JOURNAL PUNJAB ACADEMY OF SCIENCES (Peer Reviewed Open Access Journal)
JPAS, 23: 217-227 (January-December, 2023)
Available online at www.jpas.in
ISSN 2229-7014 (P)

HANDLING MISSING VALUES IN NUMERIC DATASET USING MACHINE LEARNING TECHNIQUES: A REVIEW

Kamaljeet Kaur[1], Dr. Amrit Kaur[2], Dr. Navjot Kaur[3]

Department of Computer Science and Engineering, Punjabi University Patiala, India[1,2,3]

ABSTRACT

Data mining is essential for pre-processing task to ensure the quality of the final product. These tasks include data preparation, cleaning, integration, transformation, reduction, and discretization. Missing values are a common problem that regularly occurs throughout the data cleaning process in various research fields. To complete missing values, eliminate noise and remove inconsistencies is an important process in the preparation of the data. This paper focuses on a review of several classification methods, including their benefits and shortcomings. It is used in a variety of industries, including internet marketing, healthcare, social networking, finance, and insurance. The accuracy of data imputation for machine learning classifiers such as Bayesian Networks, Decision Trees and K-Nearest Neighbors (KNN), as well as Support Vector Machines, is compared in this paper. Based on the findings, Bayesian appears to provide the most promising results when compared to the other classifiers.

Keywords: Data mining, Data classification, K-Nearest Neighbor, Decision Tree, Support Vector Machine, Naïve Bayes.

INTRODUCTION

Data mining is a current method for resolving several challenging real-world issues. Data preparation is the simple conversion of raw data into a usable format, according to Vivek. Data purification, integration, transformation, data reduction, and data discretization are the primary data pre-processing phases, as illustrated in figure 1 taken from [1]. Managing missing data is an important step in the pre-processing of data. The initial stage of preprocessing data, known as data cleansing, includes this technique [1]. Data mining is another term for the process of knowledge finding from databases. In order to extract hidden information from the available data, KDD is used. In addition to business, it is used in a variety of industries, including internet marketing, healthcare, social networking, finance, and insurance [2]. Data mining is the act of identifying connections and broad patterns that exist in enormous databases. The subset of data mining that works with spatial data is called spatial data mining. Another significant clustering technique with a purpose is the density-based algorithm [3]. Any automatable classification or regression job requires a learnable decision-making system and these systems all start with a dataset as its building block [4]. Several elements from various dataset sources contribute to this amusingness, which is the result of it. These include incorrect and wrong data entry, data availability issues, data gathering issues, missing features, missing files, insufficient information etc [5].

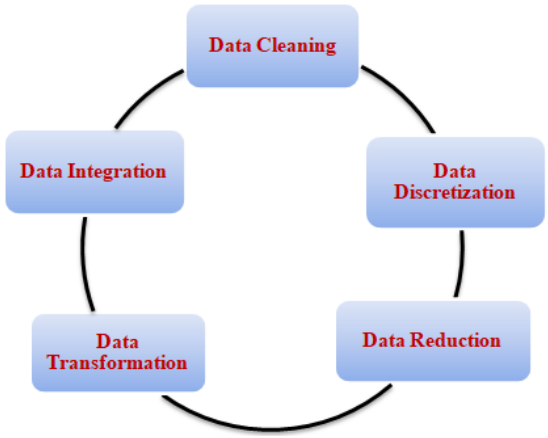

Fig.1. Data Mining Task [1]

1.1. Missing Values:

Missing data might happen if the value is not When information that would have been important to a certain instance could not be captured the information was gathered, or users disregard it as a result of Privacy issues. Research has been done on the issue of missing values, several years ago. Competent design is the only truly effective way to deal with missing data, although good analysis can reduce the issues. Issues resulting from lacking data: Accuracy loss because of insufficient information, Computation challenges brought caused by dataset gaps, Bias brought on by data distribution bias [6].To deal with missing values, a number of conventional statistical and machine learning imputation approaches, including mean, regression, K closest neighbor, and group based etc., have been developed in the literature. "Missing data patterns and techniques" is the part of research that discusses missing values. The Iris data set was then used to develop two machine learning algorithms, which are explained in the section under "Performance and training on machine learning approaches" [7]. The issue of missing data regularly occurs in all types of clinical research. Multiple imputations are one of the more advanced (imputation) approaches that manage missing data and produces substantially better outcomes. A case analysis that is complete when there are MARs for missing value is no longer based on a sample chosen at random from the source population and bias is probably present [8].

1.2. Missing data mechanism:

According to its definition, Data imputation is a method for substituting values for missing data. The three missing value methods are missing partially at random (MPAR), missing at random (MAR) and missing not at random (MNAR), according to Rubin. A missing value situation is more likely to be true when there are known values in addition to the missing value, or MAR data. MNAR, on the other hand, denotes a situation where the likelihood depends on the value of a missing value in a class instance of that variable [1]. The definition of missing data in the literature follows these processes since the majority of mechanisms that result in missing values on data have an impact on some presumptions that underlie the majority of handling strategies for missing data [7].

LITERATURE SURVEY:

In this paper, Data mining is also called Knowledge discovery from Database. Data mining is a process used by organizations to extract specific data from large amount of databases. Data mining is used to solve business problems as well as other fields like as digital marketing, health care, social media, banking and insurance. Data mining process includes stages Data cleaning, integration, Selection, Transformation and Pattern evaluation. In Data mining various techniques used such as Classification, Clustering, Prediction, Neural networks, Association, Artificial intelligence etc. [2].This paper provides a Review of Density-Based clustering on Data streams. Data mining is non-vital process that encompasses various technical methods like as Data summarization, Classification, Finding networks dependency, Detecting

inconsistency. Finding characteristic rules discriminate rules and association rules are among the knowledge challenges requiring spatial data. Density grid clustering algorithm relevant task is based on clustering data streams like as DUC Stream to find evolving clusters in a short amount of time and memory. This paper is used to reduce the noise, outliers and important information. Reduce these points to improve clustering efficiency [3]. In this paper, topic of Hybrid Prediction Model with missing value Imputation for identification of patient's disease and screening of diabetes, cancer and liver disorders. Multi layer Perception (MLP) and k-mean clustering are two types of hybrid prediction models. K-mean clustering is best method for handling the missing values. In data mining, missing values create various problems like as lack of efficiency, complexity in managing and break-down data. The model for multi-class unbalanced classification challenges will be evaluated and enhanced. This paper describes the creation of a hybrid prediction model with missing data imputation (HPM-MI) to address the challenge of clinical patient predictive categorization [4]. In this paper, Survey and study of the literature on how machine learning is affected by missing value imputation for incomplete data. The basis of any learnable decision-making system for automated regression or classification tasks is a dataset. In the dataset amusingness of various sources include incomplete information, collecting issues, lacking features, unavailable data. Three different procedures exist in the datasets for missing values, including Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR). Missing values imputation methods are divided into two parts such as statistical and ML based techniques, both techniques including few algorithms like as EMMVI, HDMVI, LLSMVI, LRMVI, MMVI, MICEMVI, BPCAMVI, SVDMVI, ANNMVI, KMCMVI, FCMMVI, CARTMVI, KNNMVI, and RFMVI. The top five ML models for evaluating indirect MVI are KNN, RF, SVM, BPCA, and DT. [5]. In this paper, three basic missing value imputation techniques have been evaluated. Issues resulting from lacking data: Accuracy loss because of insufficient information, Computation challenges brought caused by dataset gaps, Bias brought on by data distribution bias. Without missing data, k-means and fuzzy c-means performed similarly among the three clustering methods and generated a good Quality index of 0.85. The time required to cluster the original WDBC data using three separate algorithms are k-mean, FCM, SOM. When there are more missing values, FCM appears to perform little better bit in terms of clustering accuracy than k-means. But SOM took up a lot of time because it needed extensive initial training. SOM only around 25% of the data is used for network training [6]. This paper refers to deal with missing values, a number of conventional statistical and machine learning imputation approaches, including mean, regression, K closest neighbor, and group based. Missing data patterns and techniques is the part of research that discusses missing values. Using RMSE as an evaluation metric, KNN imputation outperformed RF imputation on the Iris data for two missingness ratios but RF performed better than the KNN on all missingness ratios. The review showed that the approaches for missing vales that are currently in use have several shortcomings. More research is required to examine the potential for novel approaches to handle missing data in big data applications in the real world [7]. This paper provides Imputation of missing values explained in a respectful way. When there are missing data, complete and available case analyses produce valid but inefficient results. When missing data are MCAR, let alone

MAR, other commonly used methods to handle missing data, such as overall mean imputation and the missing-indicator approach, produce biased findings. Advanced imputation techniques are relatively easy to use allow the use of standard analysis software, and base imputations on other well-known subject characteristics. In cases when missing data are MCAR or MAR, the multiple imputation technique gives objective results with accurate standard errors [8]. The goal of this paper, Investigation of the KNN and decision tree algorithms for intrusion detection systems. IDS are vital for stopping new attacks that enhance the system's vulnerability. IDS can be divided into two groups: anomaly-based and signature-based, depending on the detection method utilized. IDS come in three types depending on where the attack is occurring including Hot based, Network based and Hybrid based. In this paper, decision trees and KNN will be studied and univariate feature selection and ANOVA will be used. In Data Pre-processing the training and testing dataset is separated into 4 sections based on the type of attack after one hot encoding such as DoS, U2R, R2L, or probe attack. One of the most important processes in data preprocessing for machine learning is feature selection. The filter and wrapper groups make up the feature selection method. Filter looks for similarities between attributes and classes and Wrapper assesses the concerned attributes. The Decision tree algorithm gives a better result with an accuracy of 99.15%. As a result, the research indicates that Decision Tree delivers better outcomes overall than KNN [10]. This paper focuses on a review of the most popular classification methods used in data mining. The comparative study between various algorithms (K-NN classifier, Bayesian network and Decision tree) is used to show the efficiency and accuracy of each categorization algorithm in terms of performance effectiveness and time complexity. The results of using WEKA on the same dataset revealed that Decision Tree performs better than Bayesian classification, which has a similar accuracy to Decision Tree but that K-NN and other predictive methods do not perform well. Decision trees provide knowledge in the form of [IF-THEN] rules, which are simpler for people to understand. A specific algorithm can be chosen depending on the application and requirements [12]. This paper categorized several methods, including Decision Tree, Nearest Neighbor, Bayesian Network, and Support Vector Machine that are employed in a variety of fields (SVM). Decision trees and Support vector machines typically have distinct operating features. However, decision trees and rule classifiers share a similar operating character. The data set will be classified using a combination of several methods. This paper presents a brief overview of the many classification methods used in various data mining applications. One classification method is always more useful than another in any given field. Various classification techniques are presented in this work. One of the following approaches can be chosen based on the needed application requirements [16]. This paper suggested method uses correlation technique to address missing data before applying a genetic algorithm to a support vector machine. The proposed strategy has a greater percentage accuracy of 2% accuracy compared to existing methods, according to the results of a comparison between it and an existing neural network approach using mean identification rate [24]. In this article, numerous data mining decision tree methods are examined. On the basis of their accuracy and the attribute selection metric employed, several decision tree algorithms can be evaluated for their effectiveness. When dealing with missing values, C4.5 and CART are superior to ID3 because Data that is

absent or noisy cannot be managed by ID3. As an example, the ID3 algorithm employs information gain, the C4.5 method utilizes gain ratio, and the CART algorithm uses the GINI Index as the attribute selection measure. The paper also provides an overview of the attribute selection measures utilized by several decision tree algorithms. These decision tree induction procedures should be utilized at various times depending on the circumstance [28].

Statistical Methods for Missing Values:

List wise Deletion:

Each case that has one or more missing values is eliminated using list-wise deletion. When there are a lot of discarded cases, list-wise deletion may also lead to the loss of some critical data [7]. The most common technique for addressing missing data is list wise deletion, which has taken over as the analysis option of choice in the majority of statistical software pieces of software. List wise deletion, however, is not the best course of action when there is a small sample size or the MCAR condition is not met. The most common approach to dealing with missing data is to simply ignore the cases where the data is absent and analyses the cases where the data is there. Complete case analysis or list wise deletions are two terms used to describe this strategy [17].

Pair wise Deletion:

Only when the specific data point required to test a specific assumption is lacking can pair wise deletion remove information. Given that pair wise deletion is less biased for MCAR or MAR data, the relevant methods are included as factors. But if there are a lot of missing observations, the analysis will be inaccurate [17]. Pair wise deletion is another name for the accessible case procedure. Due to two factors, pair wise deletion is rarely used in education research. The first thing to note is that models with just one or two variables, like as one sample t tests, independent samples t tests, and one-way ANOVA, both list wise deletion and pair wise deletion produce the same results. Second, many general statistical software packages used by education researchers do not provide pair wise deletion for techniques requiring more than two variables, like two-way ANOVA and multiple regression [19].

Multiple Imputations:

Each of the many imputations results in a somewhat different solution. These m replies would be evidence that the imputation was accurate if they were remarkably near [9]. Starting with the data from the other variables that are already present, this technique predicts the missing data. The imputed data set, which is a complete data set, is then constructed by replacing the missing values with the projected values. It has been demonstrated that multiple imputation can generate statistical inference that is legitimate [17]. In order to produce a set of three or more simultaneous finished data sets, the researcher develops a number of potential values for each missing observation in the data. The creation of parallel finished data sets and the computing of multiply-imputed estimates are the two steps in the multiple imputation process. The main task in multiple imputations is to produce potential values for each missing data [18].

Maximum Likelihood:

When maximum likelihood is suggested, it is believed that the observed data came from a multivariate normal likelihood function to a linear model. This strategy has a drawback in that it needs different software, which can be difficult and time-consuming [1]. When managing missing values, advanced Mplus users have access to a vast array of choices. Utilizing complete information maximum likelihood estimate, this Mplus implementation is designed for missing

variables that are MAR [9]. The greatest likelihood method can be used with a variety of approaches to deal with missing data. The statistics describing the associations between the variables may be calculated using the maximum likelihood method when there are missing but generally full data. In other words, it is possible to estimate the missing data by utilizing the conditional probability of other factors [17].

Expectation-maximization (EM):

The Expectation maximization algorithm is another method for addressing missing data that is based on maximum probability. The EM algorithm has several appealing features. The missing mechanism must be ignorable for an EM estimator to be impartial and effective. The EM algorithm is also straightforward, user-friendly, and reliable. The MI approach is less effective than the EM algorithm and other missing data techniques based on the actual information log likelihood, like FIML. Furthermore, EM is model-specific [20]. When there are gaps in the data, the EM algorithm is a general approach for getting ML estimates. The EM algorithm consists of the two phases listed below: The step of expectation (E) and the step of maximizing (M). In order to deal with unknown variables, this framework is created and it is suitable for handling missing data. Use the **non-missing** variables per observation to calculate the ML estimate for the missing value [21].

HANDLING MISSING DATA USING MACHINE LEARNING:

K-Nearest Neighbor (KNN):

This approach is used to organize cases that are related. Additionally known as the lazy learner since it lacks a learning phase. For the KNN method to work best, feature scaling is also necessary [10]. As shown in fig 2, Finding the nearest, most comparable neighbors involves decreasing a distance function. The distance measure is an important component of the K-nn imputation approach. A popular and dependable distance function is the heterogeneous Euclidean overlap metric. It is trustworthy in terms of the variety and amount of missing data [11]. An easy-to-use algorithm that produces excellent results is k-Nearest Neighbor. It uses an instance-based learning approach that is delayed and nonparametric. The distance between the data points closest to the objects is determined using the Euclidean, Hamming and Murkowski distances, where k is the number of neighbors, which is typically odd. High accuracy is achieved in the k-result NN's prediction [22]. The K-nearest neighbors of a data sample X are scanned in order to classify it, and then X is given the class label that the large majority of its neighbors fall under. The term "closest neighbor categorization" is used if K=1.

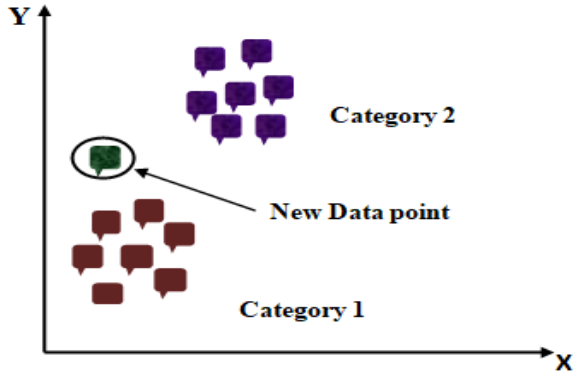

Fig.2. K- Nearest Neighbor

The K-NN classifier operates as follows:

1. Initialize value of K.
2. Determine the separation between the input sample and the practice samples.
3. Arrange the distances.
4. Select the top K-nearest neighbors.
5. Use the simple majority.
6. For the input sample, predict a class label with more neighbors [12].

Decision Tree:

In a decision tree, each node represents a feature, each link represents a choice and each leaf represents the result. It is very easy to take the data and produce some accurate interpretations because decision trees mirror human level thinking. Decision trees are straightforward because they are similar to how people make decisions. As shown in fig 3, a decision tree structure is composed of the root, internal and leaf nodes. It can solve problems with either discrete or continuous data. One of the best features of Decision Tree is its transparency. In decision trees variable screening and feature section are sufficient [26]. In data mining, decision tree models are widely used to analyze data and infer the tree and its rules that will be used to create predictions. The main benefit of using decision trees is the ability to visualize data by class. This representation is valuable because it allows people to quickly grasp the general structure of data in terms of which property most impacts the class. The goal is usually to determine the best decision tree by reducing the prediction error. WEKA employs a C4.5 method implementation known as J48, which has been employed in all of our studies [27]. The most widely used attribute selection metrics in decision trees are entropy (information gain), gain ratio, and gini index. First, A measure of the degree of uncertainty surrounding a random variable is called entropy. Information gain is used by ID3 as a criterion for choosing attributes. Information Gain is the variation between the old and new information gain requirements. Second, Gain Ratio is distinct from information gain, which evaluates information in relation to a categorization that is achieved based on some partitioning. Third, The CART decision tree method uses the Gini Index as a tool for selecting attributes [28].

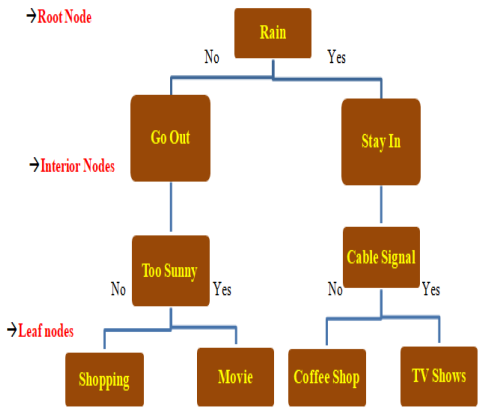

Fig.3. Decision Tree

Support Vector Machine (SVM):

The SVM is a kernel-based algorithm that makes good use of statistical learning and structural risk - reducing techniques. The primary SVR hyper parameter that must be chosen or tuned prior to running the SVR models is described by the kernel function. The SVR approach's foundation is the creation of a regression optimization algorithm based on a set of support vectors derived from training data [23]. An early application of the supervised learning method known as the support vector machine was in the two-class classification issue. It is also possible to enhance the parameters by using the kernel functions. The SVM searches for the best separating hyper-plane from a labeled training sample, maximizing the distance to the closest data points, as shown in fig 4. A hyper plane is a plane that serves as a line that is crucial in splitting the plane into two halves one for each class. SVM is used in image segmentation as well. The SVM makes it possible to recognize handwritten language as well as classify proteins in biological science. A few drawbacks apply to SVM as well. On data that are stable, the SVM algorithm does not estimate probability. Complete labeling of the given data is required [24]. The current study looks at different ways that data sets with missing values can be handled by an SVM. A least squares version of the support

vector machine is called the least squares support vector machine (LS-SVM). In this method, the estimated value of the missing value is obtained by solving a convex quadratic programming (QP) problem for traditional SVMs [25].

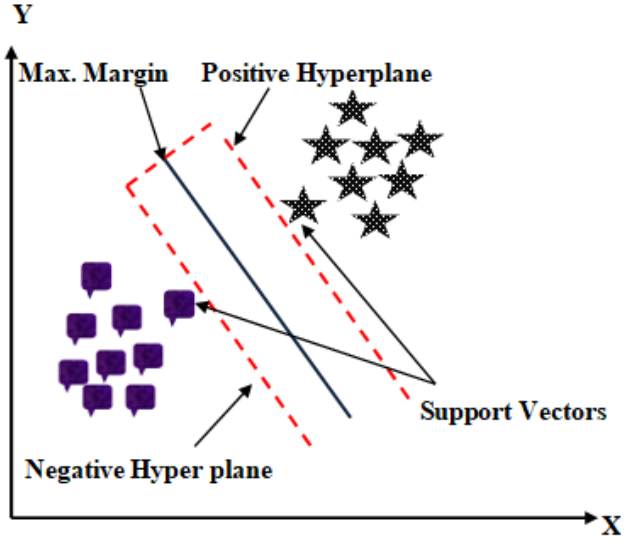

Fig.4. Support Vector Machine

Naive Bayes Classification:

Bayesian networks are a further machine learning approach utilized for data impute. The number of issues that Bayesian networks are used to solve increasing. It requires network learning and adequate discretization of all data [1]. S is a directed acyclic graph in this graphical model (DAG). As shown in fig 5, Bayes' Theorem serves as the foundation for the Naive Bayes Classifier. Theorem is, P (A/B) = (B/A). P (A)/P (B) [16]. Naive Bayesian Classifier's strengths include its capacity to anticipate predictor variables regardless of how those impacts will affect the classification. Acceptance of any large number of categories or continuous factors, cutting of high-dimensional space to one-dimensional basic density estimate, faster training and classification and lack of compassion to unhelpful variables to determine the data's numerical probability. Each parameter's mean and standard deviation must be determined [29]. The resilience of the Naive Bayes classification method is one of its key characteristics. Dependencies and missing values are common in practical data sets, yet the results indicate that using Naive bayes has no negative effects on performance. More research in the medical field, including studies on breast cancer, heart disease, thyroid, and liver disease, shows that Naive Bayes is the greatest model in comparison to its competitors. The Naïve Bayes performs remarkably well in other applications such as email spam filtering, even though the chosen features are not independent [30].

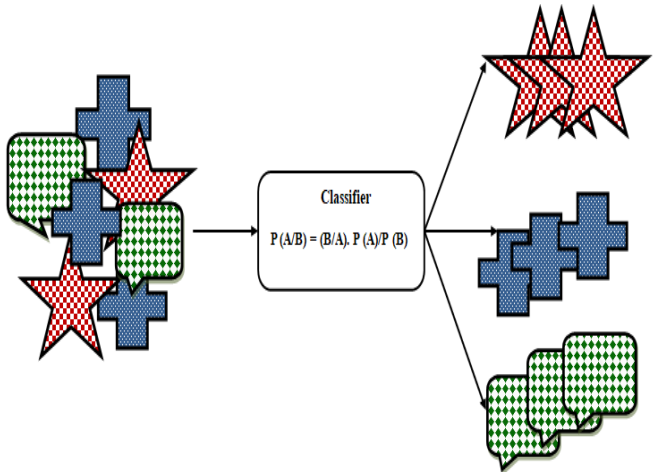

Fig.5. Naïve Bayes

Comparison Table of Classification techniques:

Algorithms	Advantages	Disadvantages
K-Nearest Neighbor(K-NN) Classifier	1. Easy to understand and implement. 2. Robust to noisy training data. 3. Training is very fast.	1. Testing is slow. 2. Expensive for large dataset. 3. Memory limitation. 4. Hard to interpret the result.

	4. Zero cost in the learning process.	
Support Vector Machine(SVM)	1. Produce very accurate classifiers. 2. Less over-fitting, robust to noise. 3. Memory intensive. 4. Easy to interpret results.	1. Sensitive to running kernel choice. 2. Very black box. 3. Computationally Expensive. 4. Run very slowly.
Decision Trees	1. Performs well with large databases. 2. Can handle missing values. 3. Handle numerical and categorical data. 4. A single tree is highly interpretable.	1. Restricted to one output attribute. 2. Less predictive in many situations. 3. Generates categorical output. 4. Unstable classifier.
Naïve Bayes	1. Can handle multiple classes. 2. Can deal with noisy and missing data. 3. Makes computational process easier. 4. Better speed and accuracy for huge datasets.	1. Sensitive to how input data is prepared. 2. Lower performance. 3. Required large data for good result. 4. Loss of accuracy.

CONCLUSION:

The main goal of data mining is to extract important information from large amounts of raw data and transform it into a form that can be used effectively and efficiently. This paper focuses on a review of several classification strategies, including their features and drawbacks. Applications for classification algorithms include customer targeting, medical disease diagnosis, social network analysis, credit card rating, Artificial intelligence and document categorization, among many more. Several major kinds of classification techniques are K-Nearest Neighbor classifier, Naive Bayes, Decision Trees and Support Vector Machine. It is true that the different machine learning algorithms have their own unique strengths. For example, SVM is good for handling missing data, KNN is easy to understand and implement, Decision Tree is good for dealing with irrelevant features and Naïve Bayes is good for handling multiple classes. Their performance is also affected by the type, size and quality of the data. Comparison Table shows that the Decision Tree outperforms and Bayesian classification having the same accuracy as of Decision Tree but other predictive methods like K-NN and SVM does not giving good results.

REFERENCE:

Agarwal Vivek, Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis, International Journal of Computer Applications, 131(4):30–36, 2015.

Jiawei Han, Micheline Kamber, Jian Pei Data Mining Concepts and Techniques, Third Edition.

Pragati Shrivastava, Hitesh Gupta, "A Review of Density-Based clustering in Spatial Data," IJACR, vol. 2, pp. 200-202, September 2012.

Purwar A, Singh SK. Hybrid prediction model with missing value imputation for medical data. Expert Syst Appl 2015; 42:5621–31.

Hasan, M.K., Alam, M.A., Roy, S., Dutta, A., Jawad, M.T. and Das, S., 2021. Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). Informatics in Medicine Unlocked, 27, p.100799.

Somasundaram, R.S. and Nedunchezhian, R., 2011. Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. International Journal of Computer Applications, 21(10), pp.14-19.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B. and Tabona, O., 2021. A survey on missing data in machine learning. Journal of Big Data, 8(1), pp.1-37.

Donders, A.R.T., Van Der Heijden, G.J., Stijnen, T. and Moons, K.G., 2006. A gentle introduction to imputation of missing values. Journal of clinical epidemiology, 59(10), pp.1087-1091.

Acock, A.C., 2005. Working with missing values. Journal of Marriage and family, 67(4), pp.1012-1028.

Pathak, A. and Pathak, S., 2020. Study on decision tree and KNN algorithm for intrusion detection system. International Journal of Engineering Research & Technology, 9(5), pp.376-381.

García-Laencina, P.J., Sancho-Gómez, J.L. and Figueiras-Vidal, A.R., 2010. Pattern classification with missing data: a review. Neural Computing and Applications, 19(2), pp.263-282.

Jadhav, S.D. and Channe, H.P., 2016. Comparative study of K-NN, naive Bayes and decision tree classification techniques. International Journal of Science and Research (IJSR), 5(1), pp.1842-1845.

Widyananda, w., purnomo, m.f.e., aswin, m., mudjirahardjo, p. And pramono, s.h., 2022. Dataset missing value handling and classification using decision tree c5. 0 and k-nn imputation: study case car evaluation dataset. Journal of theoretical and Applied Information Technology, 100(12).

Saar-Tsechansky, M. and Provost, F., 2007. Handling missing values when applying classification models.

Rahman, M.M. and Davis, D.N., 2013. Machine learning-based missing value imputation method for clinical datasets. In IAENG transactions on engineering technologies (pp. 245-257). Springer, Dordrecht.

Sharma, S., Agrawal, J., Agarwal, S. and Sharma, S., 2013, December. Machine learning techniques for data mining: A survey. In 2013 IEEE international conference on computational intelligence and computing research (pp. 1-6). IEEE.

Kang, H., 2013. The prevention and handling of the missing data. Korean journal of anesthesiology, 64(5), pp.402-406.

Pigott, T.D., 2001. A review of methods for missing data. Educational research and evaluation, 7(4), pp.353-383.

Cheema, J.R., 2014. A review of missing data handling methods in education research. Review of Educational Research, 84(4), pp.487-508.

Dong, Y. and Peng, C.Y.J., 2013. Principled missing data methods for researchers. SpringerPlus, 2, pp.1-17.

Lin, T.H., 2010. A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. Quality & quantity, 44, pp.277-287.

Kaur, H. and Kumari, V., 2022. Predictive modelling and analytics for diabetes using a machine learning approach. Applied computing and informatics, 18(1/2), pp.90-100.

Essam, Y., Huang, Y.F., Ng, J.L., Birima, A.H., Ahmed, A.N. and El-Shafie, A., 2022. Predicting streamflow in Peninsular Malaysia using support vector machine and deep learning algorithms. Scientific Reports, 12(1), p.3883.

Alhroob, A., Alzyadat, W., Almukahel, I. and Altarawneh, H., 2020. Missing data prediction using correlation genetic algorithm and SVM approach. International Journal of Advanced Computer Science and Applications, 11(2).

Sivapriya, T.R., Kamal, A.N.B. and Thavavel, V., 2012. Imputation and

classification of missing data using least square support vector machines–a new approach in dementia diagnosis. Int. J. Adv. Res. Artif. Intell, 1(4), pp.29-33.

Patel, H.H. and Prajapati, P., 2018. Study and analysis of decision tree based classification algorithms. International Journal of Computer Sciences and Engineering, 6(10), pp.74-78.

Rahman, M.M. and Davis, D.N., 2013. Machine learning-based missing value imputation method for clinical datasets. In IAENG Transactions on Engineering Technologies: Special Volume of the World Congress on Engineering 2012 (pp. 245-257). Springer Netherlands.

Gupta, B., Rawat, A., Jain, A., Arora, A. and Dhami, N., 2017. Analysis of various decision tree algorithms for classification in data mining. International Journal of Computer Applications, 163(8), pp.15-19.

Yudianto, M.R.A., Agustin, T., James, R.M., Rahma, F.I., Rahim, A. and Utami, E., 2021. Rainfall Forecasting to Recommend Crops Varieties Using Moving Average and Naive Bayes Methods. International Journal of Modern Education & Computer Science, 13(3).

Wickramasinghe, I. and Kalutarage, H., 2021. Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Computing, 25(3), pp.2277-2293.