JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-14, No.-4, July-August (2019) pp 27-40 ISSN (Print) 0973-8975

An Efficient Statistical Feature Selection Based Classification

K. Laxmi Narayanamma[1], R. V. Krishnaiah[2], P. Sammulal[3]

[1]Assistant Professor, Dept. of Information Technology, Institute of Aeronautical Engineering, Hyderabad. E-mail:laxmi.k1207@gmail.com

[2]Professor, Dept. of Electronics and communication Engineering, Institute of Aeronautical Engineering, Hyderabad. E-mail :r.v.krishnaiah@gmail.com

[3]Associate Professor, Dept. of computer science and Engineering, Jntuh College of Engineering, Hyderabad. E-mail: sam@jntuh.ac.in

https://doi.org/10.26782/jmcms.2019.08.00003

Abstract

Initial identification about pancreatic cancer (PC) will be a very challenging task due to particular symptoms of cancer happens only at an advanced phase & a dependable screening device to detect high danger patients. To know this challenge, a new method for decreasing the features might have been developed, tested & trained with the use of the health information of 800,114 defendants caught in the "national health interview survey (NHIS)"& Pancreatic, Colorectal, Lung, & "PLCO (ovarian cancer)" datasets, together risk of cancer might have been evaluated at a distinct level by including 18 characteristics under the recommended. The recognized "hybrid feature selection method" attained a true positive rate of 87.3 & 80.7% a true negative rate 0.86 & 0.85 for the training and testing associates, individually.

Keywords : American Cancer Society (ACS), Machine learning (ML), Feature selection (FS), Feature extraction (FE), pancreatic cancer (PC).

I. Introduction

With the expanding capability about health awareness centers to digitally establish the information of clinical cancer patient, novel strategies exist to investigate the cancer treatment. This information is advantageous in demonstrating complete trends in treatment of patient but it has only until lately applied to assessing how to provide good treatment for distinctive patients. Methods based on level information of patient with the use of data mining strategies have the possibility to provide for knowledge under the long term outlook &good treatment for every patient.

The aim of this manuscript is to apply the data mining strategies to information of cancer patient. The survey under modeling of associations among a history of patient, disease phase, medical file, & result will be central to survey about the best

treatment for cancer patients. The data mining strategies [Hay06]have the possibility to enhance the current methods of these associations.

For this purpose, the concentration on "pancreatic cancer (PC)". According to "American Cancer Society's 2007 statistics", PC will be fourth highest dangerous among all types of cancers in US [ACS07]. For the diagnosed with PC, there will be a 1 year survival rate of 19% & 5 year survival rate of 4%. Whether clinical resection of cancer is implemented, the 5 years survival rate improves up to 40% [DHR05]. These percentages give a considerable argument for clinical deduction of cancers from whole patients but this will be not suitable frequently.

The PC is very hard to treat & very in capacitating to patient. For instance the process of Whipple is the general surgical treatment for PC, and it might take around 8 hours to complete & might take few months to get well. Whether the tumor has spread under the patient's veins or different organs of body, much this powerful surgery will be not capable to cure the infection. In instances the place it may be not probable to eliminate the tumor, it will be more suitable to take stages to enhance the quality life of patients though existing with the infection.

If clinical treatment of PC will be applicable, a trade off should continuously made deliberating the survival time & expected quality life of patients.

PC is the 4th prominent reason of cancer death in human in US [I] despite its low incidence rate. In 2017, a "total of 43,093 associated deaths (7.17% of all cancer deaths) &a total of 53,670 new PC cases (3.18% of all new cancer cases)" were noted in US [I]. The death rate of age related cancer will be incrementing for PC &it will be anticipated that PC will become the 2nd general reason of cancer deaths by 2030 [VII]. PC has a high rate of humanity in part due to symptoms of cancer in many patients (>80%) happen only at a progressive phase [VIII].

As stated by the 2017 ACS statistics [I], the past 5-years survival rate for all phases of PC will be 8.5%. The survival rates for patients with early-stage analysis might a chance to be high as 20%, [VIII]. Nevertheless, a minor part of patients (<15%) have surgically illness at the diagnosis time [X]. Moreover, people identification with initial phase disease or at high problematic for PC will be challenging because of the absence of dependable screening devices, the absence of specific & sensitive biomarkers, and the low pervasiveness [VII] [VIII].

Lately, various investigations are concentrated on initial identification of PC over the validation & identification of guaranteeing. Moreover, the capacity to recognize precancerous progressions among high risk people via "magnetic resonance imaging (MRI)", Doppler ultrasound (DUS), "computed tomography (CT) scan", endoscopic ultrasound (EUS), or "positron emission tomography (PET)"are exhibited in few clinical investigations [XI] [XII].

Pancreatic cancers are less as 0.5 cm might be recognized with "diagnostic imaging", like MRI, CT, or EUS. Nevertheless, in spite of high sensitivity of these strategies [7] it will be not useful or monetarily attainable to execute common PC

screening in overall people because of the moderately low occurrence rate [9] though, these strategies might be utilized more proficiently and cost-effectively, whether working in "high risk subset" of people. For instance, screening conventions are connected to patients with germ line mutations connected with PC & patients with familial PC [X]. Nevertheless, main 10–20% of whole PC situations might be attributed with familial PC.

Different clinical &epidemiologic aspects are connected with PC occurrence, comprising PC family history anthropometric variables [such as BMI (body mass index)] [2] [3] inherited genetic influence/variation therapeutic comorbidities (such as diabetes) [5], [7], and lifestyle (such as drinking alcohol, smoking) [New onset diabetes will be acknowledged as the toughest predictors of PC, & various epidemiologic investigations described that the connection among recently diabetes mellitus & diagnosed PC might have been ∼50% [11] [7].

The work [13] demonstrated that the 3-years increasing prevalence of PC among patients with novel diabetes may be 8 times greater than required. Therefore, it might be expressed that diabetes connected with PC might be "Para neo plastic phenomenon" initiated by tumor [7] Smoking is also increments the danger of PC by 2 factors. Indeed the utilization of smokeless tobacco increments the risk of PC [1]. The PC family history will be likewise recognized a hazard component.

I.i Data Sources

The "national health interview survey (NHIS)" [6] might have been recognized in 1957 to display the whole health status of US through individual interviews on a expansive extend of health topics. Various epidemiologic investigations are showed utilizing NHIS [16]. In 1997 to 2017, the NHIS datasets are utilized within this survey. The population of target survey comprised about people with onset for PC less than 4 years previous to review date. Acknowledging the input features time reliance to the method, this 4-years cut off on the PC set might have been chose then cautious testing of distinctive cut offs on execution of method to strike stability among predictive power & sample size of our strategy. Then applying this cut off, we have 645,217 respondents, 131 for whom had PC.

The "Prostate, Lung, Colorectal, and ovarian (PLCO) trial may be a randomized, regulated trial exploring if specific screening exams diminish mortal from PLCO tumor. The 154,897 participants are enrolled between November 1993 and July 2001, the 767 participants are established PC throughout 13 years. To this survey, history of family, personal information of health, PC status, lifestyle, socio behavior, & dietary information are removed from PLCO datasets through an "in-house Mat lab code".

I.ii Sample Size Considerations

Whole information in NHIS dataset from 1997 to 2017 &PLCO dataset were utilized to increase the force and generalize ability of the outcomes. To explore the ANN execution on distinctive datasets, 3 datasets were built:
1. DS1 = NHIS dataset (645,217 participants, with 131 PC cases).
2. DS2 = PLCO dataset (154,897 participants with 767 PC cases).
3. DS3 = NHIS dataset + PLCO dataset (800,114 participants with 898 PC cases).
After randomizing & constructing these 3 datasets, we utilized a test/train/validate strategy. The ANN might have been trained on 70% (training dataset) of the information utilizing "10-fold cross-validation", same time the residual 30% might have been withheld for further testing (testing dataset). The sensitivity, risk of cancer, and specificity were estimated for testing & training datasets.

II. Previous Works

An extensive diversity of data mining methods present to training data method also known as training examples, which might be utilized to anticipate the value of target of an unseen example. Particularly, these systems build methods of connections among a group of I/P attributes, called as features, & concept of target. Every attributes signifies an I/P space dimension, which will be utilized to build a target concept method. The FS strategies exist to assistance diminish the amount of extents in this I/P space.

$$\equiv Entropy(I) - \sum_{vCvaluse(A)} \frac{|I_v|}{|I|} entropy(I_v)$$

There are many strategies for FS consisting ones, which choose better group of unique attributes & ones that change the I/P space to acquire a better group of features to signify the information. The relationship methods among the target concept & I/P space are constructed through utilization of ML algorithms.

II.i Feature Selection

The methods of feature selection are usually alluded to as algorithms of attribute selection to decrease the "input space dimensionality". Typically, this is completed by examining for much pertinent group of attributes. Diminishing the dimensionality of input space is generally increments predictive accuracy & efficiency of ML algorithms.

II.ii. Gain Ratio Attribute Selection

$$Entropy(TrainingIns\tan ceI) \equiv \sum_{i=1}^{numberofl\arg evalues} -p_i \log_2 p_i$$

Where:

P_1 be the classification target i probability

Formula 1: Entropy

Gain (training instances I some Attribute A)

$$\equiv Entropy(I) - \sum_{vCvaluse(A)} \frac{|I_v|}{|I|} entropy(I_v)$$

The equation for estimating the gain ratio (GR) is suggested in [Qui86] as a strategy for assessing the attributes in DT construction. A difficult factor in estimating the GR will be entropy. The entropy will be estimated in equation 1 by combining over the probability's negative product of every value of classification times its logarithm. The every classification value probability will be estimated rely on its frequency in group of training examples. Whether all classes have an similar probability, the entropy formula is return a maximum value than whether a minor subset of probable values have high probabilities than remaining. The entropy will be dimensions of degree to those classes are distinguished within the training examples. This entropy is utilized to estimate the information gain (IG) for attribute as represented in equation 2. The path of entropy is calculates the degree to those classes are distinguished within training examples might be applied to distinct attribute with the use of split data. The split data is described in equation 3 & it is estimated the similar as entropy just with attribute probability rather than the classification target probability. The split information &IG are utilized to estimate the GR for attribute will be represented I equation 4. The GR will be utilized for FS by running for each attribute.

II.iii. Principal Components Analysis

$$C^{NXM} = (\forall (c_{i,j}) 1 \square\ i \square\ n \Lambda 1 \square\ j \square\ n, c_{i.} = \text{covariance}() Attribute_{i.} Attribute_{j}))$$

Where n is the of Arrtibutes

$Arrtibutes_x$ is the x^{th} Arrtibutes

Formula5: covariance Matrix [smi02]

The significant module examination changes the I/P space from n attribute I/P space to 1 signified by the patterns among the attributes. This change needs constructing an "n*n covariance matrix" will be characterized in (Eq 5). The Eigen values & eigenvectors are estimated for the covariance grid. The eigenvectors are positioned by their Eigen values whereas the maximum Eigen values are choose 1st note, which for I/P space with n attributes, there is a n*n covariance matrix that will generate n eigenvectors. Evacuating the eigenvectors hence diminishes the amount of measurements in I/P space same time still providing each attribute some impact through the last classification.

II.iv Relief Attribute Selection

Set all weights $W[A] = 0, 0;$

For i=1 to Number of instances do:
R=instance number i
H=Find k neighbors with same class value as R
M=Find k nearest neighbors with different class value R
For a=1 to Number _of _ Attributes do:
W[A]=W[A]+(difference(A,R,M)-(difference(A,R,M))/Number_of_instancecs
End;
Difference (A,R,M)
Number different=0;
For j=1 to k do;
if value(A,X[j])!=value(A,R):
number different+=absolute value(value(A,R)-value(A,X[j])/Max value (A,R),value(A,X[j])-Min value(A,R),value(A,M[j])
Return number different/k;
End;

The selection of ReliefF attribute allocates a weight to every attribute rely on how better; which attribute will be capable to segregate among close-by examples. The ReliefF algorithm, which is showed in Figure 1might have been initially introduced over [Kon94]. For each example, the k closest neighbors of diverse class values &k closest neighbors with similar class value are discovered. These 2 sets of closest neighbors are utilized to estimate the level to every attribute distinguishes close-by illustrations. The weight for every attribute will be adjusted by distinction among the instance score value & neighbor attribute value. The weight will be expanded to neighbors having a separate class &diminished for the neighbors that have the similar class. When the weights are estimated, the attributes are positioned from highest to lowest weights.

II.v. Support Vector Machine (SVM) Attribute Selection

This attribute evaluator employment the SVM strategy is examined in ML algorithms with the "linear kernel function" to assess the significance for every attribute. The attributes with a higher impact over last arrangement are allocated weights by SVM methods, which are away from 0. These weight squares are utilized to rank the attributes.

II.vi. Machine Learning (ML) Algorithms

ML algorithms utilize the training information to build the relationship methods among a target attribute &group of input attributes. These models frequently are

utilized for either classification or regression. The regression includes mapping of I/P attributes to numeric quality whereas classification will be the mapping of the I/P attributes to a minor quality. The strategies planned for regression might be changed to execute the classifications.

II.vii. Zero R

Zero R will be the easy of classifiers &could be considered of as the default classifier. It generally envisages the greater part target value. Whether numerous target qualities tie for greater value, it subjectively selects one to anticipate.

II.viii. Logistic Regression

The logistic regression (LR) will be regularly utilized in medical group to connections model among group of attributes. Its predictive force will be trusted by medical group thus other modeling strategy utilized in this perspective must be compared against "logistic regression "to estimate if it either increments predictive accuracy or by builds a much informative method. Fundamental LR expands a method dependent upon a group of attributes to anticipate the binary classification target probability. An arrangement of regression stages creates a method whereas the inputs generate a representation of probability. For many logistics regressions, much difficult classification targets might be executed & the resultant probabilities consolidated together to envisage the similar class.

II.ix. Bayesian Approaches

$$P(T\arg et / inputVector) = \frac{p(inputVector / T\arg et)\,p(T\arg et)}{p(inputVector)}$$

Formula 6: Bayes'Theorem

Bayesian methods are based on Bayes' rule, detailed in Formula 6, which provides a way to calculate the probability of a classification target given the input vector from the probability of the classification target, the probability of the input vector, and the probability of the input vector given the classification target. Bayesian networks use directed acyclic, graphs to model the dependencies among variables. Variables are represented by nodes, each of which contains a table of probabilities. For every node, the table of probabilities provides the conditional probabilities of each of the variable's values, given each possible combination of the values of the variable's parents in the network. One type of Bayesian Network is Naïve Bayes. Naïve Bayes assumes conditional independence among all variables given the classification target. Although this approach loses conditional dependencies between variables, in practice it can be powerful for classification.

II.x. Decision Trees (DT)

The DT model is a consecutive group of decisions, which finally outcome in classification. Every decision signifies an attribute, every probable value of that attribute important to either classification or another decision. DT Development begins by choosing the attribute, which will be better capable to expand the separation of the classes by grouping the attributes into all probable values. Every level of the tree will be consecutively constructed in this way dependent upon the subset of training qualities with values of previous layers. There are few measurements for degree to that an attribute expands the separation of classes comprising GR & IG which were deliberated in "gain ratio attribute selection section".

III. Implementation

The core idea behind deep learning is that comprehensive feature representations can be efficiently learned with the deep architectures which are composed of stacked layers of trainable non-linear operations. However, because of the diversity of dataset content, it is hard to learn effective feature representations directly from images for cancer detection. Cancer detection is generally framed as an issue of binary classification. This method is named as "universal/blind detection", develops the main stream amongst most present algorithms. In the training phase, effective features extraction is extracted to highlight probable manipulation by detection.

In figure 1, a novel method is proposed to reduce the feature in feature extraction, selection, & training the classifier is implemented at the same time utilizing the "logistic regression (LR) method". The method optimizes a "feature weight vector" utilized to scale the distinctive features in "unique pattern vectors". A masker vector will be also utilized to concurrent determination of a feature subset. We utilize this method over consolidation with RESNET, & compare the outcomes with classical feature extraction & selection systems.

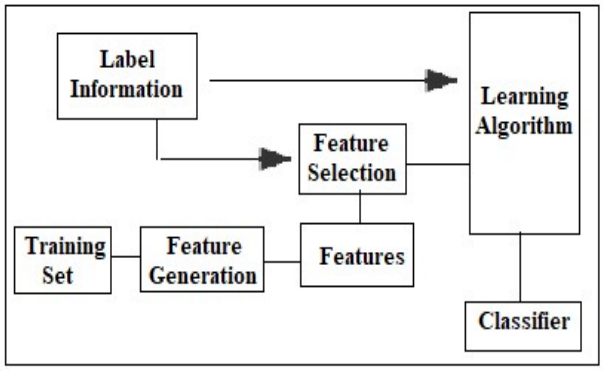


Fig.1 Proposed Designing

III.i. Logistic Regression

Recommended to consolidate learner decisions that combined generates a class label or single regression value. The knowledge of HNN is expanded to FS. Methodologies along the individual lines incorporate the feature relevance definition from "random forests" or combination of different feature rankings gotten from a "SVM-based classifier". The methodologies depend on different re samplings of learning example. Therefore, the differing of ensemble is gotten by acknowledging different subsets of training examples. We pick here to other method of generating diversity, in particular by inspecting the space of feature as stated to a distribution of probability that will be iteratively distinguished to superior method the pertinence of every feature.

III.ii. Regular LR

Logistic Regression Let x ∈Rp denotes an observation made of p feature values and let y ∈ {−1, +1} denote the corresponding binary output or class label. A logistic regression models the conditional probability distribution of the class label y, given a feature vector x as follows.

$$\Pr ob(y / x) = \frac{1}{1 + \exp(-y(w^T x + v))'} \tag{4}$$

Where the weight vector w ∈Rp and intercept v ∈ R are the parameters of the logistic regression model. The equation wTx+v = 0 defines an hyper plane in feature space, which is the decision boundary on which the conditional probability of each possible output value is equal to ½.

We consider a supervised learning task where we have n i.i.d. training instances {(xi, yi), i = 1,...,n}. The likelihood function associated with the learning sample is n i=1Prob (yi|xi), and the negative of the log-likelihood function divided by n, sometimes called the average logistic loss, is given by

$$l_{avg}(w,v) = \frac{1}{n}\sum f(y_i(w^T x_i + v)) \tag{2}$$

Where f (z) = log (1 + exp(−z)) is the logistic loss function. A maximum likelihood estimation of the model parameters w and v would be obtained by minimizing (2) with respect to the variables w ∈Rp and v ∈ R. This minimization is called the logistic regression (LR) problem. When the number n of observations is small compared to the number p of features, a logistic regression model tends to over-fit the learning sample. When over-fitting occurs many features have large absolute weight values, and small changes of those values have a significant impact on the predicted output.

The most common way to reduce over-fitting is to add a penalty term to the loss function in order to prevent large weights. Such a penalization, also known as regularization, gives rise to the l2-regularized LR problem:

$$\min_{w,v} l_{avg}(w,v) + \lambda \square w \square_2^2 = \min_{w,v} \frac{1}{n} \sum f(y_i(w^T x_i + v)) + \lambda \sum_{j=1}^{p} w_j^2$$

Here $\lambda > 0$ is a regularization parameter which controls the trade-off between the loss function minimization and the size of the weight vector, measured by its l2-norm. As discussed in [15], the l2-regularized LR worst case sample complexity grows at least linearly in the number of (possibly irrelevant) features. This result means that, to get good predictive performance, adding a feature to the model requires the inclusion of an additional learning example. For small n, large p problems the l1-regularized LR is thus usually considered instead by replacing the l2-norm $\cdot 2$ 2 in (3) by the l1-norm $\cdot 1$. This is a natural extension to the LASSO [20] for binary classification problems. The benefit of the l1-regularized LR is its logarithmic rather than linear sample complexity. It also produces sparse models, for which most weights are equal to 0, hence performing an implicit feature selection. However l1-regularized LR is sometimes too sparse and tends to produce a highly unstable feature selection. A trade-off is to consider a mixed regularization relying on the Elastic Net penalty.

$$\min_{w,v} l_{avg}(w,v) + \lambda \sum_{j=1}^{p} \left[\frac{1}{2}(1-\alpha) w_j^2 + a \left| w_j \right| \right] \tag{4}$$

Where $\alpha \in [0, 1]$ is a meta-parameter controlling the influence of each norm. For high-dimensional datasets the key control parameter is usually still the l1 penalty, with the l2 norm offering an additional smoothing. We argue in this paper that there is an alternative way of obtaining sparse and stable logistic regression models. Rather than relying on a regularization including an l1 penalty, the sparsely is obtained by constraining the model to be built on a number of features of the same order as the number of available samples. This constraint is implemented by sampling feature subsets of a prescribed size. The key ingredient of such an approach, further detailed in section 2.2, is a non-uniform sampling probability of each feature where such a probability is proportional to the estimated feature relevance.

The proposed feature selection is essentially an embedded method relying on regularized logistic regression models. Those models are built on small subsets of the full feature space by sampling at random this space. The sampling probability is directly proportional to the estimated feature relevance. The initial relevance of each feature is estimated according to a t-test ranking. Such a simple univariate ranking does not consider the dependence between features but is observed to be stable with respect to variations of the learning sample. This initial relevance index is iteratively refined as a function of the predictive performance of regularized logistic regression

models built on re sampled features. This procedure iterates until convergence of the classifier performance.

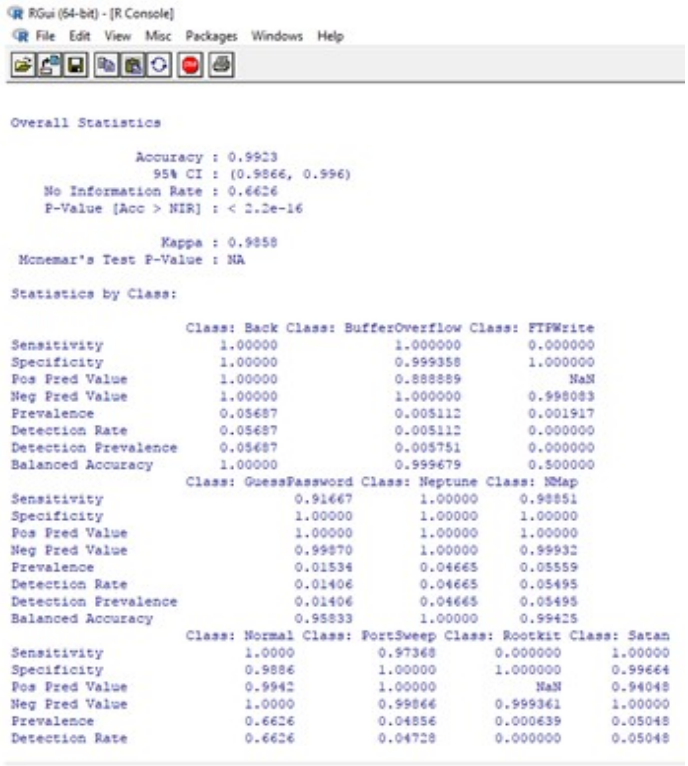
Our method relies on the l2-regularized LR as estimated by the optimization problem (3). The sparsity is not enforced here with an l1 penalty but rather by explicitly limiting the number of features on which such a model is estimated. The sample complexity result from gives us a reasonable default number of features to be equal to the number of training examples n. Those n features could be drawn uniformly from the full set of p features (with p >> n) but we will show the benefits of using a non-uniform sampling probability. We propose here to relate the sampling probability of a given feature to its estimated relevance. Since our primary application of interest is the classification of microarray data, a t-test relevance index looks to be a reasonable choice as a first guess.

This method ranks features by their normalized difference between mean expression values across classes:

$$t_j = \frac{\mu_{j+} - \mu_{j-}}{\sqrt{\sigma_{j+}^2 / m_+ + \sigma_{j-}^2 / m_-}} \tag{5}$$

Where μj+ (respectively μj−) is the mean expression value of the feature j for the m+ positively (respectively m− negatively) labeled examples, and σj+, σj− are the associated standard deviations. The score vector t over the p features is normalized to produce a valid probability distribution vector prob. We note that there is no need here to correct for multiple testing since the t-test is not used to directly select features but to define an initial feature sampling probability. At each iteration the learning sample is split into training (80%) and validation (20%) sets. Next, a subset of n features is drawn according to prob and a l2-regularized LR model is estimated on the training data restricted to those features.

IV. RESULTS



RGui (64-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
Overall Statistics

               Accuracy : 0.9923
                 95% CI : (0.9866, 0.996)
    No Information Rate : 0.6626
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9858
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Back Class: BufferOverflow Class: FTPWrite
Sensitivity              1.00000              1.000000         0.000000
Specificity              1.00000              0.999358         1.000000
Pos Pred Value           1.00000              0.888889              NaN
Neg Pred Value           1.00000              1.000000         0.998083
Prevalence               0.05687              0.005112         0.001917
Detection Rate           0.05687              0.005112         0.000000
Detection Prevalence     0.05687              0.005751         0.000000
Balanced Accuracy        1.00000              0.999679         0.500000
                     Class: GuessPassword Class: Neptune Class: NMap
Sensitivity                       0.91667         1.00000      0.98851
Specificity                       1.00000         1.00000      1.00000
Pos Pred Value                    1.00000         1.00000      1.00000
Neg Pred Value                    0.99870         1.00000      0.99932
Prevalence                        0.01534         0.04665      0.05559
Detection Rate                    0.01406         0.04665      0.05495
Detection Prevalence              0.01406         0.04665      0.05495
Balanced Accuracy                 0.95833         1.00000      0.99425
                     Class: Normal Class: PortSweep Class: Rootkit Class: Satan
Sensitivity                 1.0000          0.97368       0.000000      1.00000
Specificity                 0.9886          1.00000       1.000000      0.99664
Pos Pred Value              0.9942          1.00000            NaN      0.94048
Neg Pred Value              1.0000          0.99866       0.999361      1.00000
Prevalence                  0.6626          0.04856       0.000639      0.05048
Detection Rate              0.6626          0.04728       0.000000      0.05048
```

Figure.2 Overall statics

As shown the figure 2 comparison of Confusion matrix and statistics

V. Conclusion

We described a feature selection, which might be utilized to calculate the PC with 81.7% of specificity, 81.7% of sensitivity, and 0.87 of accuracy built singularly around personal health information. Furthermore, the efficient method might have been capable to stratify human into high, medium, and low cancer risk for much risk management & tailored screening. Contrasted with "current screening methods", this recommended may be cost-effective, non-invasive, & simple to execute with promptly accessible individual health information. More information & testing might be required to more develop the execution of the recommended to enable its application in hospital.

References

I. American Cancer Society (2017). Cancer Facts & Figures 2017. Atlanta, GA: American Cancer Society.

II. Arslan, A. A., Helzlsouer, K. J., Kooperberg, C., Shu, X.-O., Steplowski, E., Bueno-De-Mesquita, H. B., et al. (2010). Anthropometric measures, body mass index, and pancreatic cancer: a pooled analysis from the Pancreatic Cancer Cohort Consortium (PanScan). Arch. Intern. Med. 170, 791–802. doi: 10.1001/archinternmed.2010.6.

III. Association, A. D. (2014). Diagnosis and classification of diabetes mellitus. Diabetes Care. 37, S81–S90. doi: 10.2337/dc10-S062

IV. Bakpo, F., and Kabari, L. (2011). "Diagnosing skin diseases using an artificial neural network," in Artificial Neural Networks-Methodological Advances and Biomedical Applications, ed K. Suzuki (InTech), 253–270.

V. Ben, Q., Xu, M., Ning, X., Liu, J., Hong, S., Huang, W., et al. (2011). Diabetes mellitus and risk of pancreatic cancer: a meta-analysis of cohort studies. Eur. J. Cancer. 47, 1928–1937. doi: 10.1016/j.ejca.2011.03.003

VI. Blewett, L. A., Rivera Drew, J. A., Griffin, R., King, M. L., and Williams, K. C. W. (2017). IPUMS Health Surveys: National Health Interview Survey, Version 6.2 [dataset]. Minneapolis, MN: University of Minnesota.

VII. Boursi, B., Finkelman, B., Giantonio, B. J., Haynes, K., Rustgi, A. K., Rhim, A. D., et al. (2017). A clinical prediction model to assess risk for pancreatic cancer among patients with new-onset diabetes. Gastroenterology. 152, 840–850.e843. doi: 10.1053/j.gastro.2016.11.046

VIII. Boursi, S. B., Finkelman, B., Giantonio, B. J., Haynes, K., Rustgi, A. K., Rhim, A., et al. (2018). A clinical prediction model to assess risk for pancreatic cancer among patients with pre-diabetes. J. Clin. Oncol. 36(15_Suppl.). doi: 10.1200/JCO.2018.36.15_suppl.e16226

IX. Iodice, S., Gandini, S., Maisonneuve, P., and Lowenfels, A. B. (2008). Tobacco and the risk of pancreatic cancer: a review and meta-analysis. Langenbecks Arch. Surg. 393, 535–545. doi: 10.1007/s00423-007-0266-2

X. Klein, A. P., Lindström, S., Mendelsohn, J. B., Steplowski, E., Arslan, A. A., Bueno-De-Mesquita, H. B., et al. (2013). An absolute risk model to identify individuals at elevated risk for pancreatic cancer in the general population. PLoS ONE. 8:e72311. doi: 10.1371/journal.pone.0072311.

XI. Lucenteforte, E., La Vecchia, C., Silverman, D., Petersen, G., Bracci, P., Ji, B. A., et al. (2011). Alcohol consumption and pancreatic cancer: a pooled analysis in the International Pancreatic Cancer Case–Control Consortium (PanC4). Ann. Oncol. 23, 374–382. doi: 10.1093/annonc/mdr120.

XII. Michaud, D. S., Vrieling, A., Jiao, L., Mendelsohn, J. B., Steplowski, E., Lynch, S. M., et al. (2010). Alcohol intake and pancreatic cancer: a pooled analysis from the pancreatic cancer cohort consortium (PanScan). Cancer Causes Control. 21, 1213–1225. doi: 10.1007/s10552-010-9548-z.

XIII. Poley, J. W., Kluijt, I., Gouma, D. J., Harinck, F., Wagner, A., Aalfs, C., et al. (2009). The yield of first-time endoscopic ultrasonography in screening individuals at a high risk of developing pancreatic cancer. Am. J. Gastroenterol. 104:2175–2181. doi: 10.1038/ajg.2009.276

XIV. Pannala, R., Basu, A., Petersen, G. M., and Chari, S. T. (2009). New-onset diabetes: a potential clue to the early diagnosis of pancreatic cancer. Lancet Oncol. 10, 88–95. doi: 10.1016/S1470-2045(08)70337-1

XV. Pannala, R., Leirness, J. B., Bamlet, W. R., Basu, A., Petersen, G. M., and Chari, S. T. (2008). Prevalence and clinical profile of pancreatic cancer–associated diabetes mellitus. Gastroenterology. 134, 981–987. doi: 10.1053/j.gastro.2008.01.039