machine learning show that python and R language libraries are mostly used in machine learning [25][26]. Different libraries and tools targets and focus on different categories of machine learning problems in terms of scale, application, ease of use and learning curves. Scikit-learn provides a high degree of flexibility and focuses on all types of procedures that are required to pre-process and post-process the data required for running supervised / unsupervised machine learning models [27] [28]. It provides a number of methods and procedures for benchmarking and validating the algorithm [29][30] [31]. The current articles from the journals reveal that validation can be done mainly in two ways. First, is the subjective evaluation and method [32]. In this domain, experts and people with related experience are involved to review, evaluate and validate the outcomes of research work and the second method is by using statistical tests and metrics for evaluating the quality of results of algorithms [33]. Documentary and empirical evidence show that subjective evaluation is prone to inter-reliability issues and objective evaluation sometimes do not comprehensively cover all aspects of performance of machine learning models. The objective of validation method such as cross validation, Monte Carlo validation etc [34]. helps to address the problems of optimizing and finding the adjustable parameters of machine learning models. These helps in identification of appropriate machine learning models. The objective of validation is to give a chance to build system that are expert and independent with fair degree of objectiveness i.e there is no bias and inter-rater dis-agreement issues.

The field of health insurance is dominated by the policies that derive their essence from the health and clinical data. It is rare that the insurers uses a different way of computing and formulating policies. Most of the policies are made for masses i.e one for all. It is now, with advent of medical fitness devices and cloud technologies [35] , the other health parameters such as degree of physical mobility of person, degree of physical activeness and behavior aspects are been considered. An integrated system that can collect data on sleep, urine, behavior, blood, respiration etc. is still an idea that needs attention. The integration of instrument that can collect health as well as insurance related data is still a challenge. There is a tendency in the industry to shift the paradigm towards wellness insurance that ensures the use of parameters that are not necessarily clinical data. It can also be observed from the current journal that, there are citation on the use and applications of evaluation metrics such as recall, precision, accuracy, f-score etc. for understanding the models of health and insurance. These models / systems of health and insurance are using real time data to analyze risk based clinical as well as behavioral data. The analyzes helps to formulate highly personalized health insurance policies. The authors [36] are using all the matrics that are computed under the 'Receiver Operating Characteristic' (ROC). The graph drawn on the basis of ROC gives graphical information on the behavior of machine learning when it is tested for rightly accepting or rejecting the instance variables for particular class. The authors are also objectively finding ways and methods to find how fast the machine learning algorithms are in prediction and classification of the task. This way they are able to optimize and identify algorithms that suits particular needs. For example, while predicting heart failure the machine learning algorithm should have pico-second level of response in predicting task. But in a task, such as classification or prediction of insurance risk the system may tolerate some level of prediction latency. Due to this factor, authors [37] have considered this aspects of evaluating the model important beside accuracy. It was further found that many researchers preferred to evaluate and validate each stage of algorithm or machine learning modelling. They are not interested in only the final outcome or performance of the algorithm. This is because evaluating the full pipeline of operation and algorithms will ensure that in production stage the system does not create false alarms and reduces the accuracy of system in final performance. It can be often observed that in many studies the researches are evaluating the learning / training phase as well, along with the testing / validation phase. The most popular method for doing health insurance classification are neural network [38][39]. Most of the work is directed towards building life insurance policies. Researcher are infact, trying to automate the life risk factors calculations by using supervised learning algorithms and their focus is not studying the health insurance.

Extensive work has been found that demonstrates the computations of "health risk" based on particular medical condition only and limited attention is given to "health risk insurance".

III. RESEARCH METHODOLOGY

This section gives all the steps and procedures done to achieve aimed information and goal of this research work, such as missing data treatment and procedures to maintain data quality of the health insurance risk data. For a better understanding of the research methodology Fig. 1 can be referred.

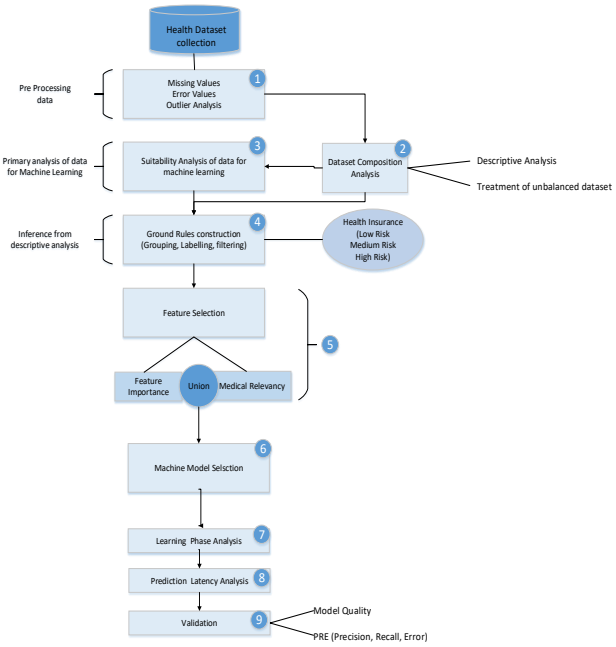

Fig.1 Steps of Research Methodology

IV. DATASET COMPOSITION ANALYSIS (DCA)

The suitability of the dataset for the machine learning algorithm is analyzed in this section using the process of dataset composition analysis.

The health risk dataset consists of sixteen attributes or the indicators of the health risk such as Birth Age(BA), Height(H), Gender(G), Weight(WT), Body Fat(BF), Visceral Fat(VF), Skeleton Muscle(SM), Body Age(BDA), Resting Metabolism(RM), Body Mass Index (BMI), Blood, Pressure Systolic(BPsys), Blood Pressure Diastolic(BPdia), Pulse(P), Sugar Fasting(SF), Sugar Postprandial(SPP) and Waist (W). The values were collected from clinical laboratories and hospitals located in Jammu, India. The numbers of instances collected were over 1000 but after following data quality guidelines the dataset of 823 instances [40] have been used in this research work and it is referred as 'Base Dataset' for building risk models in the context of our research study. The reason to refer, this as 'Base Dataset' is because it has an almost equal proportion of risk class instances and does not contain any missing or any duplicate values. The final objective of data collection is to leverage machine learning algorithms for conducting an automated health insurance risk assessment after identifying mathematical relationships between factors.

Table.1Medical standard classification of the risk level of the subject under observation

Attribute	Range value(s)		Risk Class
BMI [41]	18.5 – 25		LR
	25 – 30		MR
	> 30		HR
BP [42]	**Systolic**	**Diastolic**	
	< 120	65 - 80	LR
	> 140	> 90	MR
	> 180	> 120	HR
P [43]	60 - 100		LR
	100 – 120		MR
	> 120		HR
BF (%age) [41]	**Gender**	**Range**	
	Male	10.0 – 19.9	LR
		20.0 – 24.9	MR
		25.0 – 50.0	HR
	Female	20.9 – 29.9	LR
		30.0 – 34.9	MR
		35.0 – 50.0	HR
VF(%age) [44]	1 – 9		LR
	10 - 14		MR
	> 15		HR
SM(%age)	**Gender**	**Range**	
	Male	32.9 – 35.7	LR
		35.8 – 37.3	MR
		> 37.3	HR
	Female	25.9 – 27.9	LR
		28.0 – 29.0	MR
		>29.0	HR
BDA(Yrs.)	± 3		LR
	± 5		MR
	± 10		HR
Blood Sugar (mg/dL) [45]	**Fasting**	**After Meal (PP)**	
	80 – 120	120 - 140	LR
	120 – 140	141 – 200	MR
	>140	> 200	HR
Waist (cm) [46]	(Height (cm) / 2)	+ (0.03 * Height)	LR
	(Height (cm) / 2)	> (0.03 * Height)	MR
	(Height (cm) / 2)	> (0.05 * Height)	HR

LR – Low Risk, MR- Medium Risk, HR – High Risk

Table I gives information on the medical conditions and the logic based on which the risk indicators data were grouped and satisfied. Care has been taken to use universally accepted medical standard, definitions, and ranges for each health insurance risk indicator recorded. The process of grouping was done by converting the value (continuous) of indicators into discrete bags/buckets using range from Table I as per risk class (LR, MR, and HR). The Table II shows the discriptive statistics of dataset with their mean, standard, minimum and maximum deviation of each attribute of health risk data.

Binning of Dataset

The health insurance risk dataset is a continuous data series. It was converted into a discrete data series using the logic constructed based on medical limits. Binning is a process of segregating the data base on the intervals defined by rules. The definition of the 'intervals' in context of the problem undertaken is 'risk levels' defined as low, medium and high risk. The preliminary analysis shows that the ratio between the Low, medium and high risk data point is somewhat balanced. There is no need to treat unbalance dataset processing that involves resampling from the majority and minority class.

Therefore, we shall proceed towards conducting feature engineering [47][48] so that only appropriate features may be selected.

In routine, it is expected that there is good degree of separation and low degree of overlapping in a data. It is also expected that the level of homogeneity or similarity between the risk classes is also low, so that it is easy for the classifier to differentiate between the various classes. The metric Accuracy, Homogeneity Score Explained variance and Normalized mutual helps to find similarity between each class group. In classification it is also desired that the similarity between the group is low. Based on these four metrices it was found that the data is suitable for classification task The outcome of this test shows that the dataset is suitable for machine learning modelling.

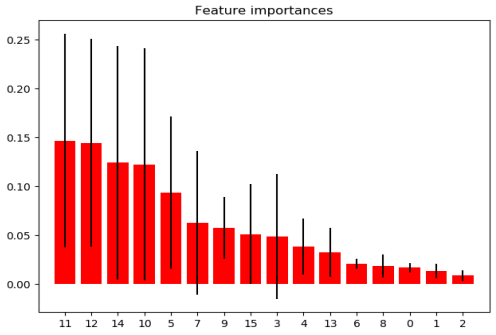

Fig.2 Feature selection

Note: 0 = BA, 1 = H, 2 = G, 3 = WT, 4 = BF, 5 = VF, 6 = SM, 7 = BDA, 8 = RM, 9 = BMI, 10 = BPsys, 11= BPdia,, 12 = P, 13 = SF, 14 = SPP, 15 = W

Table.2 Description statistics of Dataset

	BA	H	G	WT	BF	VF	SM	BDA	RM	BMI	Bpsys	Bpdia	P	SF	SPP	W
count	823															
mean	35.46	161.73	0.49	67.84	28.26	12.71	32.48	41.37	1445.1	25.85	158.20	92.91	98.75	115.94	167.39	93.05
std	13.42	8.36	0.50	15.40	7.45	8.10	6.25	16.13	238.12	5.28	52.45	20.25	20.69	36.84	54.45	13.08
min	16	138	0	35.1	8.6	1	18.3	18	916	14	78	45	52	65	75	63.5
max	80	185	1	113.5	46.9	30	55.2	84	2230	46.9	310	160	145	260	325	150

V. FEATURE SELECTION

In this section with the help of domain knowledge (medical) and insurance experts, relevant feature selection is done. This section is divided into two subparts. The first one discusses the features selection procedures[49] based on a statistical test for importance and the second part deals with features selection based on medical relevancy. The outcome of the procedure is a highly relevant feature with good discriminant power.

There are predominantly three ways to do model selection [50]. The first one is “wrapper method”. second is the “filter method” and third one is “embedded method”. The wrapper method model selection uses ‘out-of-sample” method to select feature and model. The filter method is not applied to learning algorithm on original data, but only consider statistical characteristics of the input data. The embedded method of model selection performs feature selection as part of modeling process. In this research work we are using a filtering method called “forests of trees”[51].

Based on the “forest of tree” feature selection it has been observed that some of the irrelevant features need be eliminated as per the bar graph in the Fig. 2 i.e 2 = G, 1 = H, 0 = BA, 8 = RM and 6 = SM. Therefore, final relevant attributes are shown in set A.

A= {WT, BF, VF, BDA, BMI, BPsys, BPdia, P, SF, SPP, W} where A belongs to feature set selected on the basis of “forests of trees” feature selection.

The method followed to find feature ranking is called “forests of trees”. The importance ranks value is computed on the basis of inter-tree variability. Sum of square method is used to compute variance. The next subsection gives information on how the risk attributes are medically considered.

Medically examined risk factors

A questionnaire was prepared for ranking the health risk factors on a scale of 0 to 10 from three medical experts as per their role and influence in classifying a subject to fall under different Fitness/Health Levels (Low, Medium, High). Fig. 3 gives the ranking given by each respondent. The modus operandi of collected responses for this questionnaire was based on the Delphi method [52].

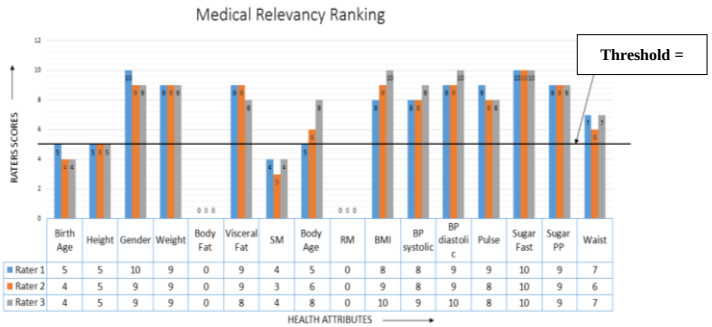

Fig.3 Medical relevancy for feature selection

The average ranking matrix shows that following set 'B' of attributes plays a major role in classifying a subject under Health/Fitness Levels (Low / Medium / High) medically as their score is about 50% of the threshold as indicated by the horizontal line in Fig. 3.

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

Therefore, based on the feature selection by machine learning and medical relevancy test the final attributes can be selected using set theory as

A= {WT, BF, VF, BDA, BMI, BPsys, BPdia, P, SF, SPP, W} where A ϵ feature set selected on the basis of Random Forest Tree selection.

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 U B = {G, WT, BF, VF, BDA, BMI, BPsys, BPdia, P, SF, SPP, W}

The selection of features is complete, it is time to discuss the selection of machine learning model for finally constructing automated health insurance risk assessment system. The next section gives an exact information on how different algorithms are evaluated to find the best one.

VI. MACHINE MODEL SELECTION

In this section, all the five algorithms are evaluated on the basis of four performance metrices. The outcome of the evaluation process is to identify that model, whose performance is stable and consistent on mutually exclusive test/validation datasets [53]. Background to understand the selection metrics are as under By using these metrices as criteria for model selection, following results were demonstrated with 16 and 12 features levels.

A.Accuracy

It is apparent from Table III that DT algorithm is showing 0.91 maximum accuracy. This may be attributed to the fact that best feature selection mechanism is embedded in machine learning modelling process of DT. SVM performs poorly while LR is giving accuracy between 0.85 to 0.86. The KNN and NB is second best in the performance.

Table.3 Considering all the 16 attributes of the health insurance risk

Metrics	LR	KNN	DT	NB	SVM
Accuracy	0.8570	0.8676	0.9164	0.8692	0.3982
Homogeneity score	0.5944	0.6062	0.7168	0.6091	0.0072
Explained variance	0.6788	0.6757	0.7508	0.6885	-0.0221
Normalized mutual score	0.5960	0.6097	0.7249	0.6088	-0.7981

There is quite a notable change happens when four features are dropped Table IV. The KNN algorithm becomes the best performer and its results are consist in 10 evaluation runs. The performance of LR drops by 2% and there is slight drop of 1% in performance of DT. And NB performance also improves by 1% but is still lower than the KNN algorithm accuracy level.

Table.4 Considering 12 attributes of the health insurance risk

Metrics	LR	KNN	DT	NB	SVM
Accuracy	0.8357	0.9132	0.9072	0.8768	0.4058
Homogeneity score	0.5568	0.7249	0.6869	0.6241	0.0121
Explained variance	0.6147	0.7850	0.7279	0.7108	-0.0045
Normalized mutual score	0.5601	0.7257	0.6970	0.6249	-0.3553

B. Homogeneity Score

If the level or degree of the homogeneity is higher in all the classes of Risk. The degree of separation will be less and for an algorithm it will be hard to identify separate boundaries of each class type. It can be seen that the score of homogeneity typically remains higher and it impacts that performance of each machine learning algorithm. The value of homogeneity in case of KNN increases when the number of features is reduced. Refer Table III and Table IV.

C. Explained variance

In this model selection criteria Table III is based on explained variance metric. It was found that DT is the best performer (0.7508). After DT it can be observed that NB is at second position with 0.6885 and LR and KNN algorithm is at third with 0.67. The SVM continues to give consistent poor outcome i.e -0.0221.

The result for the top performer changes when the overhead of four features is reduced Table IV. The KNN algorithm becomes the top performer. Its performance increases by almost 11% whereas the performance level of DT decreases by almost 3%.

D.Normalized Mutual score

The result based on the metrics mutual score shows that in almost all cases the value remains in tight range, except SVM algorithm. It can be observed that when DT algorithm selects its feature based on this matric is able to perform best among all other algorithms (0.7249) Table III. It can also be observed that other algorithm performance almost same except SVM.

In case four features are dropped the KNN algorithm performs the best with increase of approximately 12 %. Table IV, whereas the performance of DT decreases by almost 3% but still remains at the second level of performance as compared to other algorithms.

VII. VALIDATION OF RESULTS

For conducting validation, a validation sample set was holdout from the training model. Twenty percent (20%) of the dataset was used as a validation sample. We have not done a single evaluation, as it would limit the ability of the system to characterize the uncertainty of performance of the algorithm. Current literature also shows that proportionally large test sets divide the data into way increase bias in results.

An appropriate size of dataset is good enough because of the small size that model will need every possible data point value to adequately determine the performance value of the metric. Due to these facts, each algorithm was evaluated using 10-fold cross-validation process [53][55]. Because of this method, a high degree of bias can be avoided and variance also remains low.

A. Validation pseudo logic code

```
data = {   }
param ={  }                    initialization of variables
split_ratio= { }
K = {  }              // number of evaluation runs

train, test = split (data, split_ratio)
skill = list ( )

for each i  in  K
  fold_train, fold_var =  cv_split ( i, K, train)
  model = fit (fold_train, param)
  skill_estimate = evaluate (model, fold_var)
  skill append (skill_estimate)
end

# now include the model
for each model
   find (model with lowest std. and highest accuracy)
end
model = fit(train)
skill = evaluate (model, test, skill)
```

From the pseudo logic it can be understand that the algorithm needs to acquire skills by learning from the data. The level of skill acquired needs to be evaluated using different sets (training, testing). Once the learning is completed the model needs to be evaluated for fitness, which is evaluated on the basis of multiple test runs called 10 K fold/validation.
The pseudocode gives information on how the validation of the algorithm is done. The objective of k-fold validation [56] is to support a statistical evidence that the performance of the algorithm under validation is stable or not. The process by splitting the data into training and testing sets ten times, so that the algorithm can evaluate ten times. Then average is computed and a holistic picture is brought. The average is computed using interquartile range (IQR) method [57] so that bias can be avoided in averaging.

B. Learning Phase Analysis

The rate at which an algorithm acquires the ability to classify and predict the class membership of a given instance is called learning rate. This section examines the 'learning curve' of the best performing algorithm. As mentioned in the previous section that KNN has the maximum accuracy and this section examines the learning phase of KNN in detail using graphical representation. Fig. 4(1) and Fig. 4(2) shows that the behavior of KNN and DT during the learning phase using 16 features and the Fig. 5(1) and Fig. 5(2) shows the progress of DT and KNN when the number of features was reduced to 12. From the shape of the curve, it can be seen that the training scale initially starts with a value of 0.82 and then it reaches up to 0.91. Therefore, the learning rate is computed as $(y_2 - y_1) / (x_2 - x_1)$

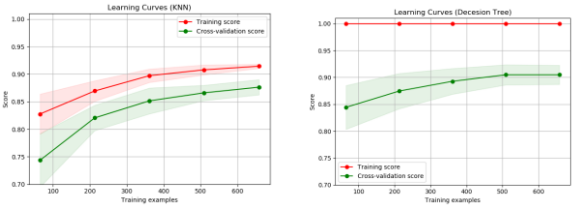

Fig.4 Learning curve using all 16 parameters

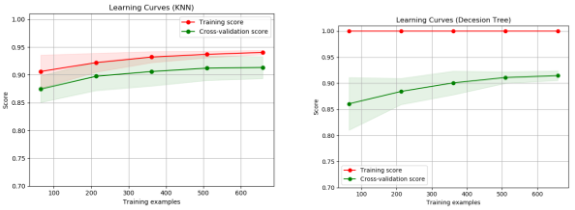

Fig.5 Learning curve using 12 parameters

Change in number of features brings positive change in DT, the cross-validation score increases by 1% and in both cases (16 and 12 features) the training score remains same. But, the improvement in KNN is more as compared to DT. The cross-validation score improves from 0.74 to 0.87 when the training sample are 100 and four features are dropped and it continues to increase till it reaches the value 0.92 with 823 samples. A comparison between DT and KNN revels that the training score of DT always remains close to 1 (constant), but training score of KNN increases above 0.95. There is almost difference of 5%. The second thing that can be observed is that cross validation score is always higher in case of DT with 16 as well as 12 features set. Typically, it is always expected that the training score and validating score should be high of the machine learning models, but in our case, it can be observed training score of DT is always close to 1 (which is higher than KNN) and its validation score is lower (which is less than KNN). It can also be seen that adding more samples will add more generalized result. In case of KNN the training score is low in the beginning and increase later on. Similarly, the validation score is low in the beginning and increases later.

C. Prediction Latency Analysis

It is always desired that the machine learning algorithm should be fast in learning and prediction. High latency of the algorithm can lead to bad user experience. Hence, this section evaluates the top performance with respect to prediction latency per instance in atomic and bulk. The atomic mode validates / tests the instances one by one where as in the bulk mode the algorithm processes in batches to evaluate prediction latency. It can be observed clearly that the lower number of features lead to a reduction in prediction latency (per instance) in both the algorithm, but DT is predicting faster as compared to a KNN. This is apparently due to the fact the DT algorithm builds multi-regression tree models which consume less time in execution Fig. 6.

This also validates the current findings from the literature survey that the main limitation of KNN is that it is a slower algorithm [58][59]. However, in our case it is giving a stable performance in terms of recall, precision, f-score and accuracy.

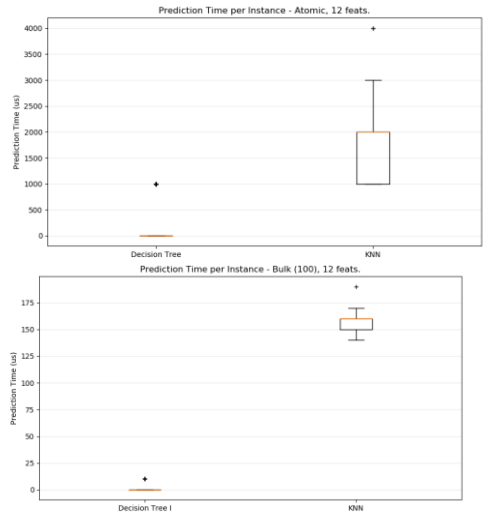

Fig.6 Latency Prediction using all 16 parameters

When both the algorithm is evaluated for prediction of latency based on bulk (100 instances in one go), the result is consistent. It is clear that DT is faster than KNN even when the number of features is 16 Fig. 7.

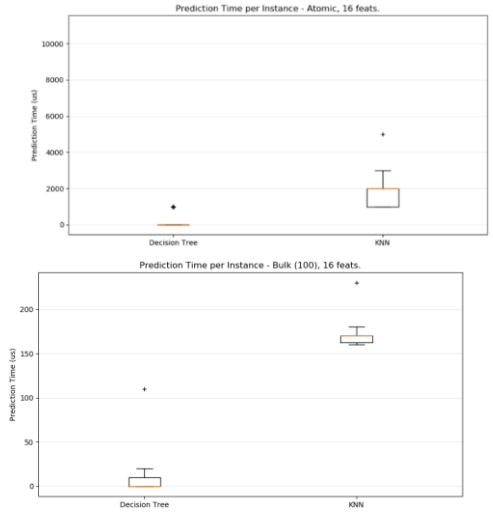

Fig.7 Latency Prediction using 12 parameters

D. Validation of Model quality

There is a fair possibility that a high-performance algorithm may have a good level of accuracy but at the same time its fitness quality is questionable. There are three possible cases. The two cases 'under fit' and 'overfit' are characterized by higher error metrics and 'good fit' is characterized by a low degree of errors. This section gives boxplots of k-fold validation done on the basis of error metric along with a tabular summary of these error metrics.

It can be seen from the Table V that in both the cases the SVM is unable to do data fitting. This is evidence from the value of R^2. The relative positive plot of all the other algorithm changes as the number of features are reduced from 16 to 12 values. The shape of KNN get longer showing that there is more variation in their performance when feature is 12 as compared to the condition of 16 features. The shape of DT and NB gets smaller which shows that their results are coming within tighter range. But their average accuracy is below KNN when considering R^2 metric.

It is always desired that machine learning under evaluation must have lower possible errors while finding feasible solution. The metric mean square error Table V attempts to give inside information about how the algorithm is finding solution. If there is large value of error metrics, it means that algorithm is finding class boundaries separation points. SVM has maximum error and DT has the minimum error value as compared to other algorithms but when four features are dropped Table V the KNN is close to the value of zero, and SVM algorithm has still maximum error value.

It is clear from the tabular summary Table IV and Table V that the other algorithms are unable to acquire low values of error during testing and validation. But KNN algorithm behavior is consistent during the training phase as well in multiple evaluation rounds of cross-validation. It seems that training loss in case of KNN is lowest and the algorithm keeps on improving its accuracy to the maximum level.

KNN's algorithm behavior does not demonstrate an 'overfitting' behavior also because the dataset does not have significant noise / outlier values. This error value is lowest as compared to other algorithms. From this, it can also be inferred that the degree of generalization of the model is optimal.

E. Validation on the basis of Precision and Recall

From the Table VI following, inference can be derived.

1. As the number of features is reduced from 16 to 12, the accuracy, precision, recall and f score of KNN and NB improves. But, KNN is better as compared to NB as its accuracy, recall, precision is higher as compared to KNN.
2. With less number of features and overhead, the KNN and NB algorithm obtains the highest level of performance in term of recall, precision, and accuracy.
3. When the values of recall, precision and f score are computed individually i.e at micro level average. The KNN algorithm has the highest value when 12 features are selected.
4. A macro-level i.e when the values of recall, precision are computed by aggregation on all variables, the KNN algorithm performs well as compared to other algorithms. This holds true and valid when the weighted average of these metrics is computed.

Table.5 Analysis on 16 and 12 features for metric R^2 and negative mean square log error

S. no.	Metrics	16 features					12 features				
		LR	KNN	DT	NB	SVM	LR	KNN	DT	NB	SVM
1	R2	0.6712	0.6630	0.7111	0.6823	-1.2139	0.6045	0.7772	0.7005	0.7031	-1.1823
2	Negative mean square log error	-0.0280	-0.0258	-0.0217	-0.0258	-0.1682	-0.0328	-0.0170	-0.0255	-0.0240	-0.1658

5. There is an average increase of 1.2% in term of recall, precision and f score when the four variable (BirthAge, Height, Resting Metabolism and Skeleton muscle) are eliminated. This also leads to reduced time in the execution of learning/training of algorithm.

F. Validation on the basis of Precision Macro / Micro / weighted

Macro precision is concerned with the overall level of precision of all the classes [60] . The Table VI shows that the performance at macro level. Following observation can be made.

1. SVM remains a poor performer as there is no impact of changing the number of features on it. A negligible increase value can be observed with reduced output feature set.
2. KNN's performance improves by 4% and its value is almost equal to DT's macro precision level.
3. The performance of LR drops by 2% approximately, with 12 features. Its average performance level remains between 83.21.
4. The performance of NB also improves just by 1%.

On basis of this metrics it can be safely conclude that KNN and DT are competing for top slot. There is a need for further investigation to finally declare the best performer.

Precision Micro is used especially when the dataset has unequal number of instances of class in the dataset. /but, in context of the problem undertaken here, the data processing process created a dataset which has equal proposition of classes. Anyway, following observation can be seen.

1. SVM remains a algorithm with poorest record of performance.
2. The performance of LR drops by 2% and NB improves by 1%. The results are similar to the outcome computed on the basis of micro precision.
3. The performance of KNN improves by 5% with 12 features sets. And, clearly is the best performer. DT remains a runner up.

In case of Precision weighted the performance of KNN reaches to level of 0.92. From this value and value calculated on the basis of macro / micro, it is safe to say that KNN is best performer.

G. Validation on the basis of Recall Macro / Micro / weighted

It can be observed that in terms of recall macro both DT and KNN have similar levels. Both algorithms are giving a value of 0.91 when the feature set size is 12, This value is about 5% higher, when the number of features is 16.

Recall Micro computes the value of recall individually for each class and then comes out with a final value. The advantage is that each class is individually evaluated. Following observations can be made.

1. KNN performs better than DT as its value is 1% higher than the DT. Infect, the KNN algorithm is the best performer and its results are stable and reproducible.
2. DT is second best and NB is on the third position.

When the value of recall is computed by giving due weightage to each class, the recall metric is termed as recall weighted metric.

1. It can be observed from Table VI that the reduction in overhead brings better levels of performance of KNN algorithm and no change in the performance level of DT is observe.
2. NB and SVM improve slightly and both of this algorithm do not match the performance of KNN, which is the best algorithm. LLR also lags by 8% when feature set size is 12.

Table.6 Validation of machine learning algorithm on PR

S. no.	Metrics	16 features					12 features				
		LR	KNN	DT	NB	SVM	LR	KNN	DT	NB	SVM
1	Precision macro	0.8528	0.8768	**0.9198**	0.8662	0.2650	0.8336	**0.9151**	0.9133	0.8762	0.3995
2	Precision micro	0.8570	0.8676	**0.9194**	0.8692	0.3982	0.8357	**0.9132**	0.9088	0.8768	0.4058
3	Precision weighted	0.8626	0.8774	**0.9186**	0.8752	0.2975	0.8465	**0.9200**	0.9086	0.8823	0.4404
4	Recall macro	0.8496	0.8641	**0.9101**	0.8671	0.3380	0.8271	**0.9121**	0.9103	0.8738	0.3450
5	Recall micro	0.8570	0.8676	**0.9057**	0.8692	0.3982	0.8357	**0.9132**	0.8996	0.8768	0.4058
6	Recall weighted	0.8557	0.8676	**0.9072**	0.8692	0.3982	0.8357	**0.9132**	0.9042	0.8768	0.4058
7	f1 macro	0.8471	0.8659	**0.9104**	0.8636	0.1993	0.8220	**0.9107**	0.9047	0.8722	0.2118
8	f1 micro	0.8570	0.8676	**0.9043**	0.8692	0.3982	0.8357	**0.9132**	0.9042	0.8768	0.4058
9	f1 weighted	0.8558	0.8683	**0.9063**	0.8692	0.2359	0.8328	**0.9138**	0.9122	0.8769	0.2493

H. Validation on the basis of F1- macro / micro / weighted

F1-score helps to identify the problem related to bias as it tries to find the averages. It can be observed that the result of F1-micro, F1-macro and F1-weighted are also most similar in nature.

The outcome is consistent with recall and precision values. It is obediently clear that KNN is scoring higher on all parameters followed by DT algorithm. The standard deviation in the result of KNN's metric values is moderate and consistent.

VIII. CONCLUSION

In this research work, data was collected for developing a framework that can identify and classify health insurance risk factor. The important analysis of all features was done using random forest algorithm and it was found that twelve risk factors are statistically relevance. A medical relevancy analysis was also done and it was found that Birth Age, Height, Body Fat, Skeleton Muscle, and Resting Metabolism factors are not important for computing insurance risk. For monitoring data quality missing value, outlier analysis and formatting of the dataset was done, so that the data can be used for making machine learning algorithm beside this suitability analysis of the dataset was done to check the nature of the dataset for machine learning. And it was found that the dataset characteristics can be used for making machine learning as it has good prediction power and a reasonable degree of separation and modular degree of overlapping.

Ground-rule based on medical limits for binning the data into the multi-class problem was done and an equal proportion of each class instances was selected for developing machine learning models. Extensive design of experiment was planned and executed and it was found that the KNN algorithm is best for the problem undertaken.

Validation of work was done using 10 K-fold cross-validation process using twelve attributes and it was found that DT is a close competitor of KNN algorithm. It was found that KNN is slower than DT in term of prediction latency. But it is highly accurate, stable and consistent with its performance.

FUTURE SCOPE

Latest research in medical science shows that human body has up to hundred trillion bacteria. The bacteria help in human cells in digestion, protection and many other functions. It has been found that without these bacteria, the human body cannot function properly. A good diversity of 'micro biodata' is essential. Using this idea and doing analysis of microbiota of each person, risk of health and insurance can be quantified. Hence for future scope it is suggested to work in this direction.

REFERENCES

1. D. Lakdawalla, A. Malani, and J. Reif, "The insurance value of medical innovation," *Journal of Public Economics*, vol. 145, pp. 94–102, 2017.
2. L. Manchikanti, S. Helm Ii, R. M. Benyamin, and J. A. Hirsch, "Evolution of US Health Care Reform," no. 1, pp. 107–110, 2017.
3. P. Singh and V. Kumar, "Insurance coverage under different health schemes in Uttar Pradesh, India," *Clinical Epidemiology and Global Health*, vol. 5, no. 1, pp. 33–39, 2017.
4. H. Nguyen and L. B. Connelly, "Cost-sharing in health insurance and its impact in a developing country-Evidence from a quasi-natural experiment," *Munich Personal RePEc Archive*, no. 44017, p. 76399, 2017.
5. M. K. Bundorf, J. Levin, and N. Mahoney, "Pricing and welfare in health plan choice," *American Economic Review*, vol. 102, no. 7, pp. 3214–3248, 2012.
6. S. Dey, "Impact of Affordable Care Act (ACA) on Health Informatics," *Proceedings - 2014 Annual Global Online Conference on Information and Computer Technology, GOCICT 2014*, pp. 36–41, 2014.
7. D. Mavalankar, "Health Insurance in India Opportunities , Challenges and Concerns," no. November, 2000.
8. A. Karan, W. Yip, and A. Mahal, "Social Science & Medicine Extending health insurance to the poor in India : An impact evaluation of Rashtriya Swasthya Bima Yojana on out of pocket spending for healthcare," *Social Science & Medicine*, vol. 181, pp. 83–92, 2017.
9. P. Weekly and P. Weekly, "Health Insurance in and India Prognosis," vol. 35, no. 4, pp. 207–217, 2012.
10. N. Devadasan, B. Criel, W. Van Damme, K. Ranson, and P. Van Der Stuyft, "protection against catastrophic health expenditure," vol. 11, pp. 1–11, 2007.
11. B. Ramesh and J. Nishant, "Factoring Affecting the Demand for Health Insurance in a Micro Insurance Scheme."
12. R. J. Koene, A. E. Prizment, A. Blaes, and S. H. Konety, "Contemporary Reviews in Cardiovascular Medicine Shared Risk Factors in Cardiovascular Disease and Cancer," no. Cvd, pp. 1104–1115, 2016.
13. M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," 2011.
14. D. S. Gangwar, "Biomedical Sensor Network for Cardiovascular Fitness and Activity Monitoring," pp. 16–18, 2013.
15. M. N. K. Boulos *et al.*, "CAALYX : a new generation of location-based services in healthcare," vol. 6, pp. 1–6, 2007.
16. T. Hayajneh *et al.*, "Secure Authentication for Remote Patient Monitoring with Wireless Medical Sensor Networks," *Sensors*, vol. 16, no. 4, p. 424, Mar. 2016.
17. M. S. Hossain and G. Muhammad, "Cloud-assisted Industrial Internet of Things (IIoT) – Enabled framework for health monitoring," *Computer Networks*, vol. 101, pp. 192–202, Jun. 2016.
18. H. Engineering, L. G. Editor, Z. Huang, G. Editors, J. M. Juarez, and X. Li, "Data Mining for Biomedicine and Healthcare."
19. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities," vol. 3536, no. c, pp. 1–10, 2017.
20. A. S. Abdullah, N. Gayathri, and S. Selvakumar, "Identification of the Risk Factors of Type II Diabetic Data Based Support Vector Machine Classifiers upon Varied Kernel Functions," *Springer*, vol. 1, pp. 496–505, 2018.
21. K. W. Degregory *et al.*, "Obesity / Data Analysis A review of machine learning in obesity," pp. 1–18, 2018.
22. Y. H. Lee *et al.*, "A cross-sectional evaluation of meditation experience on electroencephalography data by artificial neural network and support vector machine classifiers," *Medicine (United States)*, vol. 96, no. 16, 2017.
23. J. Glowacki, "Effective model validation using machine learning Model validation has been a key focus that rely on models for underwriting , learning techniques," no. May, 2017.
24. O. Article, "Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes – ELSA-Brasil : accuracy study Comparação de algoritmos de aprendizagem de máquina para construir um modelo preditivo para detecção de diabetes não di," vol. 135, no. 3, pp. 234–246, 2017.
25. F. Pedregosa, R. Weiss, and M. Brucher, "Scikit-learn : Machine Learning in Python," vol. 12, pp. 2825–2830, 2011.
26. B. Bischl, M. Lang, and Z. M. Jones, "mlr : Machine Learning in R," vol. 17, pp. 1–5, 2016.
27. D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare," vol. 5, no. 5, pp. 241–266, 2013.
28. N. Boodhun and M. Jayabalan, "Risk prediction in life insurance industry using supervised learning algorithms," *Complex & Intelligent Systems*, vol. 4, no. 2, pp. 145–154, 2018.
29. R. D. Riley *et al.*, "External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis : opportunities and challenges," pp. 27–30.
30. 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 & Mental Health*, vol. 22, no. 3, pp. 371–378, Mar. 2018.
31. B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine Learning for Medical," *Radiographics*, no. 1, pp. 1–11, 2017.

32. H. C. Ho, M. S. Wong, H. Y. Man, Y. Shi, and S. Abbas, "Neighborhood-based subjective environmental vulnerability index for community health assessment: Development, validation and evaluation," *Science of the Total Environment*, vol. 654, pp. 1082–1090, 2019.
33. D. Ellen, S. Day, C. Davies, S. Day, and C. Davies, *Statistical and Machine-Learning Data Mining*. CRC Press, 2011.
34. C. A. Ramezan, T. A. Warner, and A. E. Maxwell, "Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification," 2019.
35. 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.
36. Z. Bosnic, "ROC analysis of classifiers in machine learning : A survey ROC Analysis of Classifiers in Machine Learning : A Survey Technical report MM-1 / 2011," no. May 2013, 2015.
37. S. B. Baker, W. Xiang, and I. Atkinson, "Internet of Things for Smart Healthcare: Technologies, Challenges, and Opportunities," *IEEE Access*, vol. 5, no. c, pp. 26521–26544, 2017.
38. M. Sordo and M. G. Hospital, "Introduction to Neural Networks in Healthcare," no. June, 2015.
39. V. Raghupathi and W. Raghupathi, "Preventive Healthcare : A Neural Network Analysis of Behavioral Habits and Chronic Diseases," pp. 1–13, 2017.
40. A. Singh and K. . Ramkumar, "Health Insurance Risk Dataset." Amrik Singh, Jammu, 2018.
41. W. H. O. Bmi and T. W. H. O. Bmi, "Public health Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies," vol. 363, pp. 157–163, 2004.
42. K. T. Mills, H. He, J. Chen, P. K. Whelton, and J. He, "Systolic Blood Pressure Reduction and Risk of Cardiovascular Disease and Mortality A Systematic Review and Network Meta-analysis," vol. 70118, pp. 1–7, 2017.
43. "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].
44. H. Hsu *et al.*, "10.1017/s000711451900028x."
45. M. Komi, J. Li, Y. Zhai, and Z. Xianguo, "Application of data mining methods in diabetes prediction," *2017 2nd International Conference on Image, Vision and Computing, ICIVC 2017*, no. S Ix, pp. 1006–1010, 2017.
46. M. Tukaram, L. Col, P. Dudeja, and L. Ipsita, "ScienceDirect Correlation of visceral body fat with waist – hip ratio , waist circumference and body mass index in healthy adults : A cross sectional study," *Medical Journal Armed Forces India*, pp. 1–6, 2018.
47. F. Nargesian, H. Samulowitz, U. Khurana, E. B. Khalil, and D. Turaga, "Nargesian 等。 - 2017 - Learning Feature Engineering for Classification.pdf," pp. 2529–2535.
48. J. Heaton, "An empirical analysis of feature engineering for predictive modeling," *Conference Proceedings - IEEE SOUTHEASTCON*, vol. 2016–July, pp. 0–5, 2016.
49. A. Suresh, R. Kumar, and R. Varatharajan, "Health care data analysis using evolutionary algorithm," *Journal of Supercomputing*, pp. 1–10, 2018.
50. G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.
51. "Feature importances with forests of trees — scikit-learn 0.21.2 documentation." [Online]. Available: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html#sphx-glr-auto-examples-ensemble-plot-forest-importances-py. [Accessed: 12-Jul-2019].
52. C. Hsu and T. Ohio, "Delphi Techniques Making Sense of Consensus," *A peer-reviewed eloctronic journal Practical Assessment*, vol. 12, no. 10, p. Volume 12, ISSN 1531-7714, 2007.
53. S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," vol. 4, pp. 40–79, 2009.
54. M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using Joint Mutual Information Maximisation," *Expert Systems With Applications*, 2015.
55. Y. Zhang and Y. Yang, "Cross-validation for selecting a model selection procedure," *Journal of Econometrics*, vol. 187, no. 1, pp. 95–112, 2015.
56. A. Pe and J. A. Lozano, "Sensitivity Analysis of k -Fold Cross Validation in Prediction Error Estimation," vol. 32, no. 3, pp. 569–575, 2010.
57. University of Leicester, "Measures of variability: the range, inter-quartile range and standard deviation," pp. 1–7, 2009.
58. A. Lamba and D. Kumar, "Survey on KNN and Its Variants," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5, no. 5, 2016.
59. G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN Model-Based Approach in Classification," no. August, pp. 986–996, 2010.
60. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, pp. 1–21, 2015.