Turkish Character Usage in Text Classification

Ali Aycan KOLUKISA [1]

[1] Graduate of Izmir Katip Celebi University, Graduate School of Natural and Applied Sciences, Software Engineering, Turkey
MSc student of Dokuz Eylul University, Institute of Social Sciences, Management Information Systems, Turkey

Abstract

This study is prepared to examine the effects of Turkish character usage on text data by using multiple classifiers. Regression Classifiers, SVM, NB-Classifiers, and ANN are frequently used in supervised learning methods, especially in classification problems. Regression classifiers generally come in two types: as Linear and Logistic. There are also more than one type of Naive Bayes classifier. In our study, after mentioning the properties of Linear Regression and Logistic Regression classifiers in general terms, why Logistic Regression is much more suitable for this study is explained. Then, with the usage of "Logistic Regression", "LinearSVC", "MultinomialNB", "ComplementNB", "BernoulliNB" and "Perceptron" classifiers, the analyzing part starts. Our datasets consist of abstracts-parts from 64 Turkish articles, which have 4 different classes as Physical Sciences, Social Sciences, Educational Sciences, and Economics Administrative Sciences. The data files are all in CSV file format, however, two different data files were prepared. One with original Turkish characters, and the other with its English equivalent formation targeting the Turkish characters "Ç, ç, Ö, ö, Ü, ü, Ş, ş, İ, ı, ğ". In its English-like equivalent file, these were replaced with "C, c, O, o, U, u, S, s, I, i, g" respectively.

Keywords: Accuracy rate; bag of words; English characters; logistic regression; Turkish characters.

1. Introduction

As it is known, Regression Classifiers, SVM, NB-Classifiers, and ANN are frequently used in supervised learning methods, especially in classification problems. Regression classifiers generally come in two types as Linear and Logistic. It is possible to mention that these two classifiers have some positive and negative aspects according to their characteristics. It is also seen that there are more than one type of Naive Bayes classifier. In our study, firstly, after mentioning the properties of Linear Regression and Logistic Regression classifiers in general terms, it is explained why the Logistic Regression classifier is much more suitable for this study. Afterward, the analyzing part takes place with the usage of Logistic Regression, "LinearSVC", "MultinomialNB", "ComplementNB", "BernoulliNB" classifiers, and "Perceptron" classifiers.

Our datasets consist of abstracts-parts from 64 Turkish articles, which have 4 different class-labels such as Physical Sciences (= FEN), Social Sciences (= Sosyal), Educational Sciences (= Egitim), and Economics and Administrative Sciences (= IIBF). In collecting the data, 4 different journals have been used for each class label and 4 articles have been taken from each journal. The journal names used in this study will be given at the end of this paper. The data files have been prepared in CSV file format. And we have prepared two different types of data files. One with the original Turkish characters, and the other one with its English equivalent formation. In the second one, we have changed the original characters of the Turkish language "Ç, ç, Ö, ö, Ü, ü, Ş, ş, İ, ı, ğ" into its English-like equivalents "C, c, O, o, U, u, S, s, I, i, g" respectively. Hereby, two different-named data files, which can be regarded the same in terms of their contents but differ in the use of Turkish characters, have been made ready for accuracy analysis by the above-mentioned classifiers.

2. Classifiers Used for Text Classification

The main classifier supposed to be used in this study is Logistic Regression. However, also other classifiers are added to the study to be able to see how the other classifiers act with the same datasets.

2.1. Logistic Regression Classifier

Regression analysis is an analysis method used to examine the effect or effects of one or more independent variables on a dependent variable [1]. On the other hand, when we look at the working principles of Regression classifiers, it is seen that generally two types of results can be obtained depending on more than one variable. These results, which are generally confused as 0 - 1, are encountered especially in linear type regression classifiers. However, this is a disadvantage of linear regression classifiers as it is possible for output categories to take values between 0 and 1, such as 0.8 or 0.4. It is generally seen that these problems are overcome by setting the threshold

*Corresponding author e-mail address: kolukisa.aycan@gmail.com

Received: 12/07/2021; Accepted: 24/08/2021

value. On the other hand, if the desired results due to more than one variable are desired to be higher than 0 and 1, logistic regression is preferred because linear regression is seen to be insufficient.

At first glance, it can be assumed that logistic regression classifiers operate like linear regression classifiers. However, it appears that there are subtypes of logistic regression classifiers that can adapt to more than one output. These are the Binary Logistic Regression, Multinomial Logistic Regression, and Ordinal Logistic Regression classifiers [2]. In this way, it can give more stable results than Linear Regression.

On the other hand, when compared with the Linear Regression Classifier, there are differences in terms of Cost Function. While in Linear Regression, algorithms such as Mean Square Error, Mean Absolute Error, and Root Mean Square Error are used as cost functions, these algorithms cause various irregularities when applied in Logistic Regression [3]. For this reason, Softmax Function, which can sometimes be named as Logistics Cost Function, is generally used in Logistic Regression. On the other hand, it can be seen that due to the existing similarities of the Softmax function, it is also considered as the general form of the sigmoid function used in probability calculations on binary variables [4]. However, since the Softmax Function, which is mostly used in multiple classification problems, is a non-linear classifier [5], it takes the input data in the layer preceding it and determines which class these inputs are closer to, unlike linear regression classifiers that can distinguish with a single line, by making probability calculations [6].

Therefore, considering the above reasons, it would be appropriate to say that it would be more appropriate to prefer Logistic Regression since there are 4 different class labels of the datasets used in the study.

2.2. Other Classifiers

It is seen that statistical methods such as Regression, Logistic Regression, Time Series Analysis, and Bayesian approaches are generally used in classification problems [7]. In addition to the Logistic Regression classifier, "LinearSVC", "MultinomialNB", "ComplementNB", "BernoulliNB", and Perceptron classifiers are also used to be able to see how other classifiers act with the same datasets.

3. Datasets

The datasets in this study were generally prepared using academic journals in Turkish that have open access on the DergiPark[1] website. A ready-to-use dataset was not employed. A total of 64 articles were used. These articles have 4 different class tags. These are in the form of Science (FEN), Social Sciences (Sosyal), Educational Sciences (Egitim), and Economic and Administrative Sciences (IIBF), respectively, and articles in field journals have been used. 4 different journals were used for each field, and 4 articles(abstract parts only) were taken from each journal. In order to classify the articles, the abstract parts were taken and recorded in the data file. Although each article abstract consists of many sentences, it constitutes only 1 sample of data in the study. Therefore, there are 64 article abstracts belonging to 4 different classes in total, and there are 16 article abstracts in each class, although their lengths differ. These datasets were saved in the form of a CSV file with the name "Makale4x16(tr)" for original Turkish characters. And then in the same file, the Turkish specific characters "Ç, ç, Ö, ö, Ü, ü, Ş, ş, İ, ı, ğ" were determined and changed into their English equivalents "C, c, O, o, U, u, S, s, I, i, g" and saved as a different CSV file named as "Makale4x16". Therefore, 2 datasets consisting of 16x4 = 64 article abstracts for each, whose contents and word numbers and sequences are exactly the same, but differ only in terms of the use of Turkish characters, were made ready for analysis. The general distribution of these datasets used in the study is as follows.

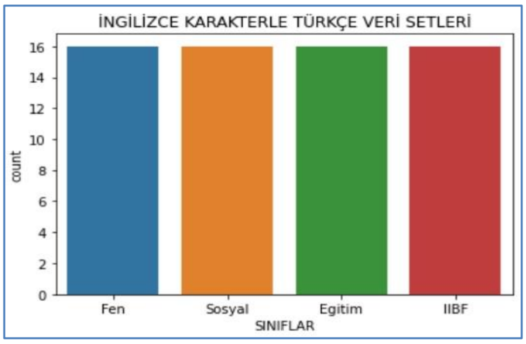


Figure 1. *English-equivalent Formation*

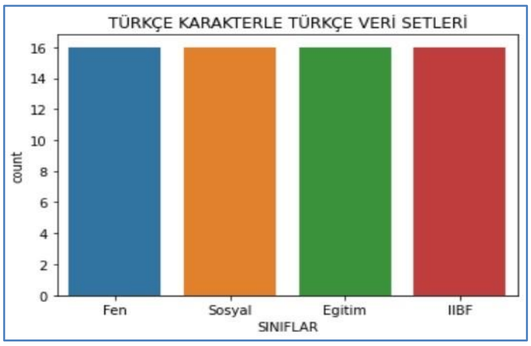


Figure 2. *Original Turkish Characters*

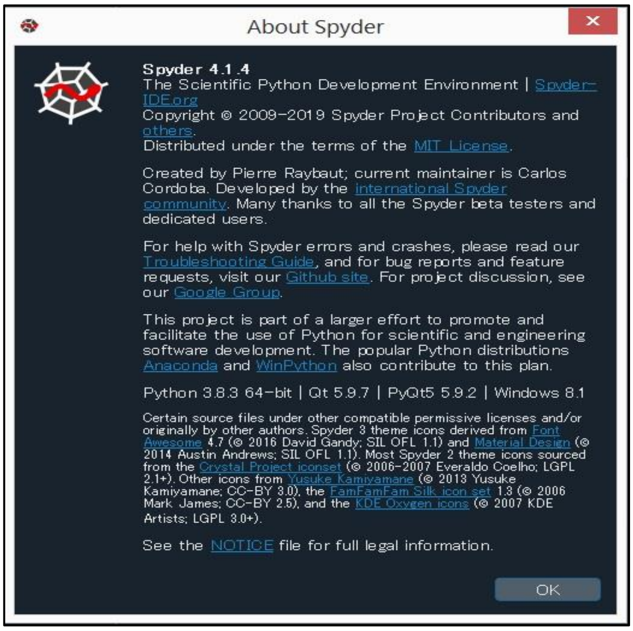
[1] https://dergipark.org.tr/tr/

As it is understood from Figures 1 and 2, the datasets have exactly the same qualification, except the Turkish character usage in the text parts of them.

4. Tools and Environment

To be able to analyze the datasets, PYTHON codes are preferred. The operating system environment is 64-bit Windows 8.1 with 10 GB of RAM - Intel Celeron 2957U@1.4 GHz. In order to run the PYTHON codes, the SPYDER interface (Figure 4) that comes with ANACONDA is preferred.

The libraries used in the application such as Pandas, Scikit-learn, Seaborn, etc. were loaded first into SPYDER via the ANACONDA command line (CMD) before the operation. The version information of SPYDER is 4.1.4 and PYTHON version used in this study is 3.8.3 (64-bit) as seen. And to be able to analyze the files, two different CSV files were loaded by using the pandas library command, as pd.read_csv('Makale4x16(tr).csv') and pd.read_csv('Makale4x16.csv').



About Spyder

Spyder 4.1.4
The Scientific Python Development Environment | Spyder-IDE.org
Copyright © 2009-2019 Spyder Project Contributors and others.
Distributed under the terms of the MIT License.

Created by Pierre Raybaut; current maintainer is Carlos Cordoba. Developed by the international Spyder community. Many thanks to all the Spyder beta testers and dedicated users.

For help with Spyder errors and crashes, please read our Troubleshooting Guide, and for bug reports and feature requests, visit our Github site. For project discussion, see our Google Group.

This project is part of a larger effort to promote and facilitate the use of Python for scientific and engineering software development. The popular Python distributions Anaconda and WinPython also contribute to this plan.

Python 3.8.3 64-bit | Qt 5.9.7 | PyQt5 5.9.2 | Windows 8.1

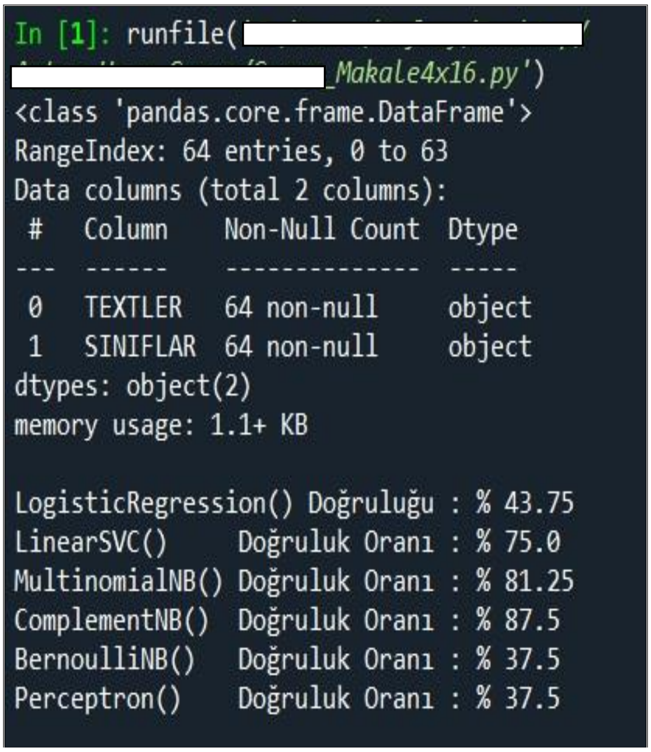
Certain source files under other compatible permissive licenses and/or originally by other authors. Spyder 3 theme icons derived from Font Awesome 4.7 (© 2016 David Gandy; SIL OFL 1.1) and Material Design (© 2014 Austin Andrews; SIL OFL 1.1). Most Spyder 2 theme icons sourced from the Crystal Project iconset (© 2006-2007 Everaldo Coelho; LGPL 2.1+). Other icons from Yusuke Kamiyamane (© 2013 Yusuke Kamiyamane; CC-BY 3.0), the FamFamFam Silk icon set 1.3 (© 2006 Mark James; CC-BY 2.5), and the KDE Oxygen icons (© 2007 KDE Artists; LGPL 3.0+).

See the NOTICE file for full legal information.

Figure 3. *Spyder Environment*

5. Operation and Analysis

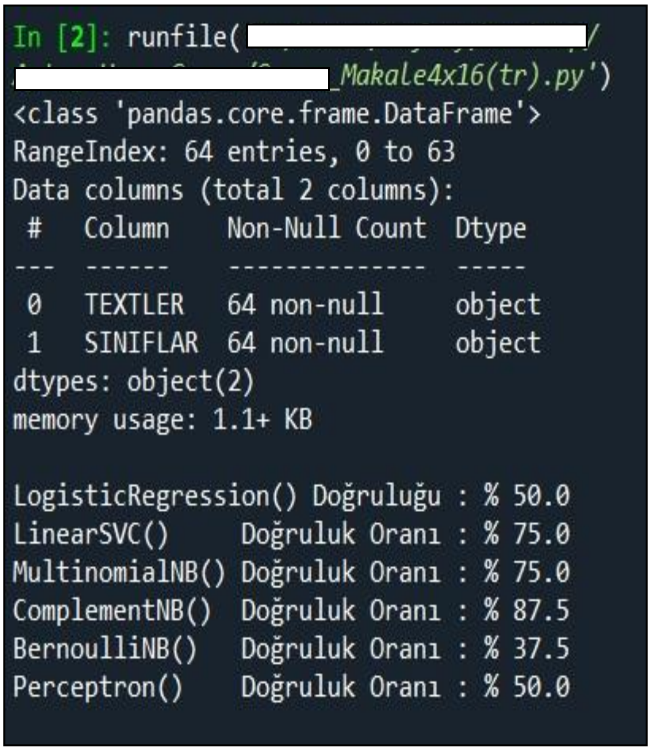
In our application, the model preparation was done first. The necessary libraries were imported and included in the application, then the modeling of Logistic Regression and other classifiers were created with Python codes. First, "Makale4x16.csv" file has been prepared to be subjected to Logistic Regression analysis. Then, the models of "LinearSVC", "MultinomialNB", "ComplementNB", "BernoulliNB" and "Perceptron" classifiers were also created and added with Python codes. However, since it is not possible for the machine to directly read the string type (textual) data, first of all, this data is converted to numerical form. For this, two methods were used. In the Logistic Regression model, these textual data were converted into numerical form by using TfidfVectorizer. For the other classifiers, the Bag of Words (BOW) model was prepared by using the CountVectorizer. Thus all the textual data was transformed into numerical data so that the machine can understand. After our models and codes were made ready for all the classifiers, the analysis phase started. First, the dataset with English characters was analyzed (Figure 4).



```
In [1]: runfile([redacted]/
[redacted]_Makale4x16.py')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64 entries, 0 to 63
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   TEXTLER   64 non-null     object
 1   SINIFLAR  64 non-null     object
dtypes: object(2)
memory usage: 1.1+ KB

LogisticRegression() Doğruluğu : % 43.75
LinearSVC()     Doğruluk Oranı : % 75.0
MultinomialNB() Doğruluk Oranı : % 81.25
ComplementNB()  Doğruluk Oranı : % 87.5
BernoulliNB()   Doğruluk Oranı : % 37.5
Perceptron()    Doğruluk Oranı : % 37.5
```

Figure 4. *Results of English-equivalent Formation*



```
In [2]: runfile([redacted]/
[redacted]_Makale4x16(tr).py')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64 entries, 0 to 63
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   TEXTLER   64 non-null     object
 1   SINIFLAR  64 non-null     object
dtypes: object(2)
memory usage: 1.1+ KB

LogisticRegression() Doğruluğu : % 50.0
LinearSVC()     Doğruluk Oranı : % 75.0
MultinomialNB() Doğruluk Oranı : % 75.0
ComplementNB()  Doğruluk Oranı : % 87.5
BernoulliNB()   Doğruluk Oranı : % 37.5
Perceptron()    Doğruluk Oranı : % 50.0
```

Figure 5. *Results of Original Turkish Characters*

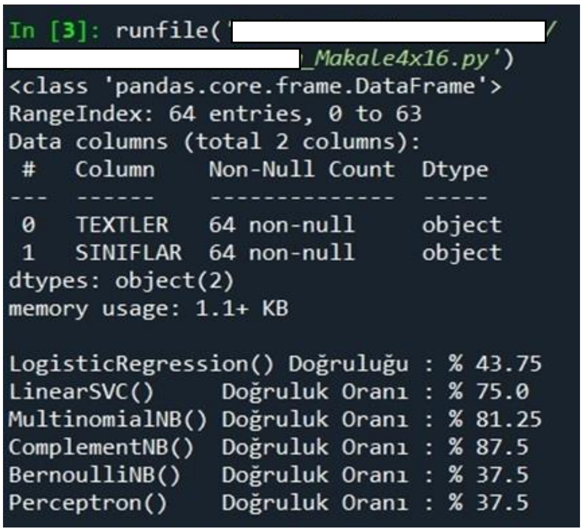
As seen in Figure 4, the accuracy rate of Logistic Regression was 43.75%, Bernoulli Naive Bayes was 37.5% and Perceptron was 37.5%, and as understood, the accuracy rates of these three classifiers were generally below 50%. On the other hand, Linear Support Vector Machine achieved an accuracy rate of 75%, Multinomial Naive Bayes was 81.25%, and Complement Naive Bayes was 87.5%, achieving an overall success rate of 75% and above. Now, with the above codes, let's replace only the part of the file to be analyzed with "Makale4x16 (tr).csv". Here, It is better to mention again that no changes have been made to the codes and they are the same as in the previous section.

This time, "Makale4x16(tr).csv" file - with Turkish original characters, was analyzed and the following results were obtained. As it is seen from the results at Figure 5, although the basic contents of our two files are the same, the detail has increased with the use of Turkish characters, and the machine classifiers have given a certain reaction to this. This time, the accuracy rate of Perceptron and Logistic Regression classifiers increased within the Turkish original character dataset file analysis and reached to 50%. However, "MultinomialNB" decreased to 75%. There has been no change in the accuracy rates of "LinearSVC", "ComplementNB" and "BernoulliNB" classifiers.

So, what does this situation tell us? Based on the above results, the use of Turkish characters increases the in-text details to a certain extent, and accordingly, Perceptron, which is the most basic and simple form of deep learning algorithms, can draw new teachings from this change, and it is not only limited to Perceptron, it can be said that it also makes sense for the Logistic Regression classifier.

6. Conclusion

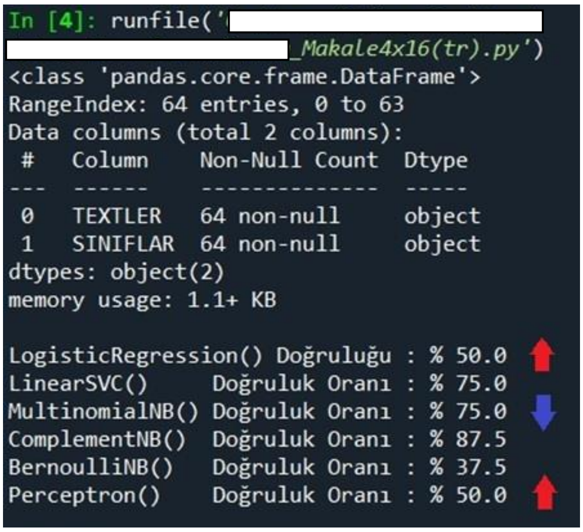
In this study, Turkish article abstracts with the same content were prepared in 2 different CSV files with Turkish and English characters, and the effects of using Turkish characters on machine learning were examined. Although the contents of both files are the same, the results show that Perceptron and Logistic Regression classifiers, which are frequently used in deep learning, have a positive response to the Turkish characters, while Multinomial Naive Bayes has a negative response under the same conditions. On the other hand, "LinearSVC", "ComplementNB" and "BernoulliNB" classifiers show no reaction to the use of Turkish characters. The results are summarized in Figures 6 and 7.



```
In [3]: runfile('[redacted]/
[redacted]_Makale4x16.py')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64 entries, 0 to 63
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   TEXTLER   64 non-null     object
 1   SINIFLAR  64 non-null     object
dtypes: object(2)
memory usage: 1.1+ KB

LogisticRegression() Doğruluğu : % 43.75
LinearSVC()     Doğruluk Oranı : % 75.0
MultinomialNB() Doğruluk Oranı : % 81.25
ComplementNB()  Doğruluk Oranı : % 87.5
BernoulliNB()   Doğruluk Oranı : % 37.5
Perceptron()    Doğruluk Oranı : % 37.5
```

Figure 6. *Results of English-equivalent Formation*



```
In [4]: runfile('[redacted]
[redacted]_Makale4x16(tr).py')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64 entries, 0 to 63
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   TEXTLER   64 non-null     object
 1   SINIFLAR  64 non-null     object
dtypes: object(2)
memory usage: 1.1+ KB

LogisticRegression() Doğruluğu : % 50.0 ⬆
LinearSVC()     Doğruluk Oranı : % 75.0
MultinomialNB() Doğruluk Oranı : % 75.0 ⬇
ComplementNB()  Doğruluk Oranı : % 87.5
BernoulliNB()   Doğruluk Oranı : % 37.5
Perceptron()    Doğruluk Oranı : % 50.0 ⬆
```

Figure 7. *Results of Original Turkish Characters*

Declaration of Interest

The authors declare that there is no conflict of interest.

References

[1] S. Alp and E. Öz, Makine Öğrenmesinde Sınıflandırma Yöntemleri ve R Uygulamaları. Nobel Akademik Yayıncılık, 2019.

[2] H.B. Akın and E. Şentürk, "Bireylerin Mutluluk Düzeylerinin Ordinal Lojistik Regresyon Analizi ile İncelenmesi", Öneri Dergisi, vol. 10, no. 37, 183-193, 2012.

[3] S. Swaminatah, "Logistic Regression- Detailed Owerview. Towards Data Science." towardsdatascience.com, 2018 [Online]. Available: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc, 2018. [Accessed: Jan. 15, 2021]

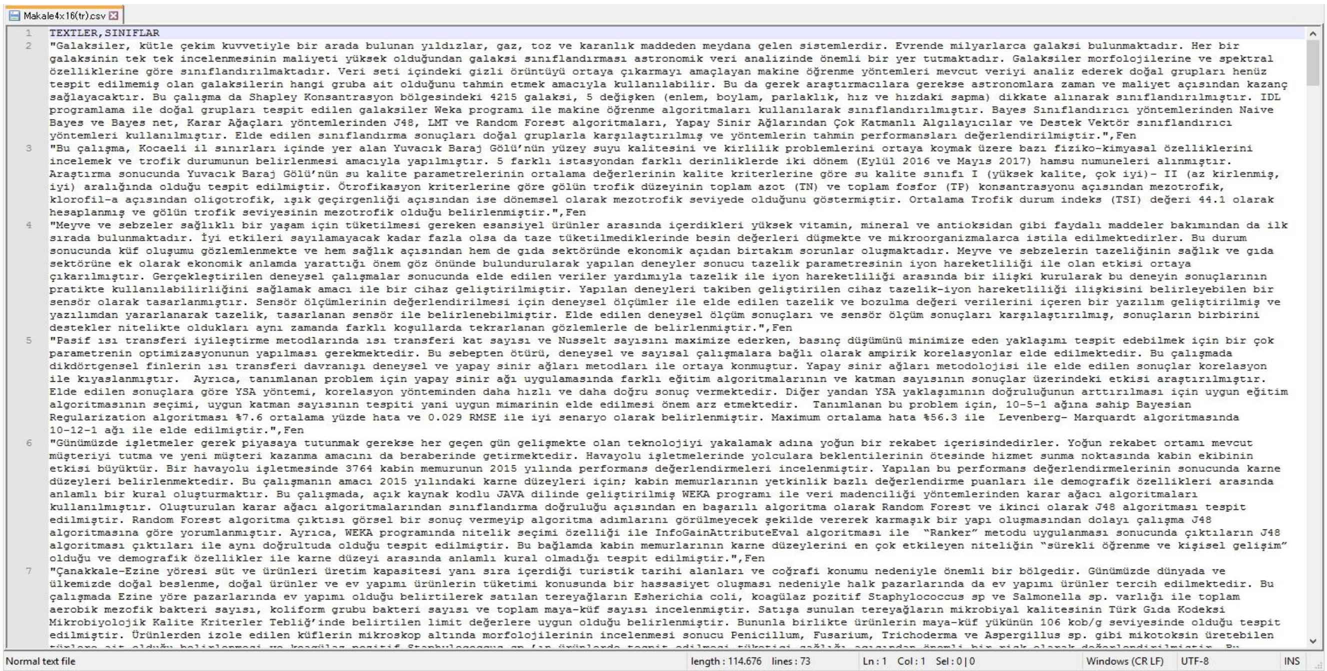
[4] Ö. Şahin, "iOS platformunda görme engelliler için TL tanıma uygulaması" Yüksek Lisans Tezi, T.C. Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya, 73, 2017.

[5] B. Aleksey, "Linear and non-linear activation, and softmax." Kaggle.com, 2018 [Online]. Available:https://www.kaggle.com/residentmario/linear-and-non-linear-activation-and-softmax, 2018. [Accessed: Jan. 13, 20201]

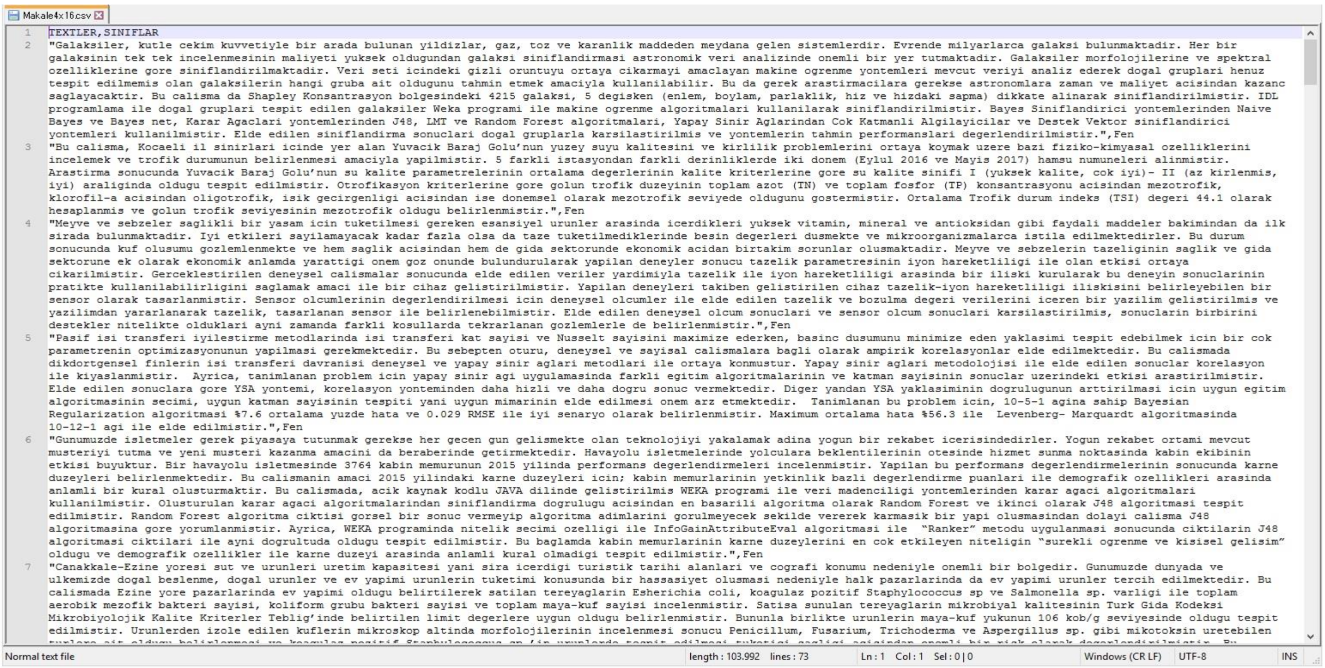
[6] F. Doğan and İ. Türkoğlu, "Derin Öğrenme Modelleri ve Uygulama Alanlarına İlişkin Bir Derleme", Dicle Üniversitesi Mühendislik Fakültesi DÜMF Dergisi, vol. 10, no. 2, 409-445, 2019.

[7] G. Silahtaroğlu, Veri Madenciliği Yöntemleri. Papatya Yayıncılık, İstanbul, 2013.

APPENDIX (parts of the datasets used)



```
TEXTLER,SINIFLAR
"Galaksiler, kütle çekim kuvvetiyle bir arada bulunan yıldızlar, gaz, toz ve karanlık maddeden meydana gelen sistemlerdir. Evrende milyarlarca galaksi bulunmaktadır. Her bir
galaksinin tek tek incelenmesinin maliyeti yüksek olduğundan galaksi sınıflandırması astronomik veri analizinde önemli bir yer tutmaktadır. Galaksiler morfolojilerine ve spektral
özelliklerine göre sınıflandırılmaktadır. Veri seti içindeki gizli örüntüyü ortaya çıkarmayı amaçlayan makine öğrenme yöntemleri mevcut veriyi analiz ederek doğal grupları henüz
tespit edilmemiş olan galaksilerin hangi gruba ait olduğunu tahmin etmek amacıyla kullanılabilir. Bu da gerek araştırmacılara gerekse astronomlara zaman ve maliyet açısından kazanç
sağlayacaktır. Bu çalışma da Shapley Konsantrasyon bölgesindeki 4215 galaksi, 5 değişken (enlem, boylam, parlaklık, hız ve hızdaki sapma) dikkate alınarak sınıflandırılmıştır. IDL
programlama ile doğal grupları tespit edilen galaksiler Weka programı ile makine öğrenme algoritmaları kullanılarak sınıflandırılmıştır. Bayes Sınıflandırıcı yöntemlerinden Naive
Bayes ve Bayes net, Karar Ağaçları yöntemlerinden J48, LMT ve Random Forest algoritmaları, Yapay Sinir Ağlarından Çok Katmanlı Algılayıcılar ve Destek Vektör sınıflandırıcı
yöntemleri kullanılmıştır. Elde edilen sınıflandırma sonuçları doğal gruplarla karşılaştırılmış ve yöntemlerin tahmin performansları değerlendirilmiştir.",Fen
"Bu çalışma, Kocaeli il sınırları içinde yer alan Yuvacık Baraj Gölü'nün yüzey suyu kalitesini ve kirlilik problemlerini ortaya koymak üzere bazı fiziko-kimyasal özelliklerini
incelemek ve trofik durumunun belirlenmesi amacıyla yapılmıştır. 5 farklı istasyondan farklı derinliklerde iki dönem (Eylül 2016 ve Mayıs 2017) hamsu numuneleri alınmıştır.
Araştırma sonucunda Yuvacık Baraj Gölü'nün su kalite parametrelerinin ortalama değerlerinin kalite kriterlerine göre su kalite sınıfı I (yüksek kalite, çok iyi)- II (az kirlenmiş,
iyi) aralığında olduğu tespit edilmiştir. Ötrofikasyon kriterlerine göre gölün trofik düzeyinin toplam azot (TN) ve toplam fosfor (TP) konsantrasyonu açısından mezotrofik,
klorofil-a açısından oligotrofik, ışık geçirgenliği açısından ise dönemsel olarak mezotrofik seviyede olduğunu göstermiştir. Ortalama Trofik durum indeks (TSI) değeri 44.1 olarak
hesaplanmış ve gölün trofik seviyesinin mezotrofik olduğu belirlenmiştir.",Fen
"Meyve ve sebzeler sağlıklı bir yaşam için tüketilmesi gereken esansiyel ürünler arasında içerdikleri yüksek vitamin, mineral ve antioksidan gibi faydalı maddeler bakımından da ilk
sırada bulunmaktadır. İyi etkileri sayılamayacak kadar fazla olsa da taze tüketilmediklerinde besin değerleri düşmekte ve mikroorganizmalarca istila edilmektedirler. Bu durum
sonucunda küf oluşumu gözlemlenmekte ve hem sağlık açısından hem de gıda sektöründe ekonomik açıdan birtakım sorunlar oluşmaktadır. Meyve ve sebzelerin tazeliğinin sağlık ve gıda
sektörüne ek olarak ekonomik anlamda yarattığı önem göz önünde bulundurularak yapılan deneyler sonucu tazelik parametresinin iyon hareketliliği ile olan etkisi ortaya
çıkarılmıştır. Gerçekleştirilen deneysel çalışmalar sonucunda elde edilen veriler yardımıyla tazelik ile iyon hareketliliği arasında bir ilişki kurularak bu deneyin sonuçlarının
pratikte kullanılabilirliğini sağlamak amacı ile bir cihaz geliştirilmiştir. Yapılan deneyleri takiben geliştirilen cihaz tazelik-iyon hareketliliği ilişkisini belirleyebilen bir
sensör olarak tasarlanmıştır. Sensör ölçümlerinin değerlendirilmesi için deneysel ölçümler ile elde edilen tazelik ve bozulma değeri verilerini içeren bir yazılım geliştirilmiş ve
yazılımdan yararlanarak tazelik, tasarlanan sensör ile belirlenebilmiştir. Elde edilen deneysel ölçüm sonuçları ve sensör ölçüm sonuçları karşılaştırılmış, sonuçların birbirini
destekler nitelikte oldukları aynı zamanda farklı koşullarda tekrarlanan gözlemlerle de belirlenmiştir.",Fen
"Pasif ısı transferi iyileştirme metodlarında ısı transferi kat sayısı ve Nusselt sayısını maximize ederken, basınç düşümünü minimize eden yaklaşımı tespit edebilmek için bir çok
parametrenin optimizasyonunun yapılması gerekmektedir. Bu sebepten ötürü, deneysel ve sayısal çalışmalara bağlı olarak ampirik korelasyonlar elde edilmektedir. Bu çalışmada
dikdörtgensel finlerin ısı transferi davranışı deneysel ve yapay sinir ağları metodları ile ortaya konmuştur. Yapay sinir ağları metodolojisi ile elde edilen sonuçlar korelasyon
ile kıyaslanmıştır.  Ayrıca, tanımlanan problem için yapay sinir ağı uygulamasında farklı eğitim algoritmalarının ve katman sayısının sonuçlar üzerindeki etkisi araştırılmıştır.
Elde edilen sonuçlara göre YSA yöntemi, korelasyon yönteminden daha hızlı ve daha doğru sonuç vermektedir. Diğer yandan YSA yaklaşımının doğruluğunun arttırılması için uygun eğitim
algoritmasının seçimi, uygun katman sayısının tespiti yani uygun mimarinin elde edilmesi önem arz etmektedir.  Tanımlanan bu problem için, 10-5-1 ağına sahip Bayesian
Regularization algoritması %7.6 ortalama yüzde hata ve 0.029 RMSE ile iyi senaryo olarak belirlenmiştir. Maximum ortalama hata %56.3 ile  Levenberg- Marquardt algoritmasında
10-12-1 ağı ile elde edilmiştir.",Fen
"Günümüzde işletmeler gerek piyasaya tutunmak gerekse her geçen gün gelişmekte olan teknolojiyi yakalamak adına yoğun bir rekabet içerisindedirler. Yoğun rekabet ortamı mevcut
müşteriyi tutma ve yeni müşteri kazanma amacını da beraberinde getirmektedir. Havayolu işletmelerinde yolculara beklentilerinin ötesinde hizmet sunma noktasında kabin ekibinin
etkisi büyüktür. Bir havayolu işletmesinde 3764 kabin memurunun 2015 yılında performans değerlendirmeleri incelenmiştir. Yapılan bu performans değerlendirmelerinin sonucunda karne
düzeyleri belirlenmektedir. Bu çalışmanın amacı 2015 yılındaki karne düzeyleri için; kabin memurlarının yetkinlik bazlı değerlendirme puanları ile demografik özellikleri arasında
anlamlı bir kural oluşturmaktır. Bu çalışmada, açık kaynak kodlu JAVA dilinde geliştirilmiş WEKA programı ile veri madenciliği yöntemlerinden karar ağacı algoritmaları
kullanılmıştır. Oluşturulan karar ağacı algoritmalarından sınıflandırma doğruluğu açısından en başarılı algoritma olarak Random Forest ve ikinci olarak J48 algoritması tespit
edilmiştir. Random Forest algoritma çıktısı görsel bir sonuç vermeyip algoritma adımlarını görülmeyecek şekilde vererek karmaşık bir yapı oluşmasından dolayı çalışma J48
algoritmasına göre yorumlanmıştır. Ayrıca, WEKA programında nitelik seçimi özelliği ile InfoGainAttributeEval algoritması ile  "Ranker" metodu uygulanması sonucunda çıktıların J48
algoritması çıktıları ile aynı doğrultuda olduğu tespit edilmiştir. Bu bağlamda kabin memurlarının karne düzeylerini en çok etkileyen niteliğin "sürekli öğrenme ve kişisel gelişim"
olduğu ve demografik özellikler ile karne düzeyi arasında anlamlı kural olmadığı tespit edilmiştir.",Fen
"Çanakkale-Ezine yöresi süt ve ürünleri üretim kapasitesi yanı sıra içerdiği turistik tarihi alanları ve coğrafi konumu nedeniyle önemli bir bölgedir. Günümüzde dünyada ve
ülkemizde doğal beslenme, doğal ürünler ve ev yapımı ürünlerin tüketimi konusunda bir hassasiyet oluşması nedeniyle halk pazarlarında da ev yapımı ürünler tercih edilmektedir. Bu
çalışmada Ezine yöre pazarlarında ev yapımı olduğu belirtilerek satılan tereyağların Esherichia coli, koagülaz pozitif Staphylococcus sp ve Salmonella sp. varlığı ile toplam
aerobik mezofik bakteri sayısı, koliform grubu bakteri sayısı ve toplam maya-küf sayısı incelenmiştir. Satışa sunulan tereyağların mikrobiyal kalitesinin Türk Gıda Kodeksi
Mikrobiyolojik Kalite Kriterler Tebliğ'inde belirtilen limit değerlere uygun olduğu belirlenmiştir. Bununla birlikte ürünlerin maya-küf yükünün 106 kob/g seviyesinde olduğu tespit
edilmiştir. Ürünlerden izole edilen küflerin mikroskop altında morfolojilerinin incelenmesi sonucu Penicillum, Fusarium, Trichoderma ve Aspergillus sp. gibi mikotoksin üretebilen
```



```
TEXTLER,SINIFLAR
"Galaksiler, kutle cekim kuvvetiyle bir arada bulunan yildizlar, gaz, toz ve karanlik maddeden meydana gelen sistemlerdir. Evrende milyarlarca galaksi bulunmaktadir. Her bir
galaksinin tek tek incelenmesinin maliyeti yuksek oldugundan galaksi siniflandirmasi astronomik veri analizinde onemli bir yer tutmaktadir. Galaksiler morfolojilerine ve spektral
ozelliklerine gore siniflandirilmaktadir. Veri seti icindeki gizli oruntuyu ortaya cikarmayi amaclayan makine ogrenme yontemleri mevcut veriyi analiz ederek dogal gruplari henuz
tespit edilmemis olan galaksilerin hangi gruba ait oldugunu tahmin etmek amaciyla kullanilabilir. Bu da gerek arastirmacilara gerekse astronomlara zaman ve maliyet acisindan kazanc
saglayacaktir. Bu calisma da Shapley Konsantrasyon bolgesindeki 4215 galaksi, 5 degisken (enlem, boylam, parlaklik, hiz ve hizdaki sapma) dikkate alinarak siniflandirilmistir. IDL
programlama ile dogal gruplari tespit edilen galaksiler Weka programi ile makine ogrenme algoritmalari kullanilarak siniflandirilmistir. Bayes Siniflandirici yontemlerinden Naive
Bayes ve Bayes net, Karar Agaclari yontemlerinden J48, LMT ve Random Forest algoritmalari, Yapay Sinir Aglarindan Cok Katmanli Algilayicilar ve Destek Vektor siniflandirici
yontemleri kullanilmistir. Elde edilen siniflandirma sonuclari dogal gruplarla karsilastirilmis ve yontemlerin tahmin performanslari degerlendirilmistir.",Fen
"Bu calisma, Kocaeli il sinirlari icinde yer alan Yuvacik Baraj Golu’nun yuzey suyu kalitesini ve kirlilik problemlerini ortaya koymak uzere bazi fiziko-kimyasal ozelliklerini
incelemek ve trofik durumunun belirlenmesi amaciyla yapilmistir. 5 farkli istasyondan farkli derinliklerde iki donem (Eylul 2016 ve Mayis 2017) hamsu numuneleri alinmistir.
Arastirma sonucunda Yuvacik Baraj Golu’nun su kalite parametrelerinin ortalama degerlerinin kalite kriterlerine gore su kalite sinifi I (yuksek kalite, cok iyi)- II (az kirlenmis,
iyi) araliginda oldugu tespit edilmistir. Otrofikasyon kriterlerine gore golun trofik duzeyinin toplam azot (TN) ve toplam fosfor (TP) konsantrasyonu acisindan mezotrofik,
klorofil-a acisindan oligotrofik, isik gecirgenligi acisindan ise donemsel olarak mezotrofik seviyede oldugunu gostermistir. Ortalama Trofik durum indeks (TSI) degeri 44.1 olarak
hesaplanmis ve golun trofik seviyesinin mezotrofik oldugu belirlenmistir.",Fen
"Meyve ve sebzeler saglikli bir yasam icin tuketilmesi gereken esansiyel urunler arasinda icerdikleri yuksek vitamin, mineral ve antioksidan gibi faydali maddeler bakimindan da ilk
sirada bulunmaktadir. Iyi etkileri sayilamayacak kadar fazla olsa da taze tuketilmediklerinde besin degerleri dusmekte ve mikroorganizmalarca istila edilmektedirler. Bu durum
sonucunda kuf olusumu gozlemlenmekte ve hem saglik acisindan hem de gida sektorunde ekonomik acidan birtakim sorunlar olusmaktadir. Meyve ve sebzelerin tazeliginin saglik ve gida
sektorune ek olarak ekonomik anlamda yarattigi onem goz onunde bulundurularak yapilan deneyler sonucu tazelik parametresinin iyon hareketliligi ile olan etkisi ortaya
cikarilmistir. Gerceklestirilen deneysel calismalar sonucunda elde edilen veriler yardimiyla tazelik ile iyon hareketliligi arasinda bir iliski kurularak bu deneyin sonuclarinin
pratikte kullanilabilirligini saglamak amaci ile bir cihaz gelistirilmistir. Yapilan deneyleri takiben gelistirilen cihaz tazelik-iyon hareketliligi iliskisini belirleyebilen bir
sensor olarak tasarlanmistir. Sensor olcumlerinin degerlendirilmesi icin deneysel olcumler ile elde edilen tazelik ve bozulma degeri verilerini iceren bir yazilim gelistirilmis ve
yazilimdan yararlanarak tazelik, tasarlanan sensor ile belirlenebilmistir. Elde edilen deneysel olcum sonuclari ve sensor olcum sonuclari karsilastirilmis, sonuclarin birbirini
destekler nitelikte olduklari ayni zamanda farkli kosullarda tekrarlanan gozlemlerle de belirlenmistir.",Fen
"Pasif isi transferi iyilestirme metodlarinda isi transferi kat sayisi ve Nusselt sayisini maximize ederken, basinc dusumunu minimize eden yaklasimi tespit edebilmek icin bir cok
parametrenin optimizasyonunun yapilmasi gerekmektedir. Bu sebepten oturu, deneysel ve sayisal calismalara bagli olarak ampirik korelasyonlar elde edilmektedir. Bu calismada
dikdortgensel finlerin isi transferi davranisi deneysel ve yapay sinir aglari metodlari ile ortaya konmustur. Yapay sinir aglari metodolojisi ile elde edilen sonuclar korelasyon
ile kiyaslanmistir.  Ayrica, tanimlanan problem icin yapay sinir agi uygulamasinda farkli egitim algoritmalarinin ve katman sayisinin sonuclar uzerindeki etkisi arastirilmistir.
Elde edilen sonuclara gore YSA yontemi, korelasyon yonteminden daha hizli ve daha dogru sonuc vermektedir. Diger yandan YSA yaklasiminin dogrulugunun arttirilmasi icin uygun egitim
algoritmasinin secimi, uygun katman sayisinin tespiti yani uygun mimarinin elde edilmesi onem arz etmektedir.  Tanimlanan bu problem icin, 10-5-1 agina sahip Bayesian
Regularization algoritmasi %7.6 ortalama yuzde hata ve 0.029 RMSE ile iyi senaryo olarak belirlenmistir. Maximum ortalama hata %56.3 ile  Levenberg- Marquardt algoritmasinda
10-12-1 agi ile elde edilmistir.",Fen
"Gunumuzde isletmeler gerek piyasaya tutunmak gerekse her gecen gun gelismekte olan teknolojiyi yakalamak adina yogun bir rekabet icerisindedirler. Yogun rekabet ortami mevcut
musteriyi tutma ve yeni musteri kazanma amacini da beraberinde getirmektedir. Havayolu isletmelerinde yolculara beklentilerinin otesinde hizmet sunma noktasinda kabin ekibinin
etkisi buyuktur. Bir havayolu isletmesinde 3764 kabin memurunun 2015 yilinda performans degerlendirmeleri incelenmistir. Yapilan bu performans degerlendirmelerinin sonucunda karne
duzeyleri belirlenmektedir. Bu calismanin amaci 2015 yilindaki karne duzeyleri icin; kabin memurlarinin yetkinlik bazli degerlendirme puanlari ile demografik ozellikleri arasinda
anlamli bir kural olusturmaktir. Bu calismada, acik kaynak kodlu JAVA dilinde gelistirilmis WEKA programi ile veri madenciligi yontemlerinden karar agaci algoritmalari
kullanilmistir. Olusturulan karar agaci algoritmalarindan siniflandirma dogrulugu acisindan en basarili algoritma olarak Random Forest ve ikinci olarak J48 algoritmasi tespit
edilmistir. Random Forest algoritma ciktisi gorsel bir sonuc vermeyip algoritma adimlarini gorulmeyecek sekilde vererek karmasik bir yapi olusmasindan dolayi calisma J48
algoritmasina gore yorumlanmistir. Ayrica, WEKA programinda nitelik secimi ozelligi ile InfoGainAttributeEval algoritmasi ile  “Ranker” metodu uygulanmasi sonucunda ciktilarin J48
algoritmasi ciktilari ile ayni dogrultuda oldugu tespit edilmistir. Bu baglamda kabin memurlarinin karne duzeylerini en cok etkileyen niteligin “surekli ogrenme ve kisisel gelisim”
oldugu ve demografik ozellikler ile karne duzeyi arasinda anlamli kural olmadigi tespit edilmistir.",Fen
"Canakkale-Ezine yoresi sut ve urunleri uretim kapasitesi yani sira icerdigi turistik tarihi alanlari ve cografi konumu nedeniyle onemli bir bolgedir. Gunumuzde dunyada ve
ulkemizde dogal beslenme, dogal urunler ve ev yapimi urunlerin tuketimi konusunda bir hassasiyet olusmasi nedeniyle halk pazarlarinda da ev yapimi urunler tercih edilmektedir. Bu
calismada Ezine yore pazarlarinda ev yapimi oldugu belirtilerek satilan tereyaglarin Esherichia coli, koagulaz pozitif Staphylococcus sp ve Salmonella sp. varligi ile toplam
aerobik mezofik bakteri sayisi, koliform grubu bakteri sayisi ve toplam maya-kuf sayisi incelenmistir. Satisa sunulan tereyaglarin mikrobiyal kalitesinin Turk Gida Kodeksi
Mikrobiyolojik Kalite Kriterler Teblig’inde belirtilen limit degerlere uygun oldugu belirlenmistir. Bununla birlikte urunlerin maya-kuf yukunun 106 kob/g seviyesinde oldugu tespit
edilmistir. Urunlerden izole edilen kuflerin mikroskop altinda morfolojilerinin incelenmesi sonucu Penicillum, Fusarium, Trichoderma ve Aspergillus sp. gibi mikotoksin uretebilen
```