20	Histology	no, yes

Experiment results and evaluation metrics

Three classification algorithms were used namely, Logistics Regression, Support Vector Machines and Naïve bayes classifiers. Logistic regression is a statistical algorithm used in the field of machine learning to solve classification and regression problems [47]. Support vector machines: a machine-based technology, is a class separation approach, which depends on statistical learning theory developed by Vapnik constructs a maximum margin separator also known as decision boundary with largest possible distance to example points. It creates a separating hyperplane in the original two-dimensional space. Maps the input variable to an n-dimensional feature space, also solves some regression problem. Support Vector Machine SVM A supervised machine learning algorithm mainly used for binary classification problem. It is trained by feeding a dataset with labeled examples (xi, yi). Where x represents features and y represent the target variable. Datasets are defined as n-dimensional feature vector that can be plotted on n-dimensional space. And Naïve Bayes: a method used for classifying objects based on closest training examples in the feature space, the most basic type of instance-based learning or lazy learning. It assumes all instances are points in n-dimensional space[48]. To measure the performance of each of the models, accuracy, precision (the proportion of positively classified results either true positive or true negative) and recall of confusion matrix were used. It is a 2 x 2 matrix which compares the predicted class with actual class. The evaluation metrics can then be defined as follows: predictive accuracy is the proportion of correctly classified outcomes either true positive or true negative.

Accuracy=(TP+TN)/(TP+FP+FN+TN) … (3)

Precision=TP/(TP+FP) …….. (4)

Recall=TP/(TP+FN) ……. …. (5)

True Positive (TP): number of patients who are classified correctly. False Positive (FP) also known as type I error, number of patients classified wrongly. False Negative also known as type II error and True Negative (TN). For experimentation, three classical machine learning algorithms were used. They are NB, LR and SVM. Support Vector Machine Logistic fits a logistic regression model to the data with a ridge estimator.

Feature selection results

In effort to obtained optimal parameter for Hepatitis disease from Hepatitis disease dataset obtained from UCI as described above, three different dimensionality reduction approaches (feature selection) were applied to the datasets. Chi-Square, PCA and genetic Algorithm and were able to select a subset of features presented in table 1 for Chi-square test and Genetic Algorithm and table 2 for principal components or optimal variables created by PCA respectively. The resulting subset of features ware used for classification where Logistics regression, Support Vector Machines and Naïve bayes classifiers were used.

Table 2: Selected subset of features.

s/n	Chi-Square	Genetic Algorithm
1	Age	Age
2	Malaise	Sex
3	Spiders	Steroid
4	Ascities	Fatigue
5	Bilirubin	Anorexia
6	Alk_phosphate	alk_phosphate
7	Sgot	Sgot
8	Albumin	Albumin
9	Histology	Protime

Table 2 above contained the selected features by Chi-square and Genetic Algorithm, we ran Genetic algorithm for 10 generations and the optimal individuals that achieve the highest validation accuracy are the features used for the classification in all the three ML classifiers. Chi-Square test also selected top 9 features based on the feature scores and the newly formed datasets was used for the classification.

Figure 3 shows the sample principal components formed as a result of dimensionality reduction from 19dimensions to 2, shows the condensed information from all features incorporated in principal component one and principal component 2.

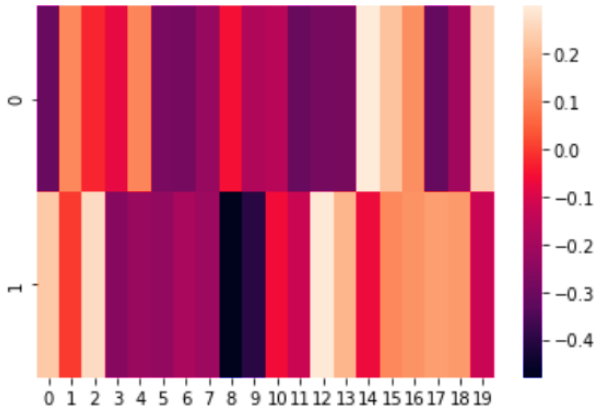

Figure 3: Principal components sample.

Classification results

After getting features selected three classifiers were built and the resulting accuracy are presented in the figures below.

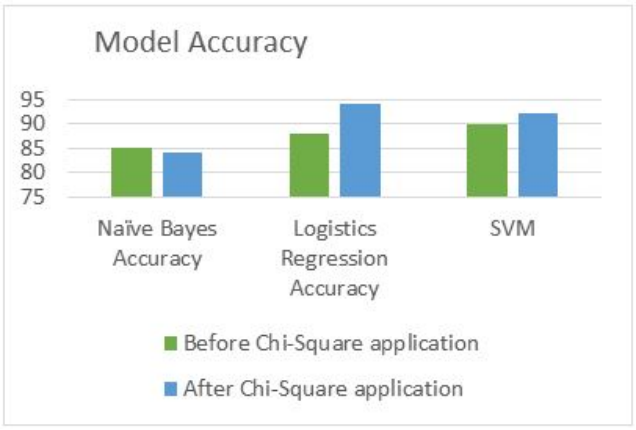

Figure 4: Model accuracy on Chi-square FS.

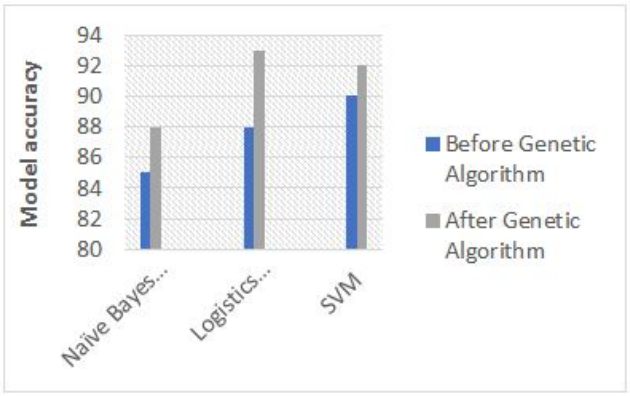

Figure 5: Model accuracy on GA FS.

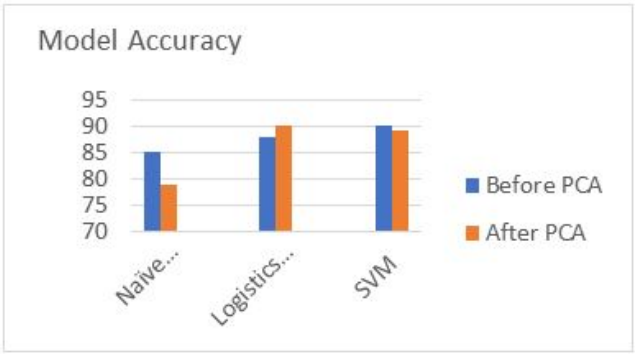

Figure 6: Model accuracy on PCA FS.

CONCLUSION

In the development of Machine learning model for classification of diagnosis purposes, the model performance is of outmost importance especially in medicine, type I and type II errors needs to be minimized. One approach to achieving this is through dimensionality reduction as reviewed in the literature. In this study, three feature selection approaches have been implemented for optimal parameter selection of hepatitis disease classification. These approaches have shown that feature selection is an important tool for improving model performance for most of machine learning algorithms. The three classifiers built the combination of Chi-square and Logistic regression produced the highest classification accuracy whereas in the case of Principal component analysis and Naïve Bayes classifier produced the least performance accuracy with lower accuracy than obtained with full dimensions of the original datasets, which is the only case where FS reduce performance accuracy. Therefore, we conclude that feature selection is a great tool for increasing performance of a model, However, the combination of a Feature selection approach with a classifier is an important one, different FS approach have good compatibility with certain classifiers and vice versa.

REFERENCES

1. http://apps.who.int/iris/bitstream/10665/255016/1/9789241565455-eng.pdf?ua=1
2. Antony Leo Jerry T, S Sundari. A study of estimation and analysis of hepatitis B antibody titres in adolescent children attending a tertiary care hospital. J Res Med Dent Sci 2021; 9:310-317.
3. Choi E, Bahadori MT, Schuetz A, et al. Doctor AI: Predicting clinical events via recurrent neural networks. InMachine learning for healthcare conference 2016.
4. Boukar M, Dane S, The effects of sex, education and marital status on alexithymia. J Res Med Dent Sci 2019; 7:82–85.
5. Abd El-Salam SM, Ezz MM, Hashem S, et al. Performance of machine learning approaches on prediction of esophageal varices for Egyptian chronic hepatitis C patients. Informatics Med Unlocked 2019; 17:100267.

6. Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to algorithms. Cambridge university press 2014.
7. Brewka G. Artificial intelligence-A modern approach by Stuart Russell, Peter Norvig, Prentice Hall. Series in artificial intelligence, Englewood Cliffs, NJ 1996; 11.
8. https://kkpatel7.files.wordpress.com/2015/04/alppaydin_machinelearning_2010.pdf
9. http://www.cs.cmu.edu/~tom/mlbook.html
10. Kumar NK, Vigneswari D. Hepatitis-infectious disease prediction using classification algorithms. Res J Pharmacy Technol 2019; 12:3720-5.
11. Suryakirani RP, Porkodi R. Comparative study and analysis of classification algorithms in data mining using diabetic dataset. Int J Scientific Res Sci Technol 2018; 4:299-304.
12. Tomar D, Agarwal S. Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes. Adv Artificial Neural Systems 2015; 2015.
13. Lanzi PL. Fast feature selection with genetic algorithms: a filter approach. InProceedings of 1997 IEEE International Conference on Evolutionary Computation (ICEC'97) 1997.
14. Guha R, Ghosh M, Kapri S, et al. Deluge based genetic algorithm for feature selection. Evolutionary Intelligence 2019; 1-1.
15. Lowe DG. Concrete making materials. Concr Prod 2002; 20.
16. Muhammed MM, Boukar MM, Aldullahi SE. The application of artificial intelligence technique (CNN-Alexnet) in diagnosing COVID-19 using chest X-ray images. J Res Med Dent Sci 2021; 9:21-26.
17. Sayed S, Nassef M, Badr A, et al. A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. Expert Systems App 2019; 121:233-43.
18. Hong JH, Cho SB. Efficient huge-scale feature selection with speciated genetic algorithm. Pattern Recognition Letters 2006; 27:143-50.
19. Ghamisi P, Benediktsson JA. Feature selection based on hybridization of genetic algorithm and particle swarm optimization. IEEE Geoscience Remote Sensing Letters 2014; 12:309-13.
20. Bhanu B, Lin Y. Genetic algorithm based feature selection for target detection in SAR images. Image Vision Computing 2003; 21:591-608.
21. Fröhlich H, Chapelle O, Schölkopf B. Feature selection for support vector machines by means of genetic algorithms. InProceeding ICTAI 2002; 3.
22. Huang J, Cai Y, Xu X. A hybrid genetic algorithm for feature selection wrapper based on mutual information. Pattern Recog Letters 2007; 28:1825-44.
23. Jadhav S, He H, Jenkins KW. An academic review: Applications of data mining techniques in finance industry. 2008.
24. Kabir MM, Shahjahan M, Murase K. A new local search based hybrid genetic algorithm for feature selection. Neurocomputing 2011; 74):2914-28.
25. Tan F, Fu X, Zhang Y, et al. A genetic algorithm-based method for feature subset selection. Soft Comput 2008; 12:111-20.
26. Spencer R, Thabtah F, Abdelhamid N, et al. Exploring feature selection and classification methods for predicting heart disease. Digital Health 2020; 6:2055207620914777.
27. Alshaer HN, Otair MA, Abualigah L, et al. Feature selection method using improved CHI Square on Arabic text classifiers: Analysis and application. Multimedia Tools App 2021; 80:10373-90.
28. Jin X, Xu A, Bie R, et al. Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. International workshop on data mining for biomedical applications. Springer 2006.
29. Huang LL, Tang J, Chen SB, et al. An efficient algorithm for feature selection with feature correlation. In International Conference on Intelligent Science and Intelligent Data Engineering. Springer 2012.
30. Khorashadizade N, Rezaei H. New method for rapid diagnosis of Hepatitis disease based on reduction feature and machine learning. J Adv Computer Sci Technol 2015; 4:148.
31. Sowmien VS, Sugumaran V, Karthikeyan CP, et al. Diagnosis of hepatitis using decision tree algorithm. Int J Eng Technol 2016; 8.
32. Avci D. An automatic diagnosis system for hepatitis diseases based on genetic wavelet kernel extreme learning machine. J Elec Eng Technol 2016; 11:993-1002.
33. Nancy P, Sudha V, Akiladevi R. Analysis of feature selection and classification algorithms on hepatitis data. Int J Adv Res Comp Eng Technol 2017; 6:19-23.
34. Chown H. A comparison of machine learning algorithms for the prediction of Hepatitis C NS3 protease cleavage sites. J Proteomics Bioinform 2019; 12:088-93.
35. Bayrak EA, Kirci P, Ensari T. Performance analysis of machine learning algorithms and feature selection methods on hepatitis disease. Int J Multidisciplinary Studies Innovative Technol 2019; 3:135-8.
36. Nilashi M, Ahmadi H, Shahmoradi L, et al. A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique. J Infection Public Health 2019; 12:13-20.
37. Meyer D, Leisch F, Hornik K. The support vector machine under test. Neurocomputing 2003; 55:169-86.
38. Yarasuri VK, Indukuri GK, Nair AK. Prediction of hepatitis disease using machine learning technique. In 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) 2019.

39. Jolliffe IT, Cadima J. Principal component analysis: A review and recent developments. Philosophical Transactions Royal Society A: Mathematical Physical Eng Sci 2016; 374:20150202.
40. https://files.isec.pt/DOCUMENTOS/SERVICOS/BIBLIO/Documentos%20de%20acesso%20remoto/Principal%20components%20analysis.pdf
41. Kong H, Wang L, Teoh EK, et al. Generalized 2D principal component analysis for face image representation and recognition. Neural Networks 2005; 18:585-94.
42. Tay FE, Shen L. A modified chi2 algorithm for discretization. IEEE Transactions on knowledge and data engineering 2002; 14:666-70.
43. H. Liu and R. Chi2: Feature selection and discretization of numeric attributes. Proc Int Conf Tools Artif Intell 1995; 388–391.
44. Mirjalili S. Genetic algorithm. Stud Comput Intell 2019; 780:43–55.
45. Mitchell M. An introduction to genetic algorithms. An Introd Genet Algorithms 2020; 1–40.
46. http://www.geocities.ws/francorbusetti/gabeasley1.pdf
47. Akkoç S. An empirical comparison of conventional techniques, neural networks and the three stage hybrid adaptive neuro fuzzy inference system (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. Eur J Operational Res 2012; 222:168-78.
48. Saxena K, Khan Z, Singh S. Diagnosis of diabetes mellitus using k nearest neighbor algorithm. Int J Computer Sci Trends Technol 2014; 2:36-43.