MAE Vet Fak Derg, 7 (3): 175-187, 2022
DOI: 10.24880/maeuvfd.1121112

Research Article

Analysis and estimation of pathological data and findings with deep learning methods

Ahmet Anıl Şakır[1], Ali Hakan Işık[1], Özlem Özmen[2], Volkan İpek[2]

[1]Department of Computer Engineering, Graduate School of Natural and Applied Sciences, Burdur Mehmet Akif Ersoy University, Burdur, Türkiye
[2]Department of Pathology , Faculty of Veterinary Medicine, Burdur Mehmet Akif Ersoy University, Burdur, Türkiye

Key Words:
classification
decision tree
deep learning
KNN
pathology

Received : 25.05.2022
Accepted : 01.09.2022
Published Online : 31.12.2022
Article Code : 1121112

Correspondence:
AA. ŞAKIR
(ahmetanilsakir@gmail.com)

ORCID
AA. ŞAKIR : 0000-0003-1317-8089
AH. IŞIK : 0000-0003-3561-9375
Ö. ÖZMEN : 0000-0002-1835-1082
V. İPEK : 0000-0001-5874-7797

The present M.Sc. thesis was supported by Burdur Mehmet Akif Ersoy University Scientific Research Projects Under the Project number of 0671-YL-20

ABSTRACT

As in human diseases, rapid diagnosis of animal diseases is of great importance. In order for the disease treatments to be carried out properly, the diagnosis must be of high accuracy, as well as the rapid diagnosis. In this study, the disease types in the data set consisting of the data examined between the years 2000-2020 belonging to the Department of Pathology of the Faculty of Veterinary Medicine of Burdur Mehmet Akif Ersoy University were estimated by using the decision tree classification model and the KNN classification model. Categories such as age, type, city, and gender in the data set were analyzed in graphics. For the estimation and analysis processes to give accurate results, the data set was corrected by going through some pre-processes and the missing data in the data set was completed. It is thought that the results obtained from the estimation and analysis will allow rapid and accurate diagnosis in animal disease diagnoses.

INTRODUCTION

The word pathology is of Latin origin, and it is a definition formed by the combination of the words pathos, which means disease, and logos, which means science (Slauson and Cooper, 1990; Carlton and McGavin, 1995; Cheville, 1999). While Medical Pathology, one of the sub-branches of pathology, deals with human diseases, Veterinary Pathology deals with animal diseases (Slauson and Cooper, 1990; Kahraman, 1996; Cheville, 1999). The pathological data, consists of the data examined between the years 2000-2020 in Burdur Mehmet Akif Ersoy University, Faculty of Veterinary Medicine, Department of Pathology.

As in human diseases, rapid diagnosis is of great importance in animal diseases. In addition to being rapid, the diagnosis should also have a high accuracy value. Today, computer applications are widely used in the field of health. One of the best examples of this is the use of artificial intelligence applications as a cancer treatment tool (Sütçü & Aytekin, 2018).

When the data in the database are evaluated according to animal species, most of the samples examined belong to ruminants, followed by animals such as dogs and cats. It was observed that digestive and respiratory system diseases were the leading causes of death in adult ruminants, and neonatal septicemia in juvenile ruminants caused significant death. Tumors were most frequently encountered in cats and dogs (Özmen, 2006).

Pandas library was used for data processing and analysis (Pandas-a, 2021). The main purpose of using the Pandas library is to use dataframe structures (Pandas-b, 2021). At the same time, it can import data in a simple way thanks to its data import feature from different formats like excel (Pandas-c, 2021). Since the Pandas library can also work with libraries such as NumPy and Matplolib, it also provides the opportunity to use these libraries. The NumPy library is a library that enables mathematical operations on arrays using matrices and arrays (Numpy-a, 2021; Numpy-b, 2021). Since the Pandas library uses dataframes as objects, it is possible to manipulate these dataframes as arrays with NumPy. Another library, MatPlotLib is a library that provides graphics creation. It can easily create various graphics such as line, bar, or pie plot over the dataframe.

Bu makaleye atıfta bulunmak için/ To cite this article:
Şakır AA. Işık AH. Özmen Ö. İpek V. Analysis and estimation of pathological data and findings with deep learning methods. MAE Vet Fak Derg. 2022; 7 (3): 175-187.

The method used to complete missing data is the SimpleImputer method in the sklearn.impute library. The SimpleImputer method uses different strategies according to the format of the data (numeric or categorical) (Scikit Learn-a, 2021). Since the data in the database are categorical data, the strategies used in the categorical data were used in the incomplete data completion processes. The methods used in the prediction of disease type are, DecisionTreeClassifier method in the sklearn.tree library and KNeighborsClassifier method in the sklearn.neighbors library (Scikit Learn-b, 2021; Scikit Learn-e, 2021).

Diagnosis has a great role for the treatment and control of diseases in animals. Evaluation of lesions in dead or live animal specimens is important so that the diagnosis can be made rapidly and accurately. Diagnosis is based on understanding general and specific pathology and the application of these categories to diagnosis (Jones & Hunt, 1993; Özmen, 2006). For the correct treatment of diseases, the diagnosis must be made precisely and correctly. Samples for diagnosis can be taken after the death of animals, or pathological samples can be obtained from live animals by appling surgical procedures. In these surgical procedures, reasons such as the inability to obtain tissues using the correct technique or obtaining faulty tissue may lead to incorrect diagnoses and therefore inability to perform treatments properly (Nakhleh, 2015; Özmen, 2021). The correct processing of the taken of pathological samples and the proper evaluation of the pathological findings not only increase the success rate in diagnosis, but also increase the success rate in the treatment process (Özmen, 2021).

As can be seen in Table 1, artificial intelligence and deep learning applications have come a long way from the past to the present, and they can be widely used not only in the field of health but also in many fields. For example, eight of hundred companies that develop artificial intelligence applications, develop applications in the field of health (Sütçü and Aytekin, 2018). As technology develops, the increase in this number and the applications produced can greatly benefit humanity in developing diagnostics that are both rapid and accurate. In this study, it is aimed to analyze animal diseases, to make a quick diagnosis depending on the estimation result, and to have a high accuracy rate in relation to the diagnosis speed. Our aim is to analyze the animal disease types, with animal species and pathological-anatomical diagnoses and their prediction using

Table 1. A timeline of specific AI innovations that resulted in the conquest of cancer (Sütçü & Aytekin, 2018)

Date	Description
1952	Marvin Minsky introduced the Stochastic Neural Analog Reinforcement Calculator (SNARC), the first connectivity neural network learning machine, and possibly the first self-learning machine.
1975	The back propagation algorithm was developed, which solves the difficulties in computer-aided machines, trains multi-layer neural networks, and provides widespread use of neural networks in the 1980s.
Around 2000	The term "deep learning" was used for the first time to describe a machine learning, the creation of networks that can learn from unstructured data in an unsupervised manner.
2011-2012	The convolutional neural network AlexNet has achieved unprecedented accuracy in visual recognition, paving the way for the deep dives into the mainstream.
January 2017	Researchers at Stanford University have developed a deep leaning technology that can visually identify cancerous skin and lesion with the same precision as a human dermatologist.
February 2017	Microsoft founded Healthcare NeXT, a startup design to apply artificial intelligence and machine learning technologies to health problems, including cancer treatment.
March 2017	Google's GoogleNet deep learning technology detected cancerous tumors with higher accuracy than human clinicans.
October 2017	Intel has announced the Nervana Neural Network Processor (NNP) chip that can accelerate deep learning tasks, including cancer diagnosis.
Around 2021-2026	Microsoft will launch an AI-powered computer inside the human body to detect and reprogram cancerous cells and render them harmless.

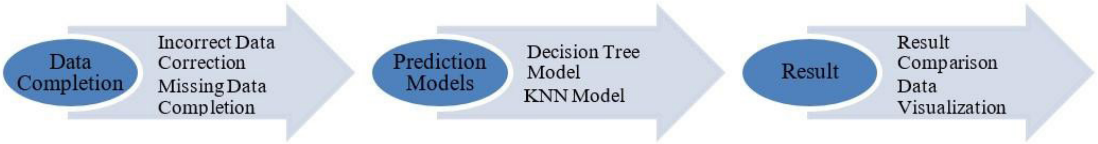
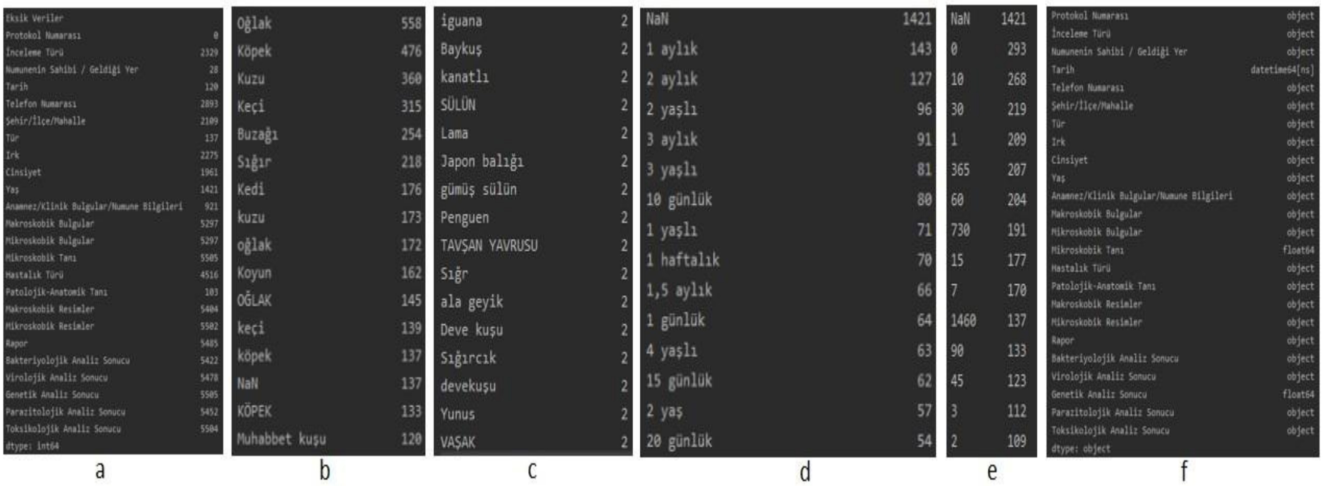


Figure 1. Flow chart followed in the study.

several categories available in the database.

MATERIALS and METHODS

The flow chart that showing of the path to be followed in this study is given in Figure 1. Firstly, data completion processes were carried out, and two different classification models were created using the completed data. The results were compared and then data visualization processes were applied.



```
Eksik Veriler
Protokol Numarası                          0
İnceleme Türü                           2329
Numunenin Sahibi / Geldiği Yer            28
Tarih                                    120
Telefon Numarası                        2893
Şehir/İlçe/Mahalle                      2109
Tür                                      137
Irk                                     2275
Cinsiyet                                1961
Yaş                                     1421
Anamnez/Klinik Bulgular/Numune Bilgileri 921
Makroskobik Bulgular                    5297
Mikroskobik Bulgular                    5297
Mikroskobik Tanı                        5505
Hastalık Türü                           4516
Patolojik-Anatomik Tanı                  103
Makroskobik Resimler                    5404
Mikroskobik Resimler                    5502
Rapor                                   5485
Bakteriyolojik Analiz Sonucu            5422
Virolojik Analiz Sonucu                 5478
Genetik Analiz Sonucu                   5505
Parazitolojik Analiz Sonucu             5452
Toksikolojik Analiz Sonucu              5504
dtype: int64
```

```
Oğlak            558
Köpek            476
Kuzu             360
Keçi             315
Buzağı           254
Sığır            218
Kedi             176
kuzu             173
oğlak            172
Koyun            162
OĞLAK            145
keçi             139
köpek            137
NaN              137
KÖPEK            133
Muhabbet kuşu    120
```

```
iguana             2
Baykuş             2
kanatlı            2
SÜLÜN              2
Lama               2
Japon balığı       2
gümüş sülün        2
Penguen            2
TAVŞAN YAVRUSU     2
Sığr               2
ala geyik          2
Deve kuşu          2
Sığırcık           2
devekuşu           2
Yunus              2
VAŞAK              2
```

```
NaN           1421
1 aylık        143
2 aylık        127
2 yaşlı         96
3 aylık         91
3 yaşlı         81
10 günlük       80
1 yaşlı         71
1 haftalık      70
1,5 aylık       66
1 günlük        64
4 yaşlı         63
15 günlük       62
2 yaş           57
20 günlük       54
```

```
NaN    1421
0       293
10      268
30      219
1       209
365     207
60      204
730     191
15      177
7       170
1460    137
90      133
45      123
3       112
2       109
```

```
Protokol Numarası                           object
İnceleme Türü                               object
Numunenin Sahibi / Geldiği Yer              object
Tarih                               datetime64[ns]
Telefon Numarası                            object
Şehir/İlçe/Mahalle                          object
Tür                                         object
Irk                                         object
Cinsiyet                                    object
Yaş                                         object
Anamnez/Klinik Bulgular/Numune Bilgileri    object
Makroskobik Bulgular                        object
Mikroskobik Bulgular                        object
Mikroskobik Tanı                           float64
Hastalık Türü                               object
Patolojik-Anatomik Tanı                     object
Makroskobik Resimler                        object
Mikroskobik Resimler                        object
Rapor                                       object
Bakteriyolojik Analiz Sonucu                object
Virolojik Analiz Sonucu                     object
Genetik Analiz Sonucu                      float64
Parazitolojik Analiz Sonucu                 object
Toksikolojik Analiz Sonucu                  object
dtype: object
```

Figure 2. Missing data in the database (a), values for type category (b1, b2), values for age category (c) age category after editing (d), data types (e).

Data Completion

Before proceeding with the analysis and estimation processes, missing or incorrect data in the data set should be corrected. In this way, it is aimed to increase the success rate of the estimation, while providing more accurate analysis. While there are approximately 5500 reviews from 24 different categories in the database, it is seen that some data are incomplete or incorrectly entered. In Figure 2-a, the values of the missing data based on categories are given.

Incorrect Data Correction

The presence of incorrect data in the database, usually due to reasons such as wrong lettering or use of small capital letters, may adversely affect the analysis result. Figure 2-b shows some of the values belonging to the Species category and how many of these values there are. For example, the "Oğlak(goat)" value is written in 3 different ways, and this same value is divided into 3 separate data sets. In another example, it can be seen in the data entered as "Sığr" instead of "Sığır(cattle)". Likewise, the example of "Deve Kuşu" and "devekuşu(ostrich)" can be given as examples of wrong data. In other words, the same species may appear more than once, and some species appear to be expressed with different words (sığır(cattle) vs. inek(cow), etc.). There are 301 clusters in the Species category. After correcting the wrong data, the number of clusters decreases to 69. With this decrease in the number of clusters, a more general and collectively categorized category has emerged, allowing clearer and more precise results in the analysis results.

Although incomplete and incorrect data are seen in the spelling of letters and similar words in most categories, the Age category may be the one with the most irregularity (Figure 2-c). The reason for this is that data is entered in multiple values (daily, weekly, monthly, etc.), instead of a single standard value. To fix this, each data must be specified over a single value. We can express this in the best way in a daily format.

After correction, the Age category can be seen in Figure 2-d. While there were 647 different categories before the correction, there are 149 different categories after the correction. This shows that it needs to be corrected to achieve a more standard and effective result.

The "NaN" value that appears in the Age category means "not a number". Some categories are originally referred to as objects, and they must be strings to be edited. When these categories are converted to string values, the null values that were originally "null" do not appear as empty strings. Therefore, "NaN" value is used instead of "null" value (Figure 2-e).

Missing Data

Before completing the missing data, data visualization was made, and it was determined which data were missing. In this way, it is aimed to complete the missing data in a simpler way. The graph shown in Figure 3 gives the distribution of missing data by categories. The protocol number appears to be complete, as it is in the form of an identifier for each entry. In the categories with a large number of missing data, those outside the category of disease type will not be used in examinations and analyzes as they will not affect the result.

Missing Data Completion

After correcting the erroneous data in the data set, data completion operations can be applied. Primarily, the bar graph showing the number of missing data after removing the missing categories and correcting the erroneous data is given in Figure 4. By dividing the number of missing data by the total number of data, we can learn the rate of missing data in the data set. To find out as a percentage value, it will be sufficient

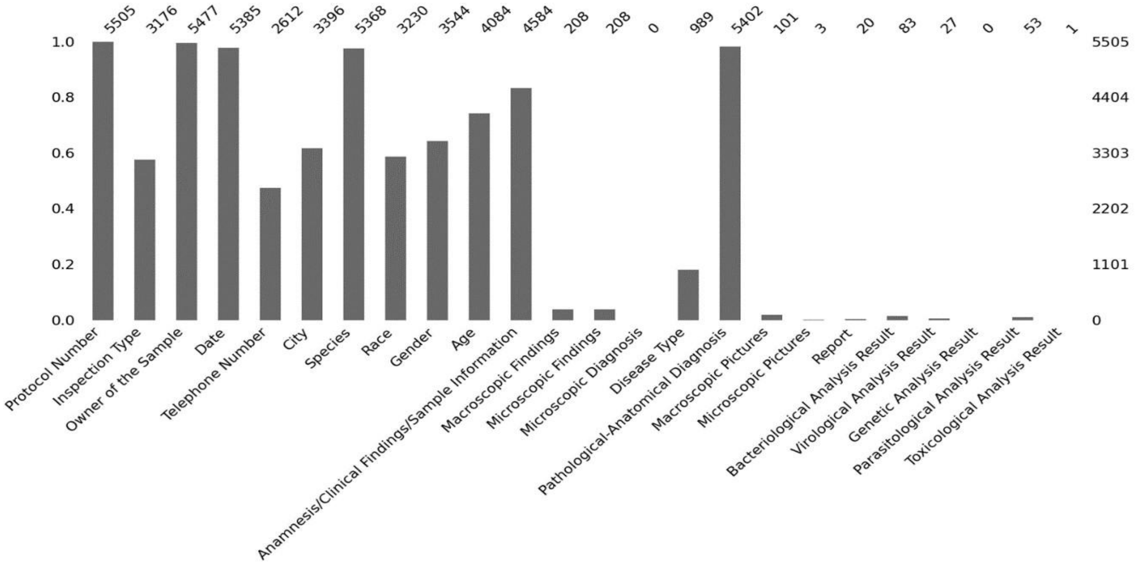


Figure 3. Bar graph representation of missing data.

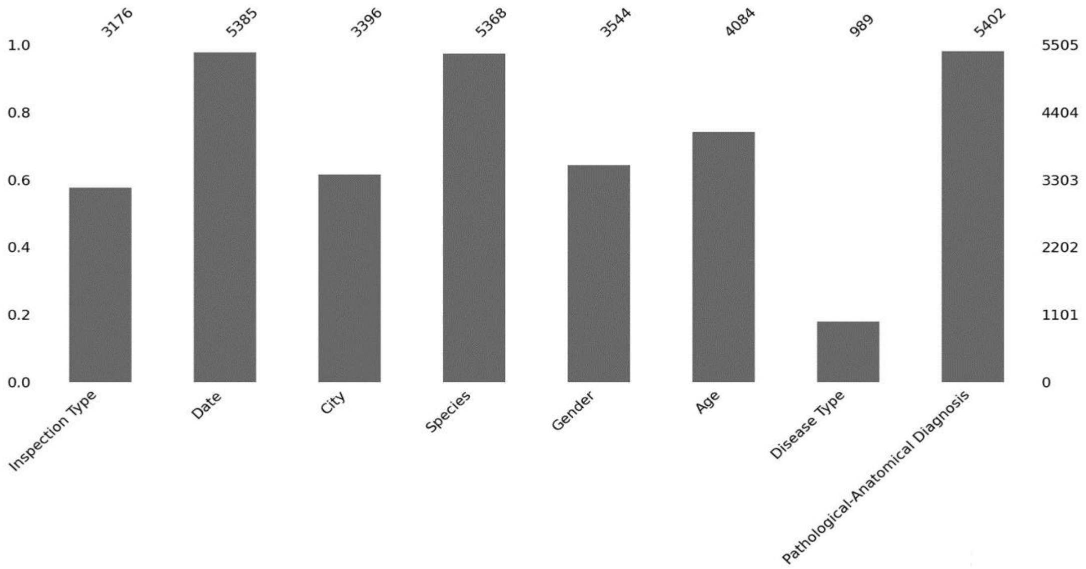


Figure 4. Number of missing data in the data set after missing categories were removed and incorrect data were corrected.

to multiply the result by 100. After performing these operations, it was calculated that there was 59.54% missing data in the data set. The amount of missing data to be used in analysis and estimation processes in the data set after the category removal process is 28.89%. This result shows that almost 1 out of every 4 data is missing.

According to the data graphics after removing the missing categories and correcting the wrong data, it is seen that more than half of the categories other than the Disease Type category are not missing. Among these categories, the Date category is the simplest as incomplete data completion. It is possible to complete the missing data by looking at the date range entered in the data set. This can be done by looking at the date of the data that comes before the missing data. With this method, the Date category is filled in completely.

Another category that can be completed manually is the City category from which the samples come. The missing data were completed by comparing the sample owners with the City categories. The Simple Imputation method of the sklearn.impute library can be used to fill in the remaining categories. This method applies data completion in accordance with both numerical and categorical data. There are two different strategies for categorical data. These are: most frequent, that is, the most frequently found data in the database, and constant, that is, entering a constant value (Scikit Learn-a, 2021). The most frequent method was applied in the remaining missing categories, and the missing data in the categories other than Disease Type were completed.

The reason why the Disease Type category is not completed with the imputation method is that the amount of data in the category is too incomplete and if it is completed with this method, the result will be the same as the most found data in the category. To complete the Disease Type category, it should be compared with the Pathologic-Anatomical Diagnosis category and the results should be taken according to their common values. As a result of this process, Disease Type category is divided into bacterial, tumoral, parasitic, viral, and other, which can be counted as 5 main categories. Other category includes the conditions in which the remaining four categories are together, as well as anomalies and traumatic lesions.

Forecasting with the Decision Tree Method

After the data completion processes were completed, the model estimation processes were performed on the complete data set. Prediction models can be expressed in two different ways: classification and regression. The prediction obtained by the decision tree method results in classification. Classification is the definition of data into predetermined classes according to their common characteristics. The Disease Type category in the data set was used for classification. Therefore, the category is divided into 5 separate classes: bacteriyal, tumoral, parasitic, viral, and other.

The Decision Tree method used for classification has a structure like the flowchart graph and allows the data to be divided into specified classes (Figure 5). The inductive method is used for classification. In the decision tree method, by performing many tests during classification, the classification result is obtained in the best way. During this process, each test creates one of the branches of the decision tree. For the test process to end, it must reach the leaf node. The order from the tree root to the leaf node containing the classification result is called the if-then rule (Emel and Taşkın, 2005).

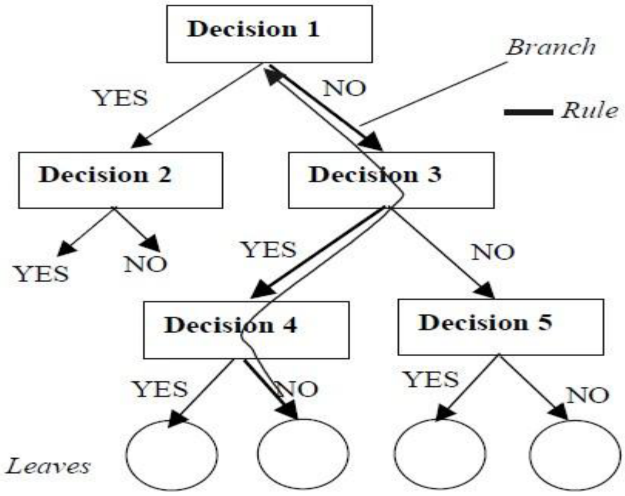


Figure 5. An example decision tree diagram (Bounsaythip and Esa, 2001).

The induction method used in the decision tree is called tree induction. This method starts with an empty tree. Tree induction used to fill this tree is an iterative method (recursive) and consists of four steps (Emel and Taşkın, 2005; Zorman et al., 2001):

Step 1: If the result of the training objects in the training set is the same, a leaf node is created using this result and continue from step 4.

Step 2: Using the heuristic evaluation function, starting from the root of the tree, on the way to the current node, the best feature is selected among the previously unused features and a split node is created for this feature. The training set is divided into subsets.

Step 3: Continue from step 1 for each sub-training set created.

Step 4: Recursion is performed by going up one level.

As a result of these four steps, basically two processes take place. These are splitting and pruning. After these processes, the stopping criterion comes to the end of the iteration method (Emel and Taşkın, 2005).

Division is a method that allows the training set to be divided into smaller subsets. In division, which is an iterative method, the first iteration covers the entire training set, including the tree root. The remaining iterations are processed using derivative nodes that include subsets of the training set. At each division step, the data is analyzed, and the best classification is selected. The most important characteristic of the division process is that it is greedy, therefore, in this method, the algorithm does not look at the steps forward on the tree to find out whether it has achieved the best result (Emel and Taşkın, 2005; Bounsaythip and Esa, 2001).

For the decision tree to be formed, the iterative method must stop. Stop criteria are used for this stopping operation. The stopping criteria usually include a few rules such as the maximum tree depth, the minimum number of items in the node considered for splitting, or the minimum number of items that should be included in the new node. It can change the parameters associated with these rules according to the data type used or the user's request. Generally, applications using this method build trees at maximum depth. While such a tree with maximum depth predicts all the objects in the training set with a certain probability, they are most likely overfit the data (Bounsaythip and Esa, 2001).

After the decision tree is formed, pruning is used to remove unwanted nodes or subsets due to overfitting. The pruning method removes partitions and their subsets and makes the decision tree more stable. Applications that create trees with maximum depth include an automatic pruning method. After the decision tree training is completed, the estimation process can be applied for the new data by using the path formed from the top of the tree until reaching any result node (Bounsaythip and Esa, 2001).

DecisionTreeClassifier belonging to the Sklearn library is a library used for classification. To make classification, firstly, the categories or columns to be used in the classification and the column containing the classification result should be divided into separate dataframes. To classify the Disease Type, Species and Pathological-Anatomical Diagnosis columns are separated. In some cases, the separated categories need to go through some pre-processing before they can be classified. Here, the one-hot method is used to convert data in categories containing text to numeric values. In the Pandas library, the one-hot method can be easily processed as a single line of code with the pandas.get_dummies method (Pandas-d, 2021). When the categories to be used in classification are ready, model training can be started. In order to determine how successful the model is, the data set should be divided into two sets as training set and test set, before model training. For this separation, the train_test_split method of the Sklearn.model_selection library was used (Scikit Learn-c, 2021). Separation rate can be determined with test_size, which is one of the parameters of this method. This allowed us to separate the data set as the test set as much as the entered parameter size and the rest as the training set. After the data set is divided into two sets as training and test, the model to be used for classification should be defined before starting the model training for classification. This definition can be done by calling the DecisionTreeClassifier function belonging to the Sklearn.tree library. After the defination process, the training is completed using the fit function (Scikit Learn-b, 2021). After the completion of the training, the success rate becomes measurable, and the model is ready to make predictions for the new data to be entered. In order to measure the success rate, the values in the test set are estimated over the trained model by using the predict parameter belonging to the DecisionTreeClassifier function. Then, the accuracy_score method of the Sklearn.metrics library was used to compare these estimates with the actual result values (Scikit Learn-d, 2021).

Forecasting with the KNN Method

The other classification method to be used in this study is KNN (k-nearest neighbors) that classifies according to the mean value of the nearest neighbors in the training set. The k value here indicates the number of neighbors to be selected. In order to find the nearest neighbor, it is necessary to calculate the distance between the points. Different distance metric measurement methods such as Euclid, Manhattan, Minkowski can be used to measure this distance. Generally, Euclidean distance is the most preferred method among these methods (Euclidian Distance, 2021; Scikit Learn-f, 2021). The Euclidean distance used to calculate the linear distance between two points (eg $P = (p_1, p_2,\dots,p_n)$ and $Q = (q_1, q_2,\dots,q_n)$ points) is given in formula 1 (Euclidian Distance, 2021).

$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2}=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}.$$

Formul 1.

The KNeighborsClassifier method of the Sklearn.neighbors library designed for KNN is a method that includes preliminary steps such as distance measurement (Scikit Learn-e, 2021). As in the decision tree method, before starting the model training, the data set must be pre-processed and divided into two sets as the training set and the test set. For the definition of the model, the KNeighborsClassifier method was used with the k parameter indicating the number of neighbors (Scikit Learn-e, 2021). After this process, the data sets are transferred to the model with the fit method and the training is completed. The predict parameter in the DecisionTreeClassifier library is also available in the KNeighborsClassifier method. Estimation is performed with this parameter. For the success rate, the accuracy_score method of the sklearn.metrics library is used (Scikit Learn-d, 2021).

Calculating Margin of Error with RMSE, MSE and MAE Methods

Among the methods used to calculate the difference between the estimated result and the actual values, there are methods such as MSE (mean squared error), RMSE (root mean squared error) and MAE (mean absolute error). The MSE, or mean absolute margin of error, is the mean value of the absolute difference between both variables in the data set. If the MAE result is close to 0, it indicates that the value gives the best result. The formula for calculating MAE is given in Formula 2 (Mean Absolute Error, 2021).

$$\mathrm{MAE}=\frac{\sum_{i=1}^{n}|y_i-x_i|}{n}=\frac{\sum_{i=1}^{n}|e_i|}{n}.$$

Formul 2.

The MSE method, which is referred to as the mean square error, gives an absolute number indicating how much the predicted results differ from the actual results in the data set. The closer the MSE value is to 0, the better the result is. MSE is given in Formula 3 (Mean Squared Error, 2021).

$$\mathrm{MSE}=\frac{1}{n}\sum_{i-1}^{n}(Y_i-\hat{Y}_i)^2.$$

Formul 3.

RMSE (or RMSD), root mean square error method is the square root of the result of the MSE method and gives smaller results than the MSE method. For this reason, it is generally preferred over the MSE method. The RMSE is given in Formula 4 (Root Mean Squared Error, 2021).

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t-1}^{T}(\hat{y}_t - y_t)^2}{T}}.$$

Formul 4.

The mean_squared_error and mean_absolute_error methods of the sklearn.metrics library were used to measure the margin of error between the predictions and the actual result in this study. By changing the squared parameter in the mean_squared_error method to true or false, it is possible to switch between MSE and RMSE methods (Scikit Learn-g, 2021; Scikit Learn-h, 2021).

RESULTS

Graphical Findings

As a result of the analyzes made on the categories belonging to the database, some graphics were produced. While some of these graphics were extracted directly from dataframes using Pandas and MatPlotLib libraries, SPSS 22.00 version program was used for some of them. In order for the graphs to be readable and smooth, the most common values in the category were shown, and the remaining values were arranged in a collective way.

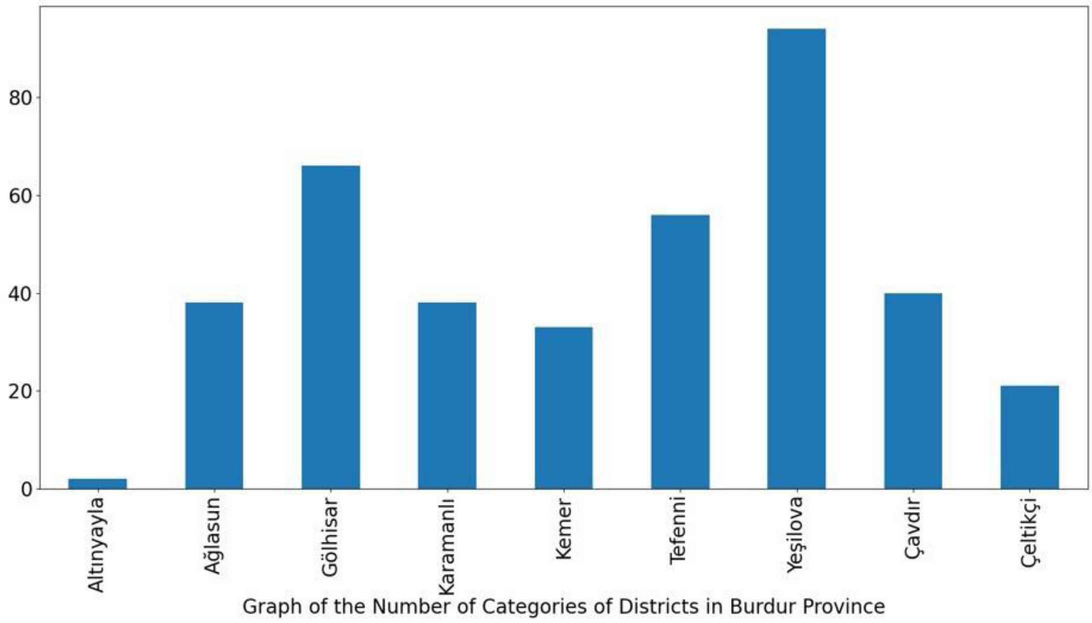


Figure 6. Graph of the number of categories of districts in Burdur province.

Charts Obtained as a Result of Analysis

There are 24 different values in the City category in the database. These are the 24 provinces from which the samples came. Except for the provinces, it is found in samples from the districts of Burdur province. These are shown in Burdur in the provincial category (Figure 6).

According to this graph, the most samples came from Yeşilova district in Burdur province. On the provincial basis, the most samples came from the province of Burdur. In the evaluation of all animal species in the data set according to Gender, the majority of the samples coming between the years 2000-2020 are female animals (61.14%).

The Inspection Type category applied to the incoming samples is divided into four main areas. These; Biopsy, Necropsy, Cytology, and Organ Sample. The combination of these four fields is shown in the other category (eg Biopsy; Organ Sample). Necropsy was the most common type of examination performed on the samples (75.59%). This is followed by Biopsy (14.28%), Cytology (5.01%), Organ Sample (3.94%) and other categories, respectively.

There are 68 different categories of species appearing in the database in the separation of incoming samples according to species. The first five categories that make up the majority of these are percentaged, while the remaining species are shown as other. According to these values, the most sampled animal species was goat (29.46%), respectively; sheep (17.68%), cattle (15.30%), dogs (14.04%) and cats (5.29%). In the evaluation made according to the age ratio of the incoming samples, it is seen that the most samples come under the age of 1 (79.46%). This is followed by over 5 years (8.37%), 1 year (5.36%), 2 years (4.05%), 4 years (2.52%) and 3 years (0.24%).

The Disease Type category, which is also a classification result category, gives four different results. These are respectively; bacterial (41.34%), tumoral (26.38%), parasitic (10.32%) and viral (6.49%). Results other than these values are given in the other category. The other category, which includes the combination of these four results, also includes anomaly and traumatic values.

According to the cities, the city with the most samples is Burdur, and the most common disease from this city seems to be bacterial. After the bacterial disease comes the tumoral disease. The most common species from Burdur was goat. Goat type respectively; cattle, dogs, sheep, and cats follow. In the evaluation made according to the examination type, it was observed that necropsy, which is the most common type of examination, was applied more in female animals. Bacterial diseases, which are the most common type of disease among animal species, are most common in goats, while tumoral diseases are most common in dogs. According to the sex ratio of the samples coming between 2000 - 2020, female animals came almost every year more than male animals.

The line chart showing the number of samples received between 2000 and 2020 is given in Figure 7. According to the graph, it is seen that the number of samples increased in 2011, 2012, 2014, 2015, 2018 and 2019. The reason for the decrease in 2020 is due to the fact that the data is until August 2020.

under. The category in which necropsy is the most common type of examination consists of animals aged 1 and younger. In the examination types other than necropsy, it is seen that animals 1 year old and younger are in the majority. Again, it is seen that female animals come more than male animals. According to the distribution of animal species according to age, it comes to the fore that mostly 1 year old and younger animal species come to the fore in all species.

Similar to other results, it is seen that animals under the age of 1 are also in the first place in the category of disease types. In addition, it is seen that bacterial and tumoral disease types are generally seen in other age categories.

Considering the ratio of the sexes of animals to animal species, it is observed that female animals are more common in goat, cattle, sheep, dog, and cat species, while the number of females in the goat category, which is the most common animal species, is even more than twice the number of males. On the other hand, in the category that includes the rest of the animal species, it is seen that the ratio of males is high. In this category, budgerigar, chicken, and fish species are the species with a higher male ratio. Finally, it is observed that the disease types are generally seen on female animals.

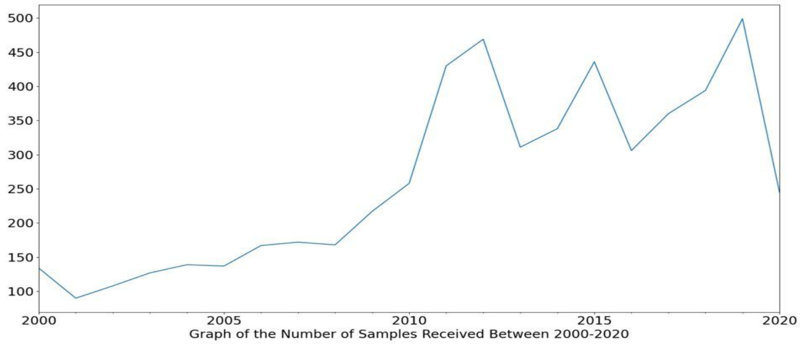


Figure 7. Graph of the number of samples received between 2000-2020.

Graphs Obtained from Statistical Analysis

Statistical analysis of the data of this study was done in SPSS (v.22.00) package program. The same program was used for the graphical representation of the data. When we look at the distribution of animal ages by years, it is seen that the most common cases are in the 1 year and under category, and they came between 2011-2015, while animals aged 10 and over usually come in 2015 and later years. It was observed that the animals coming from Burdur province were mostly 1 year old and

Decision Tree Method Results

During the decision tree classification process, the test_size value entered as a parameter is used to distinguish between training and test data sets. In this study, a test data set of 30 percent was created (3853 training sets, 1652 test sets, a total of 5505 data), and the estimation and error margin results of the models trained with these data sets are given in Table 2. In

Table 2. Table of values obtained as a result of the decision tree model

max_depth	Prediction	MAE	MSE	RMSE
5	0,643	0,752	2,032	1,426
10	0,710	0,615	1,680	1,296
15	0,719	0,568	1,493	1,222
20	0,719	0,568	1,493	1,222
None	0,719	0,568	1,493	1,222

the table, the estimated result values obtained from the models with different depths and the margin of error values of MAE, MSE and RMSE are given. According to the data in the table, it is observed that the result is stable when the depth is defined as 15 and above. The default value of the max_depth parameter to "None" indicates that the depth of the tree has reached the maximum depth it can reach. These value results are added to the last row in the table.

The max_depth value is the parameter that specifies the depth number of the decision tree, and when the default value is set to "None", it goes to the deepest point of the tree. The graph consisting of the results from the table is given in Figure 8. As can be seen from the graph, when the depth number is defined as 15 or more, a fixed value appears in the results and margins of error. For this model, 15 depth values are seen as the best value, and the graph related to this depth is given in Figure 9.

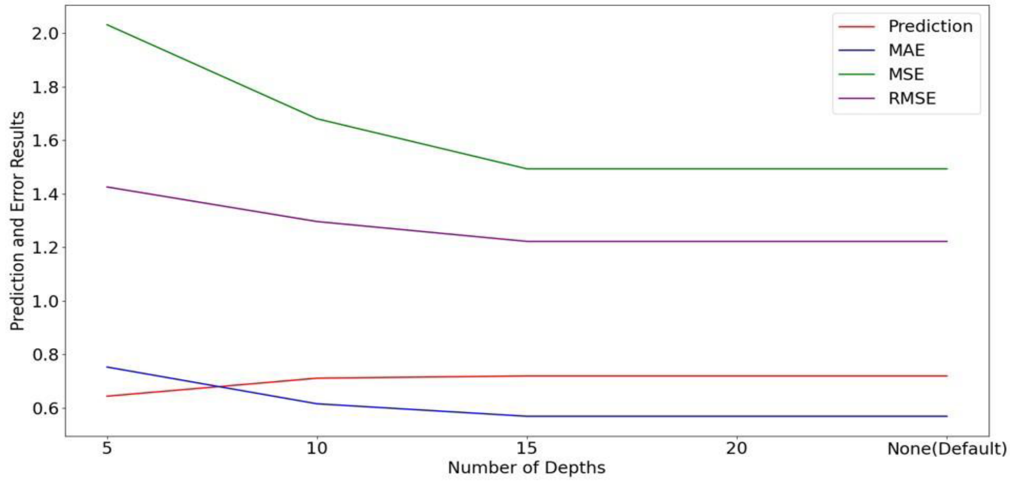


Figure 8. Decision tree model prediction and error margin graph.

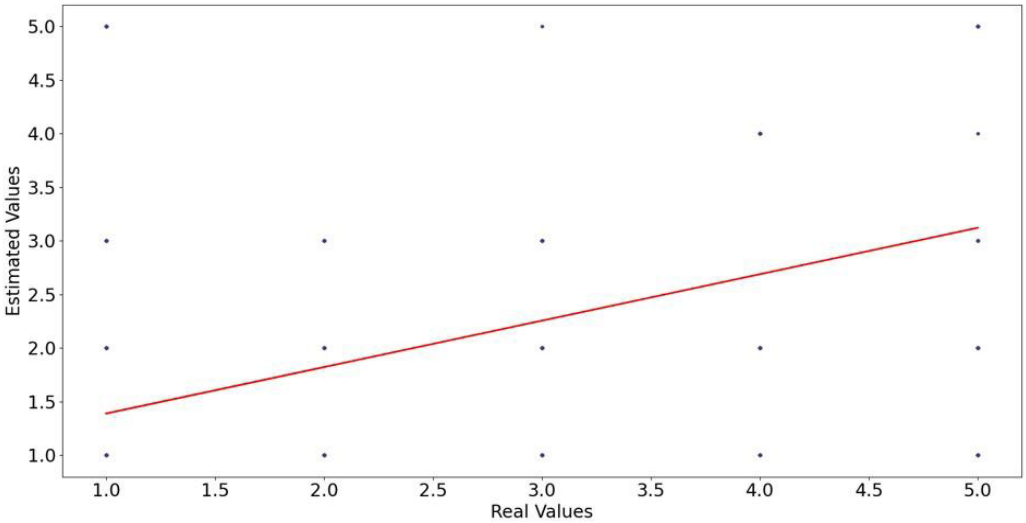


Figure 9. Pattern graphic with depth number 15.

KNN Method Results

For the KNN classification process, the training and test data sets prepared for the decision tree model are used in the same size. The estimation and margin of error results of the KNN model are given in Table 3. In the table, the estimated result values obtained from the models with different near-neighbor numbers and MAE, MSE and RMSE margin of error methods are given. According to the data in the table, the higher the number of neighbors, the better the prediction value in general, and the margin of error values seem lower than the previous value. In the KNN model, the default number of neighbors is 5, and it is seen that the values above the default number of neighbors yield better results.

The graph consisting of the results from the table is given in Figure 10. As can be seen from the graph, the higher the number of neighbors, the better the results emerge. The estimation result of the value with 8 nearest neighbors for the KNN model was the highest among the results of other

Table 3. Table of values resulting from the KNN model

n_neighbors(k)	Prediction	MAE	MSE	RMSE
3	0,700	0,665	1,844	1,358
4	0,699	0,661	1,804	1,343
5 (default)	0,699	0,661	1,804	1,343
6	0,709	0,582	1,508	1,228
7	0,704	0,601	1,591	1,262
8	0,718	0,594	1,607	1,268
9	0,710	0,579	1,492	1,221

The n_neighbors(k) value indicates the number of near neighbors, and the default value is 5.

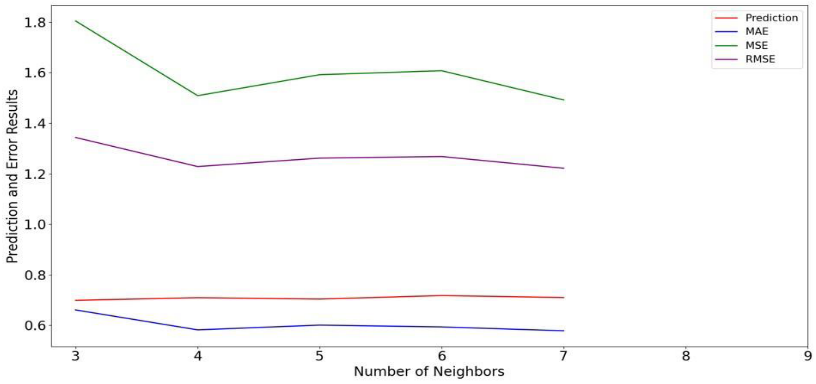


Figure 10. KNN model prediction and margin of error graph.

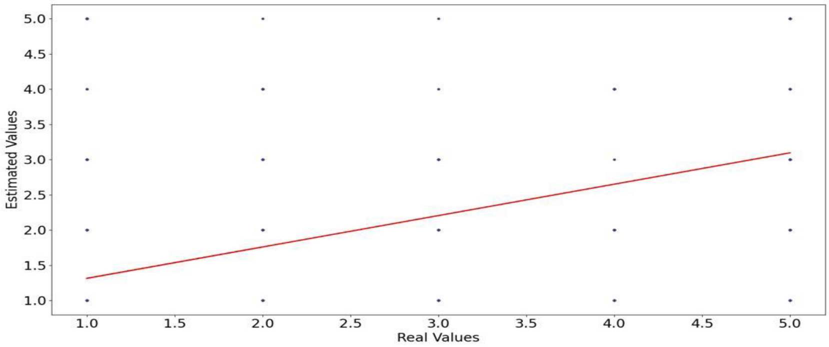


Figure 11. Model graph with 8 neighbors.

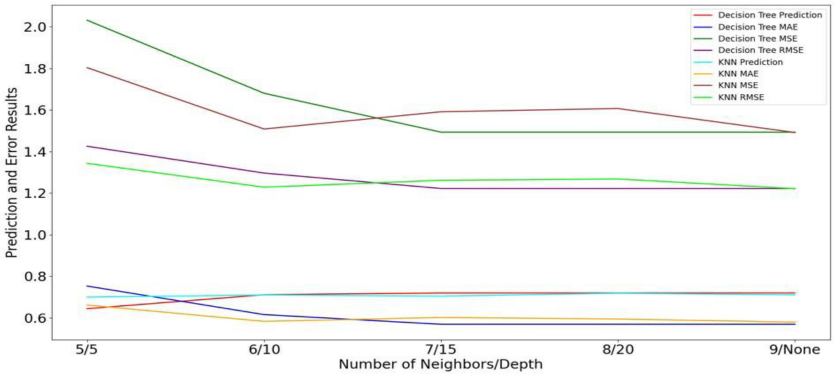


Figure 12. Model result comparison Chart.

values. The graph related to the value results with 8 neighbors is given in Figure 11.

The comparison chart of the results of the two models is given in Figure 12. Since the parameter values of the models are not equal to each other, the values with the number of neighbors of the KNN model less than 5 are not included in the graph. Although the prediction results of the models are almost the same, it is observed that the decision tree model gives a better result with a very small difference.

DISCUSSION

Depending on the developing technologies, there are continuous developments in the field of pathology today. Apart from the development and change of devices and tools used in the field of pathology, the use of computer applications is also increasing and developing. Pathological procedures usually require long-term procedures. Shortening this period may save extra time for the treatment methods to be applied with the diagnosis to be made. Today, a field called digital pathology has emerged that allows the analysis of pathological samples to be transferred to the computer environment, the results can be visualized using various applications, these results can be analyzed, and the results can be stored (Barisoni et al., 2017; Bera et al., 2019; Özmen, 2021).

Advances in the field of pathology enable rapid and accurate diagnosis. For this reason, the technological tools and devices used in the field of pathology and the techniques applied contribute to the rapid and accurate diagnosis. With the advanced computer applications that can be used, it can make a positive contribution to the studies to be done in the field of pathology and provides an opportunity to work more easily. In addition, it is possible to share these studies in the network environment and it is possible to get information from other experts in diagnosis and treatment processes (Niazi et al., 2010; Abels et al., 2019; Chang et al., 2019; McCarty et al., 2006; Özmen, 2021).

Thanks to this study, the data of Burdur Mehmet Akif Ersoy University, Faculty of Veterinary Medicine, Department of Pathology was analyzed using the data examined between 2000 and 2020, and after the analysis, the classification process was carried out. The results of the analysis are given and explained in graphics. Two different methods were used for the classification process and the results obtained from these methods were compared. During the classification process, the data set was divided into two as 30% test set and 70% training set. While obtaining the results of the models, different values were obtained in terms of obtaining the best result by changing the parameters. While 5 different results were obtained by changing the depth parameter in decision tree classification, 9 different results were obtained by changing the number of neighbors parameter in KNN classification. In the decision tree method, better results were obtained as the depth increased, while a constant value was obtained for the number of depths of 15 and above. In the KNN method, when the number of neighbors is taken less than the default value of 5, it is seen that the values are estimated at a lower rate, while the results of the model trained from the number of neighbors of 5 and above are seen to have better values. In the comparison, it was determined that the two models gave approximately 70% results. It was seen that the decision tree method gave better results with a very small difference. As a result of the high value definition of model parameters, the margin of error values gradually decreased and as a result of this decrease, an increase in the estimation results was observed.

CONCLUSION

The results of the analysis carried out in this study are shown in graphics. According to the results of the graph, the ratios of the data such as type, age, city, and gender of the samples to each other are seen. Thanks to these rates, the probability of making a more effective and faster diagnosis increases. The aim of this study is to make a more effective and rapid diagnosis in animal disease diagnoses. The results of the analysis are expected to show what kind of diseases the incoming samples or animals may encounter under certain conditions.

DECLARATIONS

Ethic Approval

Not applicable

Conflict of Interest

The authors declare that they have no competing interests

Author Contribution

Idea, concept, and design: AAŞ, ÖÖ, Vİ

Data collection and analysis: AAŞ, ÖÖ, Vİ

Drafting of the manuscript: AAŞ, AHI, ÖÖ

Critical review: AAŞ, AHI, ÖÖ, Vİ

Acknowledgement

Not applicable

REFERENCES

1. Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M. D., van der Laak, J., Bui, M. M., Vemuri, V. N., Parwani, A. V., Gibbs, J., Agosto-Arroyo, E., Beck, A. H., Kozlowski, C. (2019). Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. The Journal of Pathology, 249(3), 286-294. https://doi.org/10.1002/path.5331

2. Barisoni, L., Gimpel, C., Kain, R., Laurinavicius, A., Bueno, G., Zeng, C., Liu, Z., Schaefer, F., Kretzler, M., Holzman, L. B., Hewitt, S. M. (2017). Digital pathology imaging as a novel platform for standardization and globalization of quantitative nephropathology. Clinical Kidney Journal, 10(2), 176-187. https://doi.org/10.1093/ckj/sfw129

3. Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V., Madabhushi, A. (2019). Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nature Reviews Clinical Oncology, 16, 703-715. https://doi.org/10.1038/s41571-019-0252-y

4. Bounsaythip, C., & Rinta-Runsala, E. (2001). Overview of Data Mining for Customer Behavior Modeling, Research Report TTE1-2001-18, VTT Information Technology, Fin-

land.

5. Nakhleh, R. E., & Volmar, K.E. (2015). Error Reduction and Prevention in Surgical Pathology (2nd Edition), Springer, USA.

6. Carlton, W. W., McGavin, M. D. (1995). Thomson's Special Veterinary Pathology, Mosby-Yearbook, Inc., Missouri, USA

7. Chang, H.Y., Jung, C.K., Woo, J.I., Lee, S., Cho, J., Kim, S.W., Kwak, T., Y. (2019). Artificial intelligence in pathology. Journal of Pathology and Translational Medicine, 53(1), 1-12. https://doi.org/10.4132/jptm.2018.12.16

8. Cheville, N.F. (1999). Introduction to Veterinary Pathology, 2nd Ed. Iowa State University Press, USA.

9. Emel G., & Taşkın Ç. (2005). Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması. Eskişehir Osman Gazi Üniversitesi Sosyal Bilimler Dergisi, 6(2), 221-239.

10. Euclidian Distance. (2021). Öklid Uzaklığı. Retrieved December 23, 2021, from https://tr.wikipedia.org/wiki/Öklid_uzaklığı

11.Jones, T.C., & R.D. Hunt. (1993). Veterinary Pathology, Lea & Febiger, Philadelphia, USA.

12. Kahraman, M.M. (1996). Genel Patoloji Ders Notları, Uludağ Üniversitesi Veteriner Fakültesi, Bursa, Türkiye.

13. McCarty, J., Minsky, M.L., Rochester, N., Shannon, C.E. (2006). A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. AI Magazine, 27(4), 12-14. https://doi.org/10.1609/aimag.v27i4.1904

14. Mean Absolute Error. (2021). Mean Absolute Error. Retrieved December 25, 2021, from https://en.wikipedia.org/wiki/Mean_absolute_error

15. Mean Squared Error. (2021). Mean Squared Error. Retrieved December 25, 2021, from https://en.wikipedia.org/wiki/Mean_squared_error

16. Niazi, M. K. K., Parwani, A. V., Gürcan, M. N. (2019). Digital pathology and artificial intelligence. The Lancet Oncology, 20(5), e253-e261. https://doi.org/10.1016/S1470-2045(19)30154-8

17. Numpy-a. (2021). NumPy. Retrieved December 20, 2021, from https://numpy.org

18. Numpy-b. (2021). NumPy. Retrieved December 20, 2021, from https://tr.wikipedia.org/wiki/NumPy

19. Özmen, Ö. (2006). Veteriner Genel Patoloji Ders Notları, Mehmet Akif Ersoy Üniversitesi Veteriner Fakültesi, Burdur, Türkiye.

20. Özmen, Ö. (2016). 2000-2015 Yılları Arasında Burdur'daki Rutin Patoloji Teşhisleri. VIII. Ulusal Veteriner Patoloji Kongresi, 1-3 Eylül 2016, Samsun, Türkiye.

21. Özmen, Ö. (2021). Veteriner Fakültesi Öğrencilerinin Uygulamalı Patoloji Laboratuvar Eğitimleri ile Bilgi Düzeylerinin Arttırılması ve Çağdaş Yaklaşımlar ile Mesleğe Hazırlanması, Bilimsel Araştırma Projeleri Komisyonu, Mehmet Akif Ersoy Üniversitesi Veteriner Fakültesi, Burdur, Türkiye.

22. Pandas-a. (2021). pandas – Python Data Analysis Library. Retrieved December 21, 2021, from https://pandas.pydata.org

23. Pandas-b. (2021). pandas: powerful Python data analysis toolkit. Retrieved December 21, 2021, from https://github.com/pandas-dev/pandas

24. Pandas-c. (2021). Pandas. Retrieved December 21, 2021, from https://tr.wikipedia.org/wiki/Pandas

25. Pandas-d. (2021). pandas.get_dummies. Retrieved December 23, 2021, from https://pandas.pydata.org/docs/reference/api/ pandas.get_dummies.html

26. Root Mean Squared Error. (2021). Root Mean Squared Error. Retrieved December 25, 2021, from https://en.wikipedia.org/wiki/Root-mean-square_deviation

27. Scikit Learn-a. (2021). sklearn.impute.SimpleImputer. Retrieved December 22, 2021, from, https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html

28. Scikit Learn-b. (2021). sklearn.tree.DecisionTreeClassifier. Retrieved December 23, 2021, from https://scikit-learn.org/stable/modules/ generated/sklearn.tree.DecisionTreeClassifier.html

29. Scikit Learn-c. (2021). sklearn.model_selection.train_test_split. Retrieved December 24, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

30. Scikit Learn-d. (2021). sklearn.metrics.accuracy_score. Retrieved December 26, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

31. Scikit Learn-e. (2021). sklearn.neighbors.KNeighborsClassifier. Retrieved December 24, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier

32. Scikit Learn-f. (2021). Nearest Neighbors. Retrieved December 23, 2021, from https://scikit-learn.org/stable/modules/neighbors.html

33. Scikit Learn-g. (2021). sklearn.metrics.mean_squared_error. Retrieved December 24, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html

34. Scikit Learn-h. (2021). sklearn.metrics.mean_absolute_error. Retrieved December 24, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html

35. Slauson, D.O., Cooper, B.J. (1990). Mechanisms of Disease A Textbook of Comparative General Pathology, 2nd Ed., Williams & Wilkins.

36. Sütcü C., & Aytekin Ç. (2018). Veri Bilimi, Paloma Yayınevi, Istanbul, Türkiye.

37. Zorman, M., Vili, P., Kokol, P., Peterson, M., Sprogar, M., Ojstersek, M. (2001). Finding The Right Decision Tree's Induction Strategy for a Hard Real World Problem, International Journal of Medical Informatics, 63(1-2), 109-121. https://doi.org/10.1016/S1386-5056(01)00176-9