A Developed Prediction Approach for Small Datasets

Nuha Ahmed Salman[1], Saad Talib Hasson[2]

Software Department- College of Information Technology- University of Babylon, [1].Iraq, Hilla

[2].College of Information Technology- University of Babylon, Iraq, Hilla

Emails: [1] nuhaahmed562@gmail.com , [2]saad_aljebori@itnet.uobabylon.edu.iq

Abstract:

Using small datasets is a very challenging task. Developed approaches are significant to make accurate predictions, rankings and identify relevant attributes with a reliable model. Small dataset may lead to biased or incomplete models. Analyzing and prediction with small datasets is usually more difficult due to the restricted number of data. Identifying the relationships between the small dataset's attributes, specific attributes may have a greater chance of ranking than others due to their insufficient availability. Machine learning has a flexibility and possibility to explain different prospects for health investigation when dealing with small datasets. This paper will analyze the difficulties in making predictions from small datasets and look at several strategies and algorithms that may be applied to reduce these difficulties and increase prediction accuracy .

Keywords: (small dataset, prediction, Machine learning, attributes, correlation).

Introduction

Predictive modeling is a crucial tool for decision-making in today's data-driven society. Predictive modeling may be difficult when dealing with small datasets. Small datasets can produce biased or inaccurate models, which may result in incorrect forecasts and poor healthcare industries. For instance, because patient information is sensitive, it might be challenging to gather massive datasets in the healthcare industry [1]. Analysts and researchers must use a variety of approaches to get over the constraints of small datasets and make precise predictions [2].

The term "small dataset" refers to a dataset with only a handful of examples or instances, which might make it challenging to create precise prediction models. The small dataset can frequently be attributed to a lack of funding, ethical issues, or the specifics of the issue being investigated [3]. Because there may not be enough data for machine learning models to discover reliable patterns and because they may be overfitted or underfitted, small datasets can be a substantial barrier for them [1]. Underfitting happens if a model is too simplistic to identify the underlying trends in the data, whereas overfitting occurs when a system is too complicated for the data it is attempting to learn [2]. Investigating strategies and approaches that might enhance machine learning models' performance on limited datasets is crucial in this setting [3]. Even when there is little available data, research and data scientists may nevertheless derive significant insights and make precise forecasts in this method [4]. The difficulty of producing precise forecasts or well-informed judgments based on little amounts of data is referred to as small dataset prediction. This is a significant problem in a number of industries, including healthcare, Small datasets may be biased and provide overly simple or overfitted models, which can produce incorrect predictions. Therefore, to augment the little data and create reliable models that can precisely anticipate events or help decision-makers, analysts and researchers must employ a variety of methodologies [5].

The performance of the model should be optimized in this situation by carefully selecting the suitable algorithms for machine learning and fine-tuning their parameters. To get around the problem of small datasets, data enhancement, machine learning, and ensemble approaches can be employed. Overall, building robust models that may offer insightful information involves careful examination of the approaches and algorithms that are now available for the reliable forecasting of outcomes from small datasets [3] [5].

Machine Learning

Machine learning implementation is difficult with little datasets because the algorithms might not have enough information to identify patterns and successfully generalize to new data. However, there are certain methods that can assist machine learning models to perform better on small datasets. Most of these methods are as follows [3] [5] [6].

a. Features selection: With a small dataset, it's crucial to concentrate on the features that are most pertinent to the problem and most likely to predict the outcome.
b. Data augmentation: Creating new instances based on existing ones is a technique that may be used to artificially enhance the size of a dataset. When the dataset is small, in particular, this can aid in enhancing the model's generalization performance. This process can expand the dataset and give the model additional variance to draw from while learning.
c. Assembling: it is a method for enhancing performance by combining the predictions of many models. Assembling might be used to aggregate the predictions of multiple straightforward models on a small dataset.

Dataset

Predicting heart disease is a significant issue that can be helped by machine learning approaches. However, it might be difficult to forecast cardiac disease when dealing with a little sample. In this paper, a heart disease dataset downloaded from "**https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction**" is utilized.

This dataset is composed of 270 records 15 and attributes. It is important to improve the chances of creating a machine-learning model for heart disease prediction that works well on a small dataset. However, it's crucial to remember that working with a small dataset has its drawbacks and that the model's performance could be constrained by the quantity of data available.

SMOTE

A technique called Synthetic Minority Over-Sampling Technique (SMOTE) is used to deal with small datasets. Such datasets are unbalanced, where samples in one class are significantly underrepresented compared to samples in the other class or classes. By integrating between the actual data points, the SMOTE algorithm generates artificial points of data in the minority class [7].

Making ensuring that the created synthetic data points are accurate and true to the original data is crucial when employing SMOT on a small dataset. This is possible by carefully adjusting the SMOTE settings, such as the amount of artificial samples to produce and the interpolation distance metric employed between the initial data points [6].

It is crucial to assess the model's performance using the proper assessment measures, such as accuracy, recall, and F1-score, once it has been trained on the supplemented dataset. To make certain that the model performs consistently across several data folds, cross-validation methods may also be helpful [8][9].

SMOT is a helpful strategy for handling unbalanced datasets, but it must be used carefully and its efficacy must be determined case-by-case [10]. SMOTE is implemented on this dataset in this paper. Figure 1, shows a sample of the created data.

index	Age	Sex	Chest	BP	cholestero	FBS	EKG	Max	Exercise	ST	Slope of st	fluro	thallium	Heart
0	70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence
1	67	0	3	115	564	0	2	160	0	1.6	2	0	7	Absence
2	57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence
3	64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence
4	74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence
5	65	1	4	120	177	0	0	140	0	0.4	1	0	7	Absence
6	56	1	3	130	256	1	2	142	1	0.6	2	1	6	Presence
7	59	1	4	110	239	0	2	142	1	1.2	2	1	7	Presence
8	60	1	4	140	293	0	2	170	0	1.2	2	2	7	Presence
9	63	0	4	150	407	0	2	154	0	4	2	3	7	Presence
10	59	1	4	135	234	0	0	161	0	0.5	2	0	7	Absence
11	53	1	4	142	226	0	2	111	1	0	1	0	7	Absence
12	44	1	3	140	235	0	2	180	0	0	1	0	3	Absence
13	61	1	1	134	234	0	0	145	0	2.6	2	2	3	Presence
14	57	0	4	128	303	0	2	159	0	0	1	1	3	Absence
15	71	0	4	112	149	0	0	125	0	1.6	2	0	3	Absence
16	46	1	4	140	311	0	0	120	1	1.8	2	2	7	Presence
17	53	1	4	140	203	1	2	155	1	3.1	3	0	7	Presence
18	64	1	1	110	211	0	2	144	1	1.8	2	0	3	Absence
19	40	1	1	140	199	0	0	178	1	1.4	1	0	7	Absence
20	67	1	4	120	229	0	2	129	1	2.6	2	2	7	Presence
21	48	1	2	130	245	0	2	180	0	0.2	2	0	3	Absence
22	43	1	4	115	303	0	0	181	0	1.2	2	0	3	Absence

Figure **(1):** SMOTE data generation

Ranking of Attributes

For a specific prediction issue, an approach called attribute ranking is used to determine which attributes in a dataset are most crucial. When dealing with a limited dataset, attribute ranking may be very helpful in determining which characteristics will contribute the most information to the model. Employing a statistical analysis step, may help in indicating the relevance of certain attributes in a small dataset. An extended data after implementing a SMOTE also tested by a statistical analysis. This test can aid in locating characteristics that have a high correlation with the result variable [11].

Indicating the feature importance is contributing in determine which characteristics in a small dataset are the most useful. Correlation matrices can be used to do this [12]. These methods for ranking features in a small dataset allow to determine which characteristics are most crucial for a certain prediction issue and enhance the effectiveness of the machine learning model. Table (1) presents the attributes ranking for both the small and the extended dataset.

Table (1): Attribute ranking for small and extended datasets.

	SMOTE		Small dataset	
Rank	Attribute name	Old No.	Attribute name	Old No.
1	Thallium	14	Thallium	14
2	Number of vessels fluor	13	Number of vessels fluor	13
3	ST depression	11	Exercise angina	10
4	Exercise angina	10	Max HR	9
5	Max HR	9	Depression St	11
6	Chest pain type	4	Chest pain type	4
7	Slope of St	12	Slope of St	12
8	sex	3	sex	3
9	age	2	age	2
10	EKG results	8	EKG results	8

11	Bp	5	Bp	5
12	Cholesterol	6	Cholesterol	6
13	FBS over120	7	FBS over120	7
14	Index	1	Index	1

Table (1) shows that the attribute ranking of the extended data is very close that for the small dataset. Such results are a good indication that the generated virtual data behavior is very close or similar to the original real small or limited data. One advantage of attribute ranking is to indicate the most important attributes. The least important attributes can be eliminated or avoided in any reduction process. Also the significant attributes must be considered in any prediction model [13].

Feature reduction

It is necessary to choose suitable features for the proposed prediction model. This requires an understanding of how the data is structured and how each feature will affect the model. It is important to identify key features that will help the model make accurate predictions and rankings [14]. When dealing with a limited dataset, feature reduction is a technique used to minimize the number of features in the dataset. For feature reduction with a limited dataset, significant steps can be considered [12][14]. These steps are:

1. Making use of statistical testing to determine the importance of certain features in a limited dataset. Statistical testing can aid in locating characteristics that are safely deleted and least strongly connected with the result variable.
2. Employing correlation analysis between any attribute's pairs. A strong association with one another can be found using correlation analysis. In such circumstances, one of the traits can be eliminated with little to no information loss.
3. Using feature selection to automatically determine the most significant characteristics in a dataset. Statistical measures can be used to score the dependence or the correlation between input variables to choose the most relevant features.

By decreasing the number of features in a small dataset one may enhance the performance of the machine learning model while lowering the likelihood that it will overfit and making the model easier to understand. However, it's crucial to bear in mind that feature reduction can also lead to information loss, thus it's crucial to carefully consider how each feature removal will affect the performance of the model [8]. A correlation testing is performed on the selected dataset in this paper to indicate the relationship and the effect of each attribute on the others. Table (2) shows a sample of the calculate correlation coefficient between any attribute's pairs in a small dataset, while table (3) shows a sample from the calculated correlation coefficient between any attribute's pairs in an extended dataset. Only a strong correlation (larger or equal to 0.7) is considered in this paper analysis.

Table (2): correlation coefficient matrix for small dataset.

	ag	se	Che	Bp	Cho	FBS	EK	Ma
Thalliu	0.1	0.3	0.26	0.132	0.0288	0.0492	0.0073	0.253-
Number	0.3	0.0	0.22	0.08	0.12	0.12	0.11	0.26
Slope of	0.1	0.0	0.13	0.14	0.00	0.04	0.16	0.38
Depressi	0.1	0.0	0.16	0.22	0.02	0.02	0.12	0.34
Exercise	0.0	0.1	0.35	0.08	0.07	0.00	0.09	0.38
Max	0.4	0.0	0.31	0.03	0.01	0.02	0.07	
EKG	0.1	0.0	0.07	0.11	o.16	0.05		0.07
FBS	0.1	0.0	0.09	0.15	0.02		0.05	0.02
Choleste	0.2	0.2	0.09	0.17		0.02	o.16	0.01
Bp	0.2	0.0	0.04		0.17	0.15	0.11	0.03
Chest	0.0	0.0		0.04	0.09	0.09	0.07	0.31
sex	0.0		0.03	0.06	0.20	0.04	0.03	0.07
age		0.0	0.09	0.27	0.22	0.12	0.12	0.40

Exe	**0.09**	**0.18**	**0.35**	**0.08**	**0.07**	**0.00**	**0.09**	**0.38**		**0.27**	**0.25**	**0.15**	**0.321**
Dep	**0.19**	**0.09**	**0.16**	**0.22**	**0.02**	**0.02**	**0.12**	**0.34**	**0.27**		**0.61**	**0.25**	**0.324**
Slope	**0.16**	**0.050**	**0.137**	**0.142**	**0.005**	**0.044**	**0.161**	**0.387**	**0.256**	**0.61**		**0.109**	**0.284**
Num	**0.356**	**0.086**	**0.226**	**0.085**	**0.127**	**0.124**	**0.114**	**0.265**	**0.153**	**0.255**	**0.109**		**0.256**
Thalli	**0.106**	**0.391**	**0.263**	**0.132**	**0.0288**	**0.0492**	**0.0073**	**0.253-**	**0.321**	**0.324**	**0.284**	**0.256**	

Tables (2) and (3) gave a good indication of the important attributes and the other not significant ones. Important attributes are those affecting other attributes or correlated with others. Eliminating these attributes will affect the others and affect the dataset. Such attributes must be avoided in any reduction proposal. In this paper, any correlation coefficient greater or equal to 0.7 will be considered as a strong correlation. Other attributes can be eliminated or considered in a process of attribute reduction.

Statistical tests

Heart disease prediction often entails looking at a variety of factors that might be connected to the onset of heart disease. The importance of these factors may be determined using statistical tests, and those that are most closely linked to the outcome variable—the existence or absence of heart disease—can be found [15][16]. In this paper, certain statistical measures are also utilized such as the attribute's mean and their standard deviations. Table (5) shows the calculated statistical values for the small and extended dataset attributes.

Table (3): mean and standard deviation values.

Attribute	Std	Av	Max
Index	78.086	134.5	269
age	9.109	54.433	77
sex	0.468	0.678	1
Chest pain type	0.95	3.174	1
Bp	17.862	131.344	200
Cholesterol	51.686	249.659	564
FBS over120	0.356	0.148	1
EKG results	0.998	1.022	2
Max HR	149.678	149.678	202
Exercise angina	0.471	0.33	1
St Depression	1.145	1.05	6.2
Slope of St	0.614	1.585	3
Number of vessels Fluro	0.944	0.67	3
Thallium	1.941	4.696	7
Heart Disease			

Predictive modeling

When dealing with small datasets, low-complexity models like Naïve base, Decision tree, KNN, and Logistic Regression, may generalize the best predictions. Their measures are indicated in Table (4). The confusion matrix for each model is shown in Tables (5), (6), (7), (8), (9), and (10), respectively.

Table (4) Detailed accuracy by class

Algorithm type	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Naïve bayes	0.876 0.913 0.896	0.897 0.897 0.897	0.662 0.662 0.662	0.809 0.852 0.833	0.792 0.867 0.833	1.826 0.839 0.833	0.133 0.208 0.175	0.792 0.867 0.833	Presence Absence Weighted Avg
Decision tree (J48)	0.617 0.723 0.676	0.726 0.726 0.726	0.490 0.490 0.490	0.717 0.773 0.748	0.717 0.773 0.748	0.717 0.773 0.748	0.227 0.283 0.258	0.717 0.773 0.748	Presence Absence Weighted Avg
KNN	0.664 0.746 0,710	0.768 0.768 0.768	0.536 0.536 0.536	0.744 0.792 0.771	0.750 0.787 0.770	0.738 0.797 0.771	0.213 0.250 0.234	0.750 0.787 0.770	Presence Absence Weighted Avg
Logistic	0.887	0.899	0.662	0.810	0.800	0.821	0.140	0.800	Presence

	0.91 4 0.90 2	0.89 9 0.89 9	0.66 2 0.66 2	0.85 1 0.83 3	0.86 0 0.83 3	0.84 3 0.83 3	0.20 0 0.17 3	0.86 0 0.83 3	Absenc e Weight ed Avg

Table (5): Confusion Matrixes (Naïve Bayes)

	Presence	Absence
Presence	25	95
Absence	130	20

Table (6): Confusion Matrixes (decision tree)

	Presence	Absence
Presence	34	86
Absence	116	34

Table (7): Confusion Matrixes (KNN)

	Presence	Absence

Presence	30	90
Absence	118	32

Table (8): Confusion Matrixes (logistic regression)

	Presence	Absence
Presence	24	96
Absence	129	21

Table (9): probability for each attribute

Attributes type	Prediction numeric attributes	Normal distribution\ Probability
Index	-3.43211	0.0003
age	-1.65155	0.0495
sex	-1.50427	0.0668
Chest pain type	-1.71789	0.0436
Bp	-2.08219	0.0188
Cholesterol	0.413806	0.1591

FBS over120	0.94382	0.3264
EKG results	0.104208	0.0398
Max HR	0.289261	0.1103
Exercise angina	-0.93418	0.1762
Depression St	0.174672	0.0675
Slope of St	-1.27036	0.1020
Number of vessels Fluro	-1.65254	0.0495
thallium	0.243174	0.0948

Table (10): Algorithm accuracy

Algorithm type	Accuracy	Mean absolute Error	time
Naïve Bayes	83.3333	(0.184)	0
Decision Tree	74.8148	(0.2888)	0.04
KNN	77.037	(0.2318)	0
Logistic	83.3333	(0.6617)	0.04

Discussion

Experiment results present the performance behavior of different four machine learning models (Naïve based, Decision tree, K-Nearest Neighbor (KNN), and Logistic regression) on a heart disease small dataset.

Naive Bayes classifiers are labeled as classifiers that are relatively similar to linear models. They use Bayes' theorem to give probabilities to classes. The main limitation of these approaches is the need for independent predictors which reduces the classifier's performance with short prediction times.

A decision tree is a supervised learning algorithm. It commonly simulates a human philosophical ability while making a decision. A decision tree can be utilized for both classification and regression tasks. A decision tree is a very precise type of probability tree that permits one to make a decision about some kind of course.

KNN is one of the simple machine learning models. To make a prediction for a new data point, the algorithm finds the point that is closest to the new point in the training set. Then it assigns the label or output of this training point to the new data point.

Logistic regression represents one of the essential supervised machine learning algorithms. It is a statistical model that defines the probability of incidence of a class by fitting a logistic curve to the dataset. It can be used to predict a continuous or floating-point number.

These models' performances were estimated in terms of the average, standard deviation, and ranges. Their True Positive Rate (TP), False Positive Rate (FP), Precision, F-measure, and Matthews Correlation Coefficient (RCC) were used as evaluation metrics to evaluate the performance of these classification models.

TP is used to indicate the ratio of correctly classified positive cases to the total number of positive cases in the used dataset. FP is used to represent the ratio of incorrectly classified negative cases to the total number of negative cases. Precision represents a measure of the fraction of positive predictions that are true. It also presents the ratio of correctly classified positive occurrences to the total number of occurrences that are classified as positive. The recall is a metric to indicate the ratio of correctly classified positive cases to the total number of positive cases. It can be utilized to show the model's ability to recognize all positive cases. F-measure is an important metric (A weighted harmonic mean of recall and precision) to offer a balance between recall and

precision metrics. While MCC is used to measure the quality of both true and false positives and negatives.

Naïve Bayes represent the best from the accuracy and mean square error.

Conclusion

The rapid growth of healthcare data is considered to be a challenging task, it can be utilized in different personal care and early detection predictions. Applying Machine-Learning procedures will contribute in a positive influence on health data analysis. Small dataset prediction is a difficult task. It is feasible to utilize machine learning in creating models that can predict acceptable outcomes from small datasets.

References

using Machine Learning over Big Data", Computer Science & Engineering: An International Journal (CSEIJ), Vol.8, No.1, February 2018.

[2] Bakyarani, S., Srimathi, H., & Bagavandas, M., (2019). "A Survey Of Machine Learning Algorithms In Health Care" International Journal Of Scientific & Technology Research Volume 8, Issue 11.

[3] Azadi, A.; García-Peñalvo, F.J. Synergistic Effect of Medical Information Systems Integration: To What Extent Will It Affect the Accuracy Level in the Reports and Decision-Making Systems? Informatics 2023, 10, 12.

[4] Asif Ahmed Neloy, SazidAlam, Rafia Alifbindu, Nusrat Jahan Moni, "Machine Learning-based Health Prediction System using IBM Cloud as PaaS", Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019).

[5] Dr S. Mohan Kumar, DarpanMajumder, “Healthcare Solution based on Machine Learning Applications in IoT and Edge Computing”, International Journal of Pure and Applied Mathematics Volume 119 No. 16 2018.

[6] Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W., & Collins, G. S. (2019). Predictive analytics in health care: how can we know it works?. Journal of the American Medical Informatics Association, 26(12), 1651-1654.

[7] K.Karthika, G.Nagarajan, "Disease Prediction by Machine Learning Over Big Data from Healthcare Communities", Journal of Recent Research in Engineering and Technology, Volume 4 Issue 11 Nov 2017.

[8] Obulesu, O., Mahendra, M., & ThrilokReddy, M. (2018, July). Machine learning techniques and tools: A survey. In 2018 International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 605-611). IEEE.

[9] Imran, A.S.; Yang, R.; Kastrati, Z.; Daudpota, S.M.; Shaikh, S. The impact of synthetic text generation for sentiment analysis using GAN based models. Egypt. Inform. J. 2022, 23, 547–557.

[10] Chugh, M., Johari, R., & Goel, A. (2022). MATHS: Machine Learning Techniques in Healthcare System. In International Conference on Innovative Computing and Communications (pp. 693-702). Springer, Singapore.

[11] Alam, M.Z.; Rahman, M.S.; Rahman, M.S. A Random Forest based predictor for medical data classification using feature ranking. Inform. Med. Unlocked 2019, 15, 100180.

[12] Rashid, K.M.; Louis, J. Times-series data augmentation and deep learning for construction equipment activity recognition. Adv. Eng. Inform. 2019, 42.

[13] YichuanWang et al. “An integrated big data analytics-enabled transformation model: Application to health care”. In: Information & Management 55.1 (2018), pp. 64–79.

[14] E. Odhiambo, G. Onyango, and M. Waema, “Feature Selection for Classification using Principal Component Analysis and Information Gain,” Expert Syst. Appl., vol. 174, no. November 2020, p. 114765, 2021, doi: 10.1016/j.eswa.2021.114765.

[15] Khan, M. A., Ahmed, M. U., & Rahman, A. (2020). A review of machine learning and statistical techniques for heart disease prediction. Health informatics journal, 26(3), 2092-2113. care, 8(10), 3087-3092.

[16] Khan, M. F., Zafar, M. W., & Khan, M. A. (2017). Applications of statistical techniques in medical sciences: a literature review. Journal of data science, 15(3), 387-402.