https://dergipark.org.tr/tr/pub/ataunivbd
2019; 14(2): 209-220

Review/Derleme
DOI: 10.17094/ataunivbd.462197

Atatürk Üniversitesi Veteriner Bilimleri Dergisi
Atatürk University Journal of Veterinary Sciences

A Review on Determination of Computer Aid Diagnosis and/or Risk Factors Using Data Mining Methods in Veterinary Field

Pınar CİHAN[1a]✉, Erhan GÖKÇE[2b], Oya KALIPSIZ[3c]

1. Tekirdağ Namık Kemal University, Çorlu Faculty of Engineering, Department of Computer Engineering, Tekirdağ, TURKEY.
2. Kafkas University, Faculty of Veterinary Medicine, Department of Internal Medicine, Kars, TURKEY.
3. Yıldız Technical University, Faculty of Electrical Electronic, Department of Computer Engineering, Istanbul, TURKEY.
ORCID: 0000-0001-7958-7251[a], 0000-0003-2674-1010[b], 0000-0001-9553-669X[c]

Geliş Tarihi/Received	Kabul Tarihi/Accepted	Yayın Tarihi/Published
21.09.2018	28.03.2019	25.10.2019

Bu makaleye atıfta bulunmak için/To cite this article:
Cihan P, Gökçe E, Kalipsiz O: A Review on Determination of Computer Aid Diagnosis and/or Risk Factors Using Data Mining Methods in Veterinary Field. Atatürk Üniversitesi Vet. Bil. Derg.,14(2): 209-220, 2019. DOI: 10.17094/ataunivbd.462197

Abstract: Data mining is an interdisciplinary field. In this field, statistical analysis techniques and artificial intelligence algorithms are used. Thanks to the algorithms used, hidden information within the data is revealed and transformed into qualified information. Data mining techniques have been used effectively in health, engineering, biomedicine and many other fields for many years, and are known to contribute significantly to health sciences. However, the use of this effective method in animal health is very limited and the use of data mining methods in animal health has accelerated in recent years. Animal illness and mortality cause decrease husbandry and this impact negatively on the national economy. Interdisciplinary studies are very important to increase profitability in this field. In this review, studies on the determination of animal diseases and/or risk factors using data mining methods have been examined. Studies in the veterinary field shows the data mining methods can be successfully applied in this field.

Anahtar Kelimeler: Computer aid diagnosis, Data mining, Veterinary.

Veterinerlik Alanında Veri Madenciliği Yöntemleri Kullanılarak Bilgisayar Destekli Tanı ve/veya Risk Faktörlerinin Belirlenmesi Üzerine Bir İnceleme

Öz: Veri madenciliği disiplinler arası bir alandır. Bu alanda istatistiksel analiz teknikleri ve yapay zeka algoritmaları kullanılmaktadır. Kullanılan algoritmalar sayesinde verideki gizli bilgiler açığa çıkarılarak ve nitelikli bilgiye dönüştürülmektedir. Veri madenciliği teknikleri sağlık, mühendislik, biyomedikal ve diğer birçok alanda uzun yıllardır etkin bir şekilde kullanılmaktadır ve sağlık bilimlerine önemli ölçüde katkıda bulunduğu bilinmektedir. Ancak bu etkili yöntemin hayvan sağlığında kullanımı oldukça sınırlı olup son yıllarda hayvan sağlığında veri madenciliği yöntemlerinin kullanımı hızlanmıştır. Hayvan hastalıkları ve ölüm oranları hayvancılığın azalmasına neden olmakta ve bu da ülke ekonomisini olumsuz yönde etkilemektedir. Bu alandaki karlılığı artırmak için disiplinler arası çalışmalar çok önemlidir. Bu derlemede, veri madenciliği yöntemlerini kullanarak hayvan hastalıklarının ve / veya risk faktörlerinin belirlenmesine yönelik çalışmalar incelenmiştir. Veterinerlik alanındaki çalışmalar, veri madenciliği yöntemlerinin bu alanda başarıyla uygulanabileceğini göstermektedir.

Keywords: Bilgisayar destekli tanı, Veri madenciliği, Veterinerlik.

✉Pınar Cihan
Tekirdağ Namık Kemal University, Çorlu Faculty of Engineering, Department of Computer Engineering, Tekirdağ, TURKEY.
e-mail: pkaya@nku.edu.tr

INTRODUCTION

Data mining is classified as an interdisciplinary subfield under computer science. It involves discovering patterns through computation in large data sets combining methods used in several disciplines such as machine learning, artificial intelligence, database systems and statistics. The general objective of the data mining process is to reveal hidden predictive information and unknown data, relationships, patterns and knowledge by searching through large data sets where it is hard to search and determine using classical methods of statistics (1–3).

The health organizations generate the data so vast and complex that it is very difficult to analyze the data in order to make important decisions regarding patients' health (4). This huge bulk of data contains details about hospitals, patients, medical claims, treatment costs, etc. Therefore, the generation of a powerful tool to analyze and extract important information from this complex bulk of data is required. In order to increase their capability of making decisions on patients' health, health-care organizations use data mining techniques such as classification, clustering, regression and association.

Data mining methods are being widely used as a supporting system for diagnosis and treatment. However, it is obviously seen that data mining methods are not used sufficiently in animal health. To be more precise, the newly developed scientific solutions are not being implemented when encountering a problem in the field of veterinary. It can be said that interdisciplinary studies are inadequate (5).

Animal health is as important as human health. The data mining methods are being quite newly applied in the field of veterinary. In this review, studies in the field of veterinary will be examined under data mining models (detailed in Section 3). Classification methods were used for factors effecting fertility in Japanese quail eggs (6,7), evaluation of raw milk quality (8,9), oestrus detection (10,11), estimation of heat stress (12), estimation of milk production efficiency (13), prediction of the nutrient content in dairy manure (14), prediction diagnosis (15). Clustering methods were used for examination of goat in terms of mohair characteristics (16), classification of Holstein Friesian breed dairy cattle in terms of some milk component parameters (17), classifying the cities for sheep production (18), classify sheep according to body measurements (19), risk classification (20,21), determination of factors responsible for animal mortality (22), characterization disease (23).

The aim of this study is to allow obtaining better results in diagnosis and treatment by applying data mining methods in animal health and to introduce a different point of view to the veterinarian in terms of decision-making processes. Therefore, studies towards determination of disease diagnosis and/or risk factors using data mining methods in the field of veterinary were reviewed and the quite recent use of data mining in the field of livestock farming was discussed in detail. Before starting an application associated with data mining, it is crucial to understand the data mining related basic concepts and utilized methods in order to obtain more successful results.

1. Data Mining

Data mining comprises three steps of knowledge discovery in databases (KDD); pre-processing, modeling and post-processing of data (24). The goal of pre-processing is to prepare raw data for mining. The relationships between various data are discovered in the modeling step, in order to extract a pattern. In the post-processing step, the extracted pattern is evaluated in order to be verified and stated as knowledge. For the interactive and iterative KDD process model shown in Figure 1.

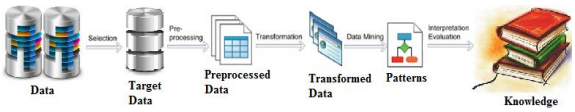

Figure 1. Summarized KDD process steps according to Fayyad, Piatetsky et al. (24,25).
Şekil 1. Fayyad, Piatetsky ve ark. göre KDD işlem adımları özeti (24,25).

1.1. Data Preprocessing

Databases may contain incomplete raw data, noise, missing values, and inconsistent data. Data mining results are affected by the quality of the data. Raw data is preprocessed, to improve the quality of the data and consequently the mining results. This increases efficiency and it eases the mining process. Data preparation and obtaining a general overview of data is aimed at the preprocessing step. In practice, 60%-90% of project time is spent on data understanding and preparation (26). So, the data preprocessing step is very important.

Data cleaning: Data becomes dirty by missing values, noisy and inconsistent data. Within the mining procedure, these dirty data may confuse the process. If some data cells are empty, this means that there are missing values. As a random part of the error, noise should be eliminated. Outlier data are those with different behavior from other data and they are to be detected. Operations such as filling in missing values, identifying or removing outliers, smoothing noisy data and resolving inconsistencies are required, in order to solve these problems. Solution methods are elimination, estimation and ignorance (3).

Data integration: Data integration comprises integration of multiple databases, files or data cubes or in order to create a single data set (3).

Data transformation: Transformation of data into forms to render them appropriate for mining. The methods of data transformation include normalization (scaled to fall within a small, z score normalization, min-max normalization and decimal scaling), smoothing (removing noise from data by binning, regression and clustering), aggregation (summarization or aggregation may be applied to data) and generalization (concept hierarchy climbing) (3).

Data reduction: This is to obtain a reduced representation of the data set, making it much smaller in volume with the same (or approximately the same) analytical results. In literature, several methods such as data cube aggregation, dimensionality reduction, data compression, numerosity reduction, discretization and concept hierarchy generation were used for data reduction (3).

Discretization and binarization: This comprises the preparation of data for classification and association algorithms. Discretization is the process of transforming a continuous attribute into a categorical attribute. It is commonly used in classification. Through binarization, a continuous or categorical attribute is mapped into one or more binary variables. Generally, it is used for association analysis (27).

1.2. Data Modeling

In the data modeling step, all tasks may be divided into predictive and descriptive categories. Predictive algorithms comprise classification and regression according to the type of target variable (discrete or continuous). Descriptive algorithms are divided into clustering and association rule mining categories (28).

Classification is the estimation of data that does not have a certain class using existing data. Classification algorithms are supervised algorithms and consist of two steps. In the first step the algorithm determines the relationships between the class information and other attributes and it establishes a model to implement it on the whole data. In the second step, this model is operated on unclassified data and the classes of these data are estimated. To measure the success of the generated model of the classification algorithm, specific criteria are used which are accuracy, error rate and sensitivity (1,3,28).

Regression, unlike classification, is used for estimation of continuous data and it is a data analysis method establishing models to indicate important data classes that estimate future data. Each of the models used for estimations is mining functions and they estimate the probable values of classes. (1, 3, 28).

Clustering sorts data into clusters consisting of similar elements and it obtains homogeneous sub-data groups from a heterogeneous data cluster. This process is done completely utilizing the similarity without any learning process (1,3,28).

Association rule mining is used to discover relationships between large transactional data for exploring sequential data. Through this method, analysis of association rules or frequently appearing items is obtained. Various criteria are taken into consideration such as support and confidence through performance evaluation on discovered rules (29).

1.3. Data Post-processing

The extracted knowledge is visualized and evaluated in this step. Visualization is known as the essential matter in data mining. It is considered as an advantage for an algorithm. Various predictive and descriptive algorithms are used to extract knowledge from raw data. Conclusively, the performance evaluation methods comprise single scalar and graphical methods. In the first group, accuracy, sensitivity and specificity are categorized. This group is simple to be practiced but not efficient in covering various aspects of evaluation. The other group comprises Receiver Operating Characteristic (ROC) Curve, Cost-Line and Lift (1,3).

Data mining reflects a process that comprises several steps, methods and algorithms. In order to generate successful results through the study, methods and algorithms should be selected according to the data set.

2. Literature Review

Automated animal diagnosis helps the veterinary surgeon to calculate the correct disease with less time. While the articles scanning we selected keywords: “analysis of animal diagnosis”, “predict animal diagnosis using decision tree”, “diagnosis of disease veterinary”, “evaluation of risk factor for animal diseases”, “animal medical data analysis”, “disease decision support animal”, “predict animal diagnosis data mining”, “analysis small ruminant data mining”, “naive bayes in veterinary”, “data mining predict animal disease”, “weka in veterinary”, etc. and their combinations are used. The studies highlight the foremost objectives of the authors working in the field of predicting animal disease(s) and/or determination of risk factors in the veterinary field using data mining methodology.

2.1. Clustering Studies

Hermann-Bank et al. (23), aimed to investigate if the new neonatal porcine diarrhea (NNPD) is associated with the composition of gut microbiota. K-means, principle component analysis (PCA) and linear discriminant analysis (LDA) methods are used. In the conducted study, analyses were performed on a data set consisting of 50 control (without NNPD) and 52 NNPD piglets. According to LDA; 70% of case samples (diarrhea) and 83% of control samples (without diarrhea) were classified correctly. When the microbial composition in neonatal piglets of different ages (three to seven days old) was examined, case and control groups were observed to be better distinguished with age. According to PCA, new neonatal porcine diarrhea (NNPD) was observed to be associated with the bacterial composition and a big variation was observed among diarrhoeic piglets. The study concludes that NNPD indicates that bacteria could be the etiology of this diarrhea which affects piglets during the first week of life.

2.2. Classification/Regression Studies

Yazdanbakhsh et al. (30), aimed to develop an intelligent system for prediction of bovine respiratory diseases in livestock farming. Chi-squared automatic interaction detection (CHAID), logistic regression (LogR) and analysis of variance (ANOVA)

methods were used. The analyses were performed on sensor data in terms of body and environment temperature obtained from sensors attached to cows. At the prediction of disease in cows for the next 7 days logistic regression was observed to have the lowest false-positive rate (sensitivity 42.3%, specificity 92.7%), MetaRBFN1 was observed to have the lowest false-negative rate (sensitivity 92.3%, specificity 63%), and MetaRBFN2 was observed to deliver the nearest point to the optimal classifier within the ROC area (sensitivity 73.1%, specificity 78.8%). The study concluded reporting that the best performance was achieved by the ensemble classifier (the voting method combining MetaRBFN1, MetaRBFN2, and Logistic Regression) with a sensitivity of 80.8% and a specificity of 80%.

Saidani et al. (31), aimed to demonstrate the effect of climate on warble infestation. Radial basis function network (RBFN), support vector machine (SVM), sequential minimal optimization (SMO), decision tree (DT) and LogR methods are used. The prevalence and infestation intensity (38.23%; 21.57±11.98) were found significantly higher than humid areas (20.74%; 14.84±7.86). Through the CHAID algorithm, the main factors were observed as climate followed by the husbandry system and breed. Through logistic regression and multivariate ANOVA, intrinsic factors (age, sex, breed) and extrinsic factors (husbandry system, treatment), as well as climate, were observed to be associated with both prevalence and infestation intensity. The study concludes that semi-arid areas are more favorable than humid ones at the free stages of Hypoderma spp. life cycle (pupae and adult flies).

Awaysheh et al. (32), aimed to develop models about the influence of inflammatory bowel (IBD) and alimentary lymphoma (ALA) diseases on serum chemistry and to distinguish these diseases by means of classification algorithms. Naïve bayes (NB), DT and artificial neural network (ANN) methods are used. The study was conducted on results of complete blood count (CBC) and serum chemistry (SC) of 120 cats consisting of 40 normal cats, 40 cats with IBD and 40 cats with ALA. With sensitivities of 70.8% and 69.2% respectively, naive bayes and artificial neural networks were observed to have a higher performance in comparison to the decision tree with a sensitivity of 62%. By classification according to normal, IBD and ALA status, the area under the ROC curve was observed to be 83% for NB, 79% for DT and 82% for ANN. By classification of cats in 3 categories, NB with 10 variables and ANN with 4 variables have shown a better performance than the decision tree with 5 variables. The study concluded reporting that both NB and ANN provide a good algorithm alternative to create prediction models in this field.

Boujenane et al. (33), aimed to analyze the rate of incidence and risk factors of clinical mastitis. LogR and gamma regression models (GRM) methods are used. The analyses were performed on 1725 Holstein cows. According to logistic regression; the obtained result showed that cows with a parity of 2 had a mastitis risk of 65% and those with parity 3 and 4 had a risk of 88% and 115% respectively and that the risk was higher than the parity 1. The highest mastitis risk was observed in cows that calved from October to January and the highest number of mastitis incidences were observed from July to September. According to gamma regression parameter estimates, the onset time of mastitis 1 and 2 in cows that calved from January to October were 16.3 and 8.5 days, respectively.

Amrine et al. (34), aimed comparison of classification algorithms to predict outcomes of feedlot cattle in Bovine respiratory disease. DT, bayesian network (BN), meta-classifiers, ANN and LogR methods are used. In this study, the analyses were performed on over one million animals. Different classification algorithms in estimating post-treatment results of individual calves based on data obtained from the first diagnosis and treatment of the disease were compared in this study. Classifier performances were observed to range from 63% to 95% according to the dataset. The study concluded reporting that feedlot managers might perform correct predictions by pairing the correct classifier with the available data.

Urbach et al. (35), aimed automated phenotyping and advanced data mining exemplified in rats transgenic for Huntington's disease. Multivariate analysis (MVA), PCA and Partial least squares discriminant analysis (PLS-DA) methods are used. Previously undetected aspects (such as circadian activity, energy metabolism, rearing) were traced for the phenotype of tgHD rats by automated systems and spontaneous free rearing rats correlated with individual performance were also traced in the accelerated test. In juvenile tgHD rats that differed from adult animals, discrimination by genotype was revealed through PCA, it was further resolved through PLS-DA to detect "temperature" (juvenile) and "rearing" (adult) as phenotypic key variables in the tgHD model. In this study, it was also observed that MVA was extremely useful in comparison to PCA in terms of identifying and scoring genotype-related traits in animals' behavior and metabolism. The study has concluded reporting that MVA combined with intra-home-cage phenotyping might characterize a complex phenotype through identifying highly sensitive and standardized novel physiological and behavioral markers while using fewer human resources.

Babcock et al. (22), aimed to assess the ability to predict the cumulative risk of bovine respiratory disease complex (BRDC). LogR and mixed negative binomial regression (MNBR) methods are used. In this study, for model training 12735, for validation 6221 and for testing 6476 cohort data set were used. Subset 1 (feedlots that reported risk code) and subset 2 (feedlots that reported daily mortality counts) were generated from the data set. Using logistic regression models, correctly classified cohort percentage per day was illustrated, then arrival weight, arrival month, and feedlot were used to evaluate the effect of day. Cattle arriving in April had the highest (77%) number of lots which were correctly classified and cattle arriving in December had the lowest one (28%). At the arrival, classification accuracy varied according to initial weight and ranged from 17% (<182 kg) to 91% (> 409 kg). When risk code was known, predictive accuracy has improved from 64% to 74% in 8 days from arrival and when risk classification was unknown the accuracy at arrival was 56% and 69% at 8 days on feed. The study concluded demonstrating the improvement in the predictive ability of models utilizing more refined data on the cohorts' history. So, these models become more useful for commercial feedlot operators.

Ahmed et al. (36), aimed to determine the influence of epidemiological factors on prevalence and infestation rate of Hypoderma spp. CHAID and ANOVA methods are used. A total of 1000 animals were analyzed and the influence of age, sex, breed, management and previous exposure factors on hypodermosis were investigated. In the study, the prevalence was shown to be higher in young and female animals. The management was found to be the most important factor in warble fly infection and it was observed that warbles were first detected in September reaching a maximum in December and disappearing in January. According to CHAID algorithm, the most effective factor in warble fly prevalence was the grazing pattern followed by district locality. Along with significantly effective factors such as previous exposure to parasite, parasitation intensity, the management system was reported to be another factor in hypodermosis prevalence.

Piwczynski et al. (39), aimed to determine risk factors responsible for lamb mortality. Classification trees (CT) and LogR methods are used. The analyses were performed on 20.044 lambs. At the study, factors such as flock, year of lambing, age of dams at lambing, body weight of dams at the age of 12 months, birth type of lambs were observed to be effective on lamb mortality at both logistic regression and classification tree methods. At the conclusion, the study reported that classification tree would be more advantageous than logistic regression by presenting the information in the form of a tree structure making it more understandable in a straightforward way.

Geenen et al. (38), aimed to discriminate between classical swine fever (CSF)-infected and non-infected pig herds. NB method is used. The analyses were performed on 245 CSF-negative herds and 245 CSF-positive herds. Naive Bayesian classifier's accuracy was observed to be 69% when all features were applied, the classifier's accuracy was observed to be 67% after the filter method was applied, the classifier's accuracy was observed to be 70% after the wrapper method was applied. Because the obtained results had a better performance than that of a previously published diagnostic rule (63%), the study reported that the NB classifier was a promising tool in modeling diagnostic problems.

Lopez et al. (39), aimed to detect and classify determinant factors on sheep protostrongylid infection. CHAID and general linear model (GLM) methods are used. The analyses were performed on 93 fecal samples collected from 74 commercial meat ovine flocks. CHAID method was used to evaluate the risk factors in sheep on being infected by lungworms and the GLM method was used for evaluating infection intensity. While Dictyocaulus filaria infection was the principal factor on protostrongylid prevalence, other factors were observed to be age, introducing external animals in the flocks, mixed management with goats and animal density in pastures. In addition, treatment effects on prevalence were only observed in flocks that did not introduce ewes. The lowest protostrongylid prevalence was observed in flocks without D. Filaria infection and when contact with goats was avoided.

Kuncheva et al. (15), aimed to reduce the false positive rate within the set of suspects. 18 classification methods (Bagging, AdaBoost, Random Forest, etc.) are used. The analyses were performed on a data set consisting of 3113 sheep. 14 clinical signs such as abnormal head position, change of temperament, nibbling biting and poor condition were recorded by veterinary officials for each animal and post-mortem diagnosis was performed. Then 18 different classification methods were used to identify healthy animals according to only clinical signs. The study reported that classifier ensembles performed lower than expected, while logistic classifier, NB and linear classifier models gave better results.

Dawson et al. (40), aimed to characterize the cytology of bronchoalveolar fluid. DT method is used. The analyses were performed on bronchoalveolar lavage fluid (BAL) collected from 113 sheep lungs consisting of 44 normal lungs, 12 lungs with maedi, 9 with adenomatosis, 27 parasitic, and 21 pneumonic lungs. According to the study, the basic problem in taking clinical decisions was reported to be whether the cytological analysis of the BAL fluid could diagnose lung diseases in sheep. Therefore, the decision tree method was used to interpret the differential cytology of the BAL fluid. According to the decision tree; it was observed that 30% of cases with cytological profiles suggesting no disease might have lung lesions which means there was a remarkable coincidence between normal BAL profiles and disease-related profiles. 37% of BAL profiles suggesting parasitism could be derived from lungs with no evidence of parasitism, but in sheep with a maedi suggesting profile only 10 percent would be misclassified. 57% of sheep profiles suggesting adenomatosis will be derived from lungs with other pathologies. Therefore, any predictions made on the basis of the differential cytology of BAL fluid are subject to considerable uncertainty.

Sandholm et al. (41), aimed to compare classifier induction algorithms on the diagnosis of equine colic and the prediction of its mortality. Linear regression (LinR), DT, K-nearest neighbors (KNN) and LogR methods are used. The study reported that the diagnosis of the disease was easy for all methods and the average accuracy of the methods differed between 94.7% and 99.3%. Small differences were observed between the accuracies of the methods and logistic regression and neural networks methods had the highest accuracies. The study reported that mortality prediction was difficult for all methods and the average accuracy differed between 62% and 72%. Neural networks and model class selection methods were observed to have the highest accuracy. In

addition, reducing the number of features was reported to lower the accuracy of logistic regression and disease classification, while increasing mortality prediction from 65% to 73%.

Table 1 presents the comparative analysis of different data mining techniques and algorithms which have been used in examined studies.

Table 1. Comparison of used data mining techniques.
Tablo 1. Kullanılan veri madenciliği tekniklerinin karşılaştırılması.

Author	Tool	Disease	Data Size	Data Mining Techniques								Successful method
				Classification/Regression						Clustering		
				DT	ANN	Bayesian	SVM	KNN	Other	Partition	Hierarchical	
Yazdanbakhsh et al. (30)	Weka, Matlab	Bovine Respiratory	31	√			√		RBFN, LogR			RBFN
Saidani et al. (31)	R, SPSS	Hypodermosis epid.	1.635	√					LogR, Anova			-
Awaysheh et al. (32)	Weka	IBD and ALA	120	√	√	√						DT
Bank et al. (23)	R	Bacterial gut	102							√		-
Boujenane et al. (33)	SAS	Mastitis	1.725						LogR			-
Amrine et al. (34)	Weka	Bovine respiratory	468.734	√	√	√			RF, MB, LB, FC, DS, VP			-
Urbach et al. (35)	Phe-noMaster	Transgenic Huntington's	12						MVA, PCA, PLS-DA			PCA
Babcock et al. (22)	Stata	Bovine Respiratory	25.432						LogR, MNBR			MNBR
Ahmed et al. (36)	SPSS	Hypoderma spp.	1.000	√					Anova			DT
Piwczyński et al. (39)	Enterprise Miner	-	20.044	√					LogR			DT
Geenen et al. (38)	Dazzle	Swine fever	490			√						NB
López et al. (39)	SPSS	Protostrongylid infection	2.093	√					GLM			
Kuncheva et al. (15)	Weka	Scrapie	3.113	√		√	√	√	Bagging, RF, adaBoost etc.			NB
Dawson et al. (40)	Minitab	Lung diseases	113	√								DT
Sandholm et al. (41)	-	Gastrointestinal colic	105	√	√			√	LinR, LogR			ANN

ALA = Alimentary Lymphoma, ANN = Artificial Neural Network, ANOVA = Analysis of Variance, BN = Bayesian Networks, CHAID = Chi-squared Automatic Interaction Detection, CT = Classification Trees, DS = Decision Stump, DT = Decision Tree, FC = Filtered Classifier, GLM = General Linear Model, GRM = Gamma Regression Models, IBD = Influence of İnflammatory Bowel, KNN = K-Nearest Neighbors, LB = Lobitboost, LinR = Lineer Regression, LogR = Logistic Regression, MB = Multiboost, MNBR = Mixed Negative Binomial Regression, MVA = Multivariate Analysis, NB = Naïve Bayes, PCA = Principle Component Analysis, PLS-DA = Partial Least Squares Discriminant Analysis, RBFN = Radial Basis Function Network, RF = Random Forest, SMO = Sequential Minimal Optimization, SVM = Support Vector Machine, VP = Voted Perceptron.

3. Data Mining Tools

Thanks to the rapidly developing technology, many transactions executed virtually or in real-time are stored electronically. The experts use data mining methods to extract the relationships and rules which would help us forecasting the future within big data clusters, utilizing those data in large databases, in order to solve current problems, to make critical decisions or to do estimations towards the future. A computer program is required for data mining applications. There are many software applications developed in this scope. Frequently used computer programs in literature are shown in Table 2.

Table 2. Comparison of data mining tools.
Tablo 2. Veri madenciliği araçlarının karşılaştırılması.

Tool	Description	Approches
R (42)	It's a free computer program for statistical calculation and graphics and it is also a programming language. Users may add packages to expand the capabilities of R. You can use this tool after some data mining experience and with some more programming skills.	Statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.
Matlab (43)	It isn't a open-source software programming. MATLAB already supports various implementations of different stages of the data mining process, including various toolboxes created by experts in the field.	Classification, classifiers, data mining, image processing, machine learning, pattern recognition, statistics.
Weka (44)	It's a open-source and platform independent software. Weka is a computer program developed in Java language and includes machine learning algorithms.	Data preprocessing, classification, clustering, association rule extraction.
Orange (44)	It's an open-source data visualizaztion and analysis for novice and experts. Orange is a component based data mining and machine learning software süite written in the Python language.	Data management and preprocessing, classification, regression, association, ensembles, ensembles, evaluation, projections.
Rapid Miner (44)	It's an open-source system for novice and experts. RapidMiner provides an integrated environment for data mining, machine learning, text mining, business analytics and predictive analytics.	Data transformation, data preprocessing, visulization, modeling, evaluation, deployment.
Knime (44)	It's an open-source platform for novice and experts. KNIME is a user friendly and comprehensive data analytics framework.	Data acsess, data transformation, initial investigation, powerful predictive analtics, visualisation and reporting.
SPSS Climentine (45)	Clementine is a module developed by the company SPSS for data mining applications. Thus, SPSS allows utilizing statistical functions whereby providing artificial intelligence based algorithms.	Classification discovery, cluster discovery, regression discovery, association discovery, text mining, outlier discovery, data visualisation.
Enterprise miner (46)	Enterprise miner is a graphical user interface designed by the company SAS considering the requirements in data mining.	Classification, regression, two-stage models, clustering, time series, association.

CONCLUSION

Data mining algorithms are being widely used in many fields to reveal hidden information within the data. In this study, current studies conducted for disease diagnosis and/or determination of risk factors using data mining methods in the field of veterinary were examined in detail. Furthermore, information was provided related to the data mining

process and some frequently used tools that might be implemented while mining the data.

We have observed that the use of data mining methods in the veterinary field gained momentum in recent years. There are very successful algorithms in the literature as well as those that are still being developed. Moreover, we have observed that the tools were mostly open-source and they did not require any prior knowledge of programming languages.

We have seen that popular methods such as decision trees, regression and artificial neural networks were used in the related studies. Besides, the latest developed methods in data mining were not implemented in order to solve the problems in the field of veterinary.

The performance of algorithms may vary in studies conducted on different data sets. The most important reason is the fact that it depends on the data source, the preprocessing on data and the selection of algorithm parameters. In studies, the researchers may use successfully tested algorithms and they may also use any algorithms which are proper for the dataset. These algorithms may produce different performances depending on the dataset.

We conclude that data mining which has been already implemented intensely in the fields of economy, industry and human health may produce successful results in livestock farming even though it is used rarely in this field. Thus, it has been observed that successful results are obtained from the data mining methods in the studies investigated within the scope of the article.

There are many areas and trends in the veterinary field where data mining experts can work. Automation technologies have the potential to enable and advance scientific knowledge. But in the veterinary field, there are limited studies. For example; Early detection / automate the detection of compromised health and welfare in animals using sensors studies, automatically diagnosing the health status of animals used by images studies, animal detection by video camera studies and automatically animal identification studies.

Conflict of interest

The authors declare that they have no conflict of interest.

REFERENCES

1. Alpaydin E., 2014. Introduction to machine learning. 3rd ed., MA: The MIT Press, Cambridge.
2. Witten IH., Frank E., Hall MA., Pal CJ., 2014. Data Mining: Practical machine learning tools and techniques. 4th ed., Morgan Kaufmann Publishers, San Francisco.
3. Han J., Pei J., Kamber M., 2011. Data mining: concepts and techniques. 3rd ed., Morgan Kaufmann Publishers, Amsterdam.
4. Tomar D., Agarwal S., 2013. A survey on Data Mining approaches for Healthcare. J BioSci Biotechnol, 5, 241–266.
5. Cihan P., Gökçe E., Kalıpsız O., 2017. A review of machine learning applications in veterinary field. Kafkas Univ Vet Fak Derg, 23, 673-680. DOI:10.9775/kvfd.2016.17281
6. Küçükönder H., Üçkardeş F., Narinç D., 2014. A data mining application in animal breeding: Determination of some factors in Japanese Quail Eggs affecting fertility. Kafkas Univ Vet Fak Derg, 20, 903-908.
7. Uckardes F., Narinc D., Kucukonder H., Rathert TC., 2014. Application of classification tree method to determine factors affecting fertility in Japanese quail eggs. J Anim Sci Adv, 4, 1017-1023.
8. Mehraban Sangatash M., Mohebbi M., Shahidi F., Vahidian Kamyad A., Qhods Rohani M., 2012. Application of fuzzy logic to classify raw milk based on qualitative properties. Int J Agrisci, 2, 1168-1178.
9. Akıllı A., Atıl H., Kesenkaş H., 2014. Çiğ süt kalite değerlendirmesinde bulanık mantık yaklaşımı. Kafkas Univ Vet Fak Derg, 20, 223-229.
10. Memmedova N., Keskin İ., 2011. İneklerde bulanık mantık modeli ile hareketlilik ölçüsünden

yararlanılarak kızgınlığın tespiti. Kafkas Univ Vet Fak Derg, 17, 1003-1008.

11. Zarchi HA., Jonsson R., Blanke M., 2009. Improving oestrus detection in dairy cows by combining statistical detection with fuzzy logic classification. In Proceedings of the 7th Workshop on Advanced Control and Diagnosis (ACD), Zielona, 20.
12. Brown-Brandl TM., Jones DD., Woldt WE., 2005. Evaluating modelling techniques for cattle heat stress prediction. Biosyst Eng, 91, 513-524.
13. Sharma AK., Sharma RK., Kasana HS., 2007. Prediction of first lactation 305-day milk yield in Karan Fries dairy cattle using ANN modeling. Appl Soft Comput, 7, 1112-1120.
14. Chen LJ., Cui LY., Xing L., Han LJ., 2008. Prediction of the nutrient content in dairy manure using artificial neural network modeling. J Dairy Sci, 91, 4822-4829.
15. Kuncheva LI., del Rio Vilas VJ., Rodriguez JJ., 2007. Diagnosing scrapie in sheep: A classification experiment. Comput Biol Med, 37, 1194-1202.
16. Çelik B., Akçapınar H., 2006. Ankara Keçisinin tiftik özellikleri yönünde kümeleme analizi. Lalahan Hay Arast Enst Derg, 46, 19-27.
17. Küçükönder H., Ayaşan T., Hizli H., 2015. Classification of holstein dairy cattles in terms of parameters some milk component belongs by using the fuzzy cluster analysis. Kafkas Univ Vet Fak Derg, 21, 601-606.
18. Gevrekçi Y., Ataç FE., Takma Ç., Akbaş Y., Taşkın T., 2011. Koyunculuk açısından Batı Anadolu illerinin sınıflandırılması. Kafkas Univ Vet Fak Derg, 17, 755-760.
19. Kılıç İ., Özbeyaz C., 2010. Bulanık kümeleme analizinin koyun yetiştiriciliğinde kullanımı ve bir uygulama. Kocatepe Vet J, 3, 31-37.
20. Niemi JK., Lyytikainen T., Sahlström L., Virtanen T., Lehtonen H., 2009. Risk classification in animal disease prevention: Who benefits from differentiated policy? Selected Paper Prepared for Presentation at the Agricultural and Applied Economics Association 2009 AAE and ACCI Join Annual Meeting, Milwaukee, Wisconsin.
21. Lyytikainen T., Kallio ER., 2008. Risk classification of finnish pig farms by simulated foot and mouth disease spread. Society for Veterinary Epidemiology and Preventive Medicine, Proceedings of a Meeting Held at Liverpook, 285-300.
22. Babcock AH., White BJ., Renter DG., Dubnicka SR., Scott HM., 2013. Predicting cumulative risk of bovine respiratory disease complex (BRDC) using feedlot arrival data and daily morbidity and mortality counts. Can J Vet Res, 77, 33-44.
23. Hermann-Bank ML., Skovgaard K., Stockmarr A., Strube ML., Larsen N., Kongsted H., Ingerslev HC., 2015. Characterization of the bacterial gut microbiota of piglets suffering from new neonatal porcine diarrhoea. BMC Vet Res, 11, 139.
24. Fayyad U., Piatetsky-Shapiro G., Smyth P., 1996. From data mining to knowledge discovery in databases. AI magazine, 17, 37.
25. Niaksu O., 2015. Development and application of data mining methods in medical diagnostics and healthcare management. Technological Sciences, Vilnius University, Vilnius.
26. Silver C., Lewins A., 2014. Using software in qualitative research: A step-by-step guide. 2nd ed., Sage publications, Los Angles.
27. Cios KJ., Pedrycz W., Swiniarski RW., 2007. Data mining: A Knowledge Discovery. Springer, New York.
28. Tan PN., Steinbach M., Kumar V., 2006. Introduction to Data Mining, Pearson Addison-Wesley, Boston.
29. Liu B., Hsu W., Ma Y., 1998. Integrating classification and association rule mining. The Fourth International Conference on Knowledge Discovery and Data Mining, New York City, 1-7.
30. Yazdanbakhsh O., Zhou Y., Dick S., 2017. An intelligent system for livestock disease surveillance. Information Sciences, 378, 26–47.
31. Saidani K., Lopez-Sandez C., Pablo DF., 2016.

Effect of climate on the epidemiology of bovine hypodermosis in Algeria. Kafkas Univ Vet Fak Derg, 22, 147-154.

32. Awaysheh A., Wilcke J., Elvinger F., Rees L., Fan W., Zimmerman KL., 2016. Evaluation of supervised machine-learning algorithms to distinguish between inflammatory bowel disease and alimentary lymphoma in cats. J Vet Diagn Invest, 28, 679-687.
33. Boujenane I., El Aimani J., 2015. Incidence and occurrence time of clinical mastitis in Holstein cows. Turk J Vet Anim Sci, 39, 42-49.
34. Amrine DE., White BJ., Larson RL., 2014. Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease. Comput Electron Agric, 105, 9-19.
35. Urbach YK., Raber KA., Canneva F., Plank AC., Andreasson T., Ponten H., Kullingsjö J., 2014. Automated phenotyping and advanced data mining exemplified in rats transgenic for Huntington's disease. J Neurosci Methods, 234, 38-53.
36. Ahmed H., Khan MR., Panadero-Fontan R., Sandez CL., Asif S., Mustafa I., Qayyum M., 2013. Influence of epidemiological factors on the prevalence and intensity of infestation by Hypoderma spp. (Diptera: Oestidae) in cattle of Potowar Region, Pakistan. Pakistan J Zool, 45, 1495-1500.
37. Piwczynski D., Sitkowska B., Wisniewska E., 2012. Application of classification trees and logistic regression to determine factors responsible for lamb mortality. Small Rumin Res, 103, 225-231.
38. Geenen PL., van der Gaag LC., Loeffen WLA., Elbers ARW., 2011. Constructing naive Bayesian classifiers for veterinary medicine: A case study in the clinical diagnosis of classical swine fever. Res Vet Sci, 91, 64-70.
39. Lopez CM., Fernandez G., Vina M., Cienfuegos S., Panadero R., Vazquez L., Diez-Banos P., Morrondo P., 2011. Protostrongylid infection in meat sheep from Northwestern Spain: Prevalence and risk factors. Vet Parasitol, 178, 108-114.
40. Dawson, S., Else RW., Rhind SM., Collie DDS., 2005. Diagnostic value of cytology of bronchoalveolar fluid for lung diseases of sheep. Vet Rec, 157, 433-436.
41. Sandholm T., Brodley C., Vidovic A., Sandholm M., 1996. Comparison of regression methods, symbolic induction methods and neural networks in morbidity diagnosis and mortality prediction in equine gastrointestinal colic. AAAI 1996 Spring Symp Ser Artif Intell Med Appl Curr Technol, California, 154-159.
42. Ripley BD., 2001. The R project in statistical computing. MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network 2001, 1, 23-25.
43. Trewartha D., 2006. Investigating data mining in MATLAB. Rhodes University, Department of Science, South Africa.
44. Rangra K., Bansal KL., 2014. Comparative study of data mining tools. Int J Adv Res Comput Sci Softw Eng, 4, 216-223.
45. Pujari P., Gupta JB., 2012. Exploiting data mining techniques for improving the efficiency of time series data using spss-clementine. Res World J Arts Sci Commer, 3, 69.
46. Matignon R., 2007. Data mining using SAS enterprise miner. 1st ed., John Wiley & Sons, New York.