A PERFORMANCE ANALYSIS OF A DATA SET ACQUIRED BY MEANS OF BOOTSTRAP TECHNIQUE WITH MACHINE LEARNING ALGORITHMS

Tolga Demirhan, Özlem Uçar

Trakya University, Balkans Campus, 22030 Edirne, Turkey

Abstract

A common problem which is encountered in the data analysis phase of scientific studies with special groups is a small sample size. In order to eliminate this problem, resampling techniques, which allow for dependable predictions and replicability of the results with small sample sizes, are used. In this study, large samples acquired through a data set with a small sample size by means of Resampling Excel-add-in macro developed with Bootstrap technique on Excel 2013 were subjected to performance analysis with weka, which is a machine learning tool.

Key words: *bootstrap, weka, hearing-impaired, resampling*

1. INTRODUCTION

A small sample size is an unremitting problem in the analysis of data in academic studies with special groups. According to Sümbüloğlu V, Sümbüloğlu K, an insufficient sample size signifies that the study lacks scientific basis. (Orekkıcı Temel, G., Erdoğan,S., 2011)

Therefore, without regard to how meticulously the scientific research study is conducted, the acquired results may be interpreted as belonging to a special sample and this reduces the value of the study. In that respect, replicating the study with different groups is how the reproducibility of the results can be obtained. However, this method is also not preferred since it necessitates extra time, money and labour in cases which the sample size cannot be increased. In such cases where the sample size is small, resampling techniques, which are also used in the fields of social and medical sciences, allow for the reproducibility of the results without replicating the study and study how to acquire accurate predictions, are used. (Orekkıcı Temel, G., Erdoğan,S., 2012)

Various resampling methods are reported in the literature. Among those, Jackknife and Bootstrap techniques are the most widely preferred prediction techniques especially in the research studies with small samples. (Bekiroğlu,N., Konyalıoğlu, R., Karahan, D.,2013)

Bootstrap is a simple and reliable method which is used in statistical analyses such as standard deviation, confidence interval and nonparametric prediction problems. The formal logic behind the method is based on the creation of new datasets through the random relocation and resampling of the observations within an existing dataset of any size. (Zaman,T., Alakus, K., Bulut,H.,2015)

In this way, as much information as possible is extracted from the dataset. Within the context of the method, a bootstrap sample is created by selecting instances within the original sample through substitution. The confidence interval of the predictor and the statistical significance of the test is computed through these instances.

Although in statistics, the parameters which are investigated are predicted under the assumption of normal distribution frequently, they can be predicted regardless of the assumption of distribution through the extraction of replacement instances from the data set in this method. Even inadequate or small datasets can be used for statistical inferences. (Wehrens R, Putter H, Buydens,2000)

Fu et al. use the bootstrap method in the prediction of misclassification in groups of 20, 30, 50 and 100 observations. In their study, it is concluded that the bootstrap method is more accurate than the other methods. (Fu, W.J., Carrol, R.J. and Wang, S., 2005)

In a study by Heller and Venkatraman (1996), the power of the test is computed as a result of 5000 simulation tests with Monte Carlo simulation technique and 2000 rounds of resampling. In the study,

the data sizes are set to 20:20 and 50:50 and Printice-Wilcoxon, Permutation and Bootstrap tests are run. As a result, it is reported that Bootstrap and Permutation tests are advisable in the cases where the assumption of normality is rejected. (Heller, G., Venkatraman, E.S., 1996)

Tang et al. utilize various resampling methods for bilateral dichotomous data, which is frequently used in modern comparative medical studies. In their study, they evaluate the performances of the Wald-type statistic, Fieller's Theorem, Likelihood Ratio, Score Statistic and Bootstrap Method and conclude that the Bootstrap Method produces reliable results in small samples as much as it does in large samples. Tang et al. create 1000 bootstrap instances out of 8, 15, 30 and 50 instances, as a result, they put forth that the bootstrap method is an interesting model among those which are investigated for small samples. (Nian-Sheng Tang, Shi-Fang, Man-Lai Tang,Yan-Bo Pei Qiu, 2011)

Temel and Erdoğan aim to introduce the resampling methods and to compare these in their study. Frequently encountered limitations in the application phase are the heterogenous distribution within the data and the incapability of the dataset in meeting the distribution assumptions or a small sample size. Reproduction of the same dataset using resampling techniques without damaging the original structure of the data enables the utilization of the data set successfully. Cross-validation and Bootstrap methods do not assume the normal distribution of the data. Moreover, in the Bootstrap method, the selection process of the sample is carried out through replacement unlike the Jackknife and Cross-validation methods. Replacement sampling, which allows the same object or individual to be selected, produces more accurate results in comparison with the non-replacement sampling methods. (Orekkıcı Temel, G., Erdoğan,S., 2011)

Takma and Atıl argues in their study that the Bootstrap method is a simple method to apply, providing information on the use of the method in confidence intervals, hypothesis testing and regression analysis. In the study, 1000 bootstrap instances are created out of 25 observations and it is concluded that the results of the classical calculation and Bootstrap methods produce similar results. The study confirms the superiority of the Bootstrap method over classical methods through the applications of the researchers. Furthermore, Bootstrap method does not have assumptions for the data set. In the classical methods, the results are considered unreliable in the case of a doubt that the assumptions of certain analyses are violated and even in such circumstances, the method produces more reliable solutions. Hence, the Bootstrap method is superior to the analysis of variance in very small datasets or with non-normally distributed data.

Bootstrap method is widely used in the field of applied statistics. The basis of this method is to carry out resampling to create very large datasets out of an existing dataset. Thus, a variety of statistical tests can be run on small samples, too. (Takma,Ç., Atıl,H., 2006)

In his study, Koşkan selects tests with 2 and 3 groups, creates samples from related populations and computes the error probability through simulations with those samples. As a result of the study which has a resampling number of 100000, it is emphasized that the resampling method produces more accurate results in small samples than the other available methods. (Koşkan,Ö., 2007)

In this study, a performance analysis of the large samples acquired through a small dataset via Resampling Excel-add-in Macro of Bootstrap Method, written on Excel 2013, is conducted on Weka, which is a machine learning tool.

2. MATERIALS AND METHODOLOGY

The study was conducted in two stages, the creation of the data set and the administration of the performance analysis.

2.1. The Creation of the Data Set

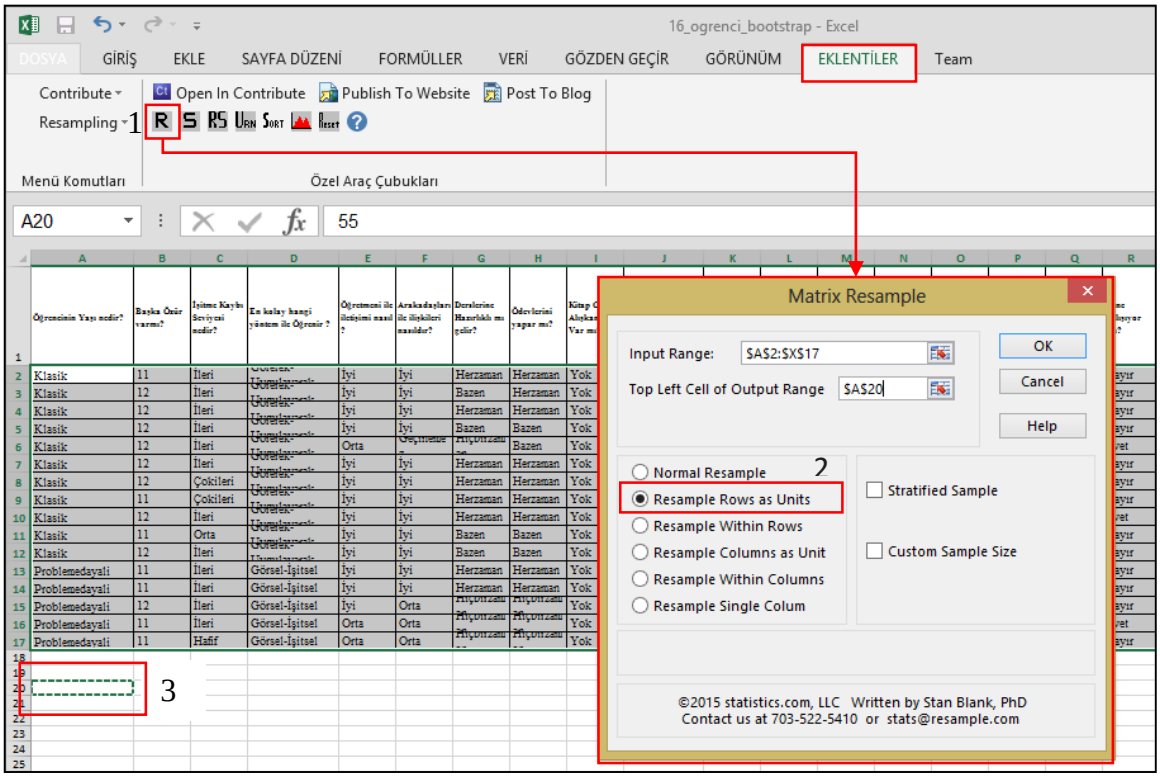
The present study was carried out with 16 volunteer 4th year students in two schools for the deaf in Edirne and İstanbul. Within the context of the study, an achievement test for the "Our Planet Earth"

chapter of the Science and Technology course and a questionnaire were given to the students. Using the Resampling Excel-add-in Macro of the Bootstrap Method, written on Excel 2013, resampling data of 100, 300, 500, 700, 900 and 1000 instances were created.

The creation of 100 instances out of the data set of 16 instances through the Bootstrap Excel Macro for resampling is explained above with pictures step by step.

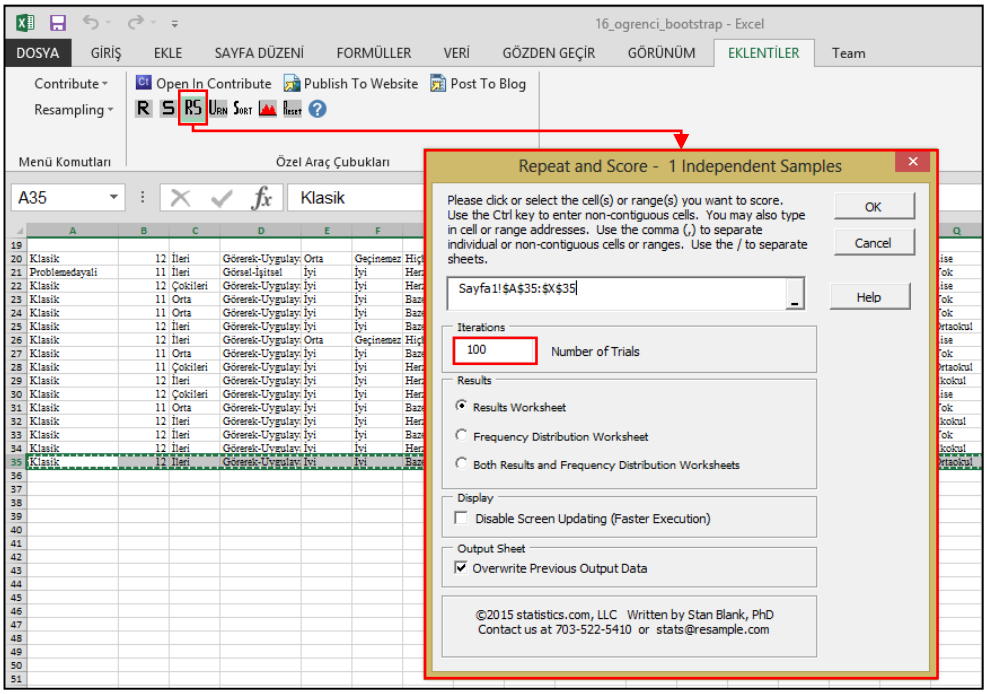
The Process of Bootstrap Resampling by 100 Using Authentic Observations;

1. Upon selecting the authentic observations in the Excel worksheet, the **Resample with replace** command button in the Plug-ins tab. (1)
2. In the **Matrix Resample** dialog box that is opened, the type of resampling (Resample Rows as Units) is selected (2) and the reference for the first cell in which the created instances will be listed is typed and OK command button is clicked (3).

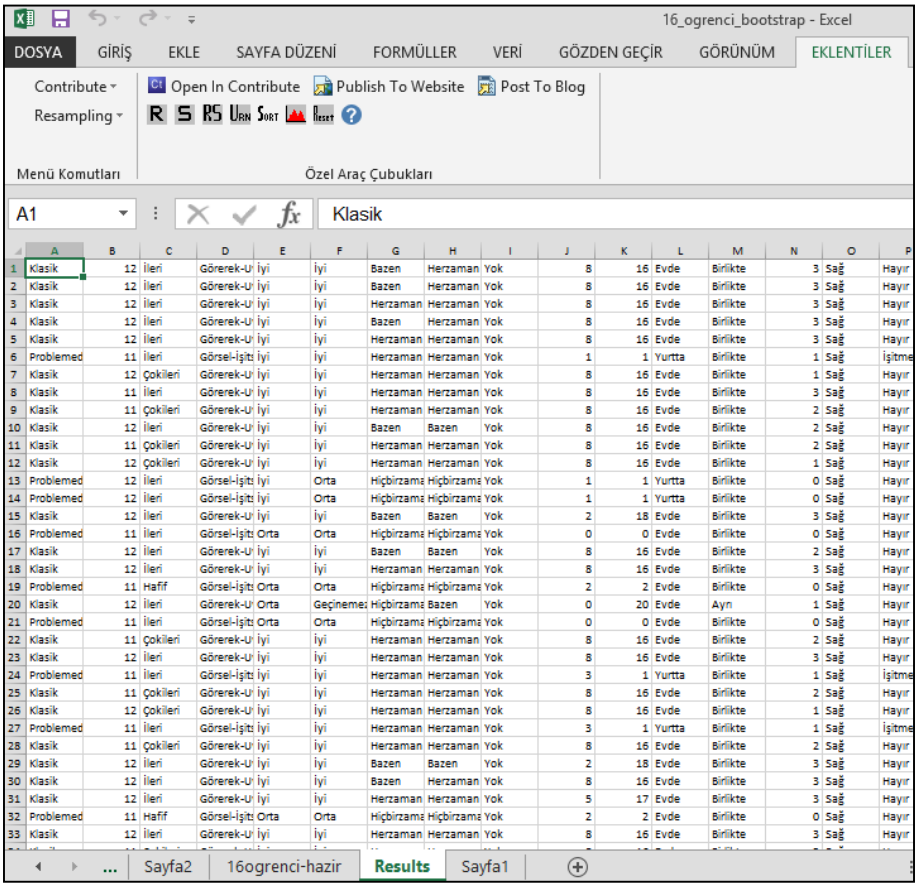


Picture 1. Matrix Resample Dialog Box

3. After the bottom row in the list of authentic observations is completely selected, Repeat and Score command button is clicked. Then the number of instances to be resampled are typed in and OK command button is clicked to create the new instances in a worksheet named ‘result’.



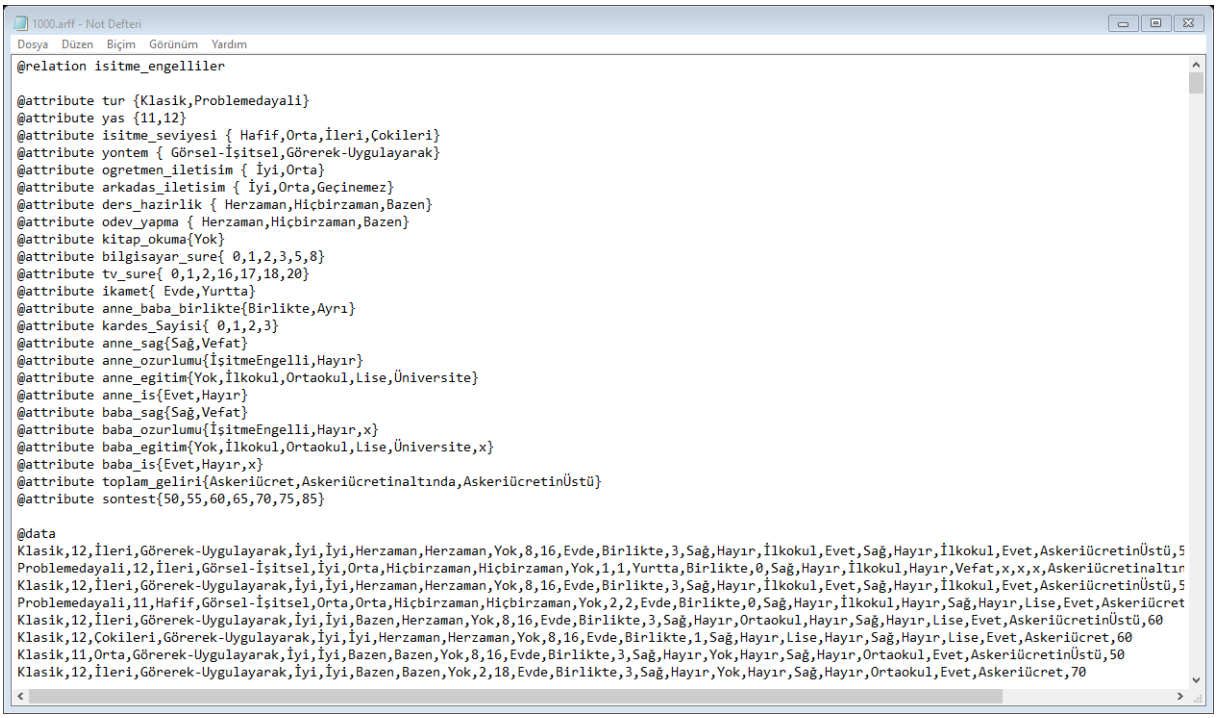
Picture 2. Repeat and Score



Picture 3. Result Worksheet

2.2. Analysis of the Data Set

There are numerours tools used for data mining and one of the most popular tools among them is Weka. By using Weka, it is possible to run various algorithms on the desired data set. The present study also utilized the classification algorithms of Weka 3.7 for the analysis of the data sets of different sizes, created within the context of the study. For this reason, the data sets were converted into .arff format so that weka could read the data. The picture below displays one of these .arff files.


1000.arff - Not Defteri

Dosya Düzen Biçim Görünüm Yardım

```
@relation isitme_engelliler

@attribute tur {Klasik,Problemedayali}
@attribute yas {11,12}
@attribute isitme_seviyesi { Hafif,Orta,İleri,Çokileri}
@attribute yontem { Görsel-İşitsel,Görerek-Uygulayarak}
@attribute ogretmen_iletisim { İyi,Orta}
@attribute arkadas_iletisim { İyi,Orta,Geçinemez}
@attribute ders_hazirlik { Herzaman,Hiçbirzaman,Bazen}
@attribute odev_yapma { Herzaman,Hiçbirzaman,Bazen}
@attribute kitap_okuma{Yok}
@attribute bilgisayar_sure{ 0,1,2,3,5,8}
@attribute tv_sure{ 0,1,2,16,17,18,20}
@attribute ikamet{ Evde,Yurtta}
@attribute anne_baba_birlikte{Birlikte,Ayrı}
@attribute kardes_Sayisi{ 0,1,2,3}
@attribute anne_sag{Sağ,Vefat}
@attribute anne_ozurlumu{İşitmeEngelli,Hayır}
@attribute anne_egitim{Yok,İlkokul,Ortaokul,Lise,Üniversite}
@attribute anne_is{Evet,Hayır}
@attribute baba_sag{Sağ,Vefat}
@attribute baba_ozurlumu{İşitmeEngelli,Hayır,x}
@attribute baba_egitim{Yok,İlkokul,Ortaokul,Lise,Üniversite,x}
@attribute baba_is{Evet,Hayır,x}
@attribute toplam_geliri{AskeriÜcret,AskeriücretinaltındaAskeriücretinÜstü}
@attribute sontest{50,55,60,65,70,75,85}

@data
Klasik,12,İleri,Görerek-Uygulayarak,İyi,İyi,Herzaman,Herzaman,Yok,8,16,Evde,Birlikte,3,Sağ,Hayır,İlkokul,Evet,Sağ,Hayır,İlkokul,Evet,AskeriücretinÜstü,5
Problemedayali,12,İleri,Görsel-İşitsel,İyi,Orta,Hiçbirzaman,Hiçbirzaman,Yok,1,1,Yurtta,Birlikte,0,Sağ,Hayır,İlkokul,Hayır,Vefat,x,x,x,Askeriücretinaltın
Klasik,12,İleri,Görerek-Uygulayarak,İyi,İyi,Herzaman,Herzaman,Yok,8,16,Evde,Birlikte,3,Sağ,Hayır,İlkokul,Evet,Sağ,Hayır,İlkokul,Evet,AskeriücretinÜstü,5
Problemedayali,11,Hafif,Görsel-İşitsel,Orta,Orta,Hiçbirzaman,Hiçbirzaman,Yok,2,2,Evde,Birlikte,0,Sağ,Hayır,İlkokul,Hayır,Sağ,Hayır,Lise,Evet,Askeriücret
Klasik,12,İleri,Görerek-Uygulayarak,İyi,İyi,Bazen,Herzaman,Yok,8,16,Evde,Birlikte,3,Sağ,Hayır,Ortaokul,Hayır,Sağ,Hayır,Lise,Evet,AskeriücretinÜstü,60
Klasik,12,Çokileri,Görerek-Uygulayarak,İyi,İyi,Herzaman,Herzaman,Yok,8,16,Evde,Birlikte,1,Sağ,Hayır,Lise,Hayır,Sağ,Hayır,Lise,Evet,Askeriücret,60
Klasik,11,Orta,Görerek-Uygulayarak,İyi,İyi,Bazen,Bazen,Yok,8,16,Evde,Birlikte,3,Sağ,Hayır,Yok,Hayır,Sağ,Hayır,Ortaokul,Evet,AskeriücretinÜstü,50
Klasik,12,İleri,Görerek-Uygulayarak,İyi,İyi,Bazen,Bazen,Yok,2,18,Evde,Birlikte,3,Sağ,Hayır,Yok,Hayır,Sağ,Hayır,Ortaokul,Evet,Askeriücret,70
```

Picture 4. The .arff Formatted Data Set File with 1000 Instances

Different criteria can be adopted for the measurement of the performance of a classification algorithm. In the present study, the criteria used to measure the performance of the algorithm are prediction accuracy, kappa and f-measure values.

Prediction accuracy is the simplest and most popular method in measuring the model success. Prediction accuracy value uses the observations in the data set as both training and test data, thus giving the rate of prediction accuracy. (Özhan E.2014)

Kappa Statistic (KS) is a criterion used to quantify the level of agreement among the predicted and observed classifications in a data set (Witten et al., 2005; Aydın, F. 2011). KS takes values between -1 and 1. A KS value of -1 indicates a total disagreement or negative relationship and 1 indicates a perfect agreement. (Kılıçaslan,Y. et al., 2009). A KS value of 0.4 and above can be interpreted as an acceptable level of agreement beyond chance. (Landis et al., 1977; Aydın, F. 2011).

F-Measure is expressed as the harmonic mean of precision and recall values. F-measure is a significant criterion especially with the aim of increasing the performance of the classifier in the process of the preparation of the training data. In the performance analysis of the classifiers, the relationship between f-measure and data size was put forward (Kılıçaslan, Y. et al., 2009) and a happy graph was obtained from the learning curve plotted between the f-measure and data size (Russell vd., 2003; Aydın, F.2011). In addition, f-measure was utilized to decide if the classes were sufficient for diagnosis in the application developed in the present study. An acceptable f-measure value is usually a minimum of 0.5, which is also used as a criterion in the present study. (Aydın, F. 2011)

In this study, the classification algorithms of Weka are used in the default values to avoid changes that would affect the model success. Furthermore, Cross-Validation 10 is preferred as the test method. The results of the analyses show that NaïveBayes, LWL and HoeffdingTree algorithms are the ones which produce the most accurate results with the valid data sets. These algorithms are briefly explained below.

Naïve Bayes Algorithm: Naïve Bayes Algorithm is based on the Bayes' probability theorem and works according to the probability of accepting the hypotheses based on particular set of data. Taking into account the available data, the most probable hypothesis is selected. Every attribute is equally important and independent of one another (Kaya, Y., Pehlivan, H.,2014)

Naive Bayes classifiers can be separated into two according to the computation method of term probabilities for a known class. The first method, polynomial method, takes the number of term occurences into account, however, the multivariate method only checks if the terms occur or not. (Gürcan,F., 2009; Kaya,Y., Pehlivan,H.,2014)

HoeffdingTree: Hoeffding Tree algorithm is a decision tree algorithm which runs efficiently on large data sets through reading every instance once at most and processing it in an appropriate time interval. Moreover, Hoeffding Tree algorithm eliminates the storage problem observed in traditional decision tree algorithms such as ID3, C4.5 and SLIQ, and allows for the creation of complex decision trees with an acceptable computational cost. The algorithm utilizes the statistical value called 'Hoeffding Bound' in the decision of how to split each one of the decision tree nodes. One of the important features of the Hoeffding Tree algorithm is that the decision tree created by the algorithm is almost identical to the classifiers which use all the instances to test each node. (Domingos, P.,Hulten, G.,2000; Onan,A.,2014)

LWL: According to Frank et al. (2003), LWL (Locally Weighted Learning) is similar to other lazy learning methods. However, it acts differently while classifying a new instance. LWL algorithm forms a new Naive Bayes model using the data weight of the learning instances. On experimentally tested data sets, LWL outperforms both standard Naive Bayes and similar methods. (Gebeyehu,S., Rao,V.,S., 2014)

3. FINDINGS

Niave Bayes, HoeffdingTree, LWL algorithms were run for data sets of varying sizes and the results are tabulated below.

Sample Size	Correctly Classified Instances %	Kappa	F-measure
16	0	-0,26	0,00
100	81	0,75	0,79
300	87	0,84	0,86
500	95	0,94	0,95
700	90	0,88	0,90
900	93	0,92	0,93
1000	94	0,92	0,94

Table 1. NaïveBayes Results

As seen in Table1, the most accurate result obtained from the data sets of different sizes, on which the NaïveBayes algorithm was run,is acquired from the data set of 500 instances. The results obtained from this data set is as follows: Correct Classification Rate 95%, Kappa value 0,94, F-measure value 0,95.

Sample Size	Correctly Classified Instances %	Kappa	F-measure
16	63	-0,19	0,04
100	81	0,75	0,79
300	87	0,84	0,86
500	95	0,94	0,95
700	90	0,88	0,90
900	93	0,92	0,93
1000	94	0,92	0,94

Table 2. HoeffdingTree Results

Table 2 demonstrates that the most accurate results among the data sets was obtained in the data set with 500 instances. In this data set, the Correct Classification Ratio is 95%, Kappa value is 0,94 and F-measure value is 0,95.

Sample Size	Correctly Classified Instances %	Kappa	F-measure
16	0	-0,25	0,00
100	92	0,89	0,90
300	90	0,88	0,90
500	100	1,00	1,00
700	92	0,90	0,92
900	98	0,98	0,98
1000	99	0,99	0,99

Table 3. LWL Results

Table 3 illustrates that the most accurate results are obtained in the data set of 500 instances among the data sets of varying sizes on which the LWL algorithm was run. In this data set, the rate of Correct Classification is 100%, Kappa value is 1 and F-measure value is 1.

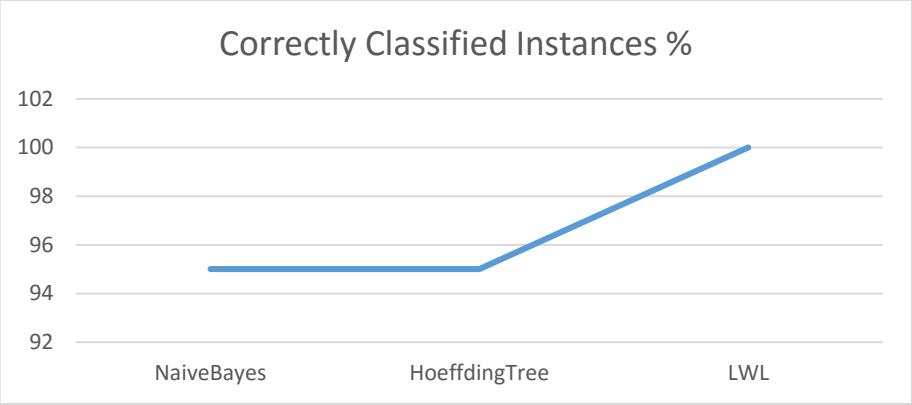
4. CONCLUSION

16 observations acquired within the context of the study entitled An Achievement Analysis of Problem-Based Learning Method in the Education of the Hearing-Impaired with Machine Learning Algorithms were converted into a data set of 1000 instances using the Bootstrap Method, which is one of the resampling methods.

The results of the analysis performed on data sets of 16, 100, 300, 500, 700, 900 and 1000 instances with Weka classification algorithms, it was seen that the LWL algorithm produced a prediction accuracy of 100%, F-measure value of 1 and Kappa Statistic value of 1 for the data set with 500 instances, which is more successful than the other algorithms. In the ranking, NaiveBayes and Hoeffding Tree algorithms shared the second position with the values they produced.

Algorithms	**Correctly Classified Instances %**	**Kappa**	**F-measure**
NaiveBayes	95	0,94	0,95
HoeffdingTree	95	0,94	0,95
LWL	100	1,00	1,00

Table 4. Most Accurate Algorithms



Graphic 1. Correct Classification Ratio

REFERENCES

Aydın F., Kalp Ritim Bozukluğu Olan Hastaların Tedavi Süreçlerini Desteklemek Amaçlı Makine Öğrenmesine Dayalı Bir Sistemin Geliştirilmesi, Yüksek Lisans Tezi,2011;39-41

Bekiroğlu,N., Konyalioğlu, R., Karahan, D., Çoklu Doğrusal Regresyon Sonuçlarının Jackknife Tekniği İle Tekrarlana bilirliğinin Değerlendirilmesi- Marmara Medical Jourcal 2013;26:63-7 Doi:10.5472/Mmj.2013.02499.1.

Fu, W.J., Carrol, R.J. And Wang, S. 2005. "Estimating Missclasification Error With Small Samples Via Bootstrap Cross-Validation", Bioinformatics, 21(9); 1979-1986

Gebeyehu,S., Rao,V.,S., A Two Step Data Mining Approach for Amharic Text Classification ,American Journal of Engineering Research (AJER) e-ISSN : 2320-0847 p-ISSN : 2320-0936 Volume-03, Issue-04, 2014; 251-259

Heller, G., Venkatraman, E.S., 1996, "Resampling Procedures To Compare Two Survival Distributions İn The Presence Of Rigt-Consered Data", Biometrics, 52,(4); 1204-1213

Kaya Y., Pehlivan H., Knn, Nn, Bayes, Dt Ve Svm Kullanılarak Ekg Vurularının Sınıflandırılması, Isıtes2014 Karabük - Türkiye,2014; 1418

Kılıçaslan Y., Güner E. S., Yıldırım S., Learning-based pronoun resolution for Turkish with a comparative evaluation; Computer Speech & Language Volume 23, Issue 3, July 2009;311-331.

Koşkan,Ö., Grup Karşılaştırmalarında Yeniden Örnekleme(Resampling) Yaklaşımı, Ankara Üniversitesi Fen Bilimleri Enstitüsü, Doktora Tezi, 2007.

Nian-Sheng Tang , Shi-Fang, Man-Lai Tang ,Yan-Bo Pei Qiu, 2011, "Asymptotic Confidence Interval Construction For Proportion Difference in Medical Studies With Bilateral Data",Statistical Methods İn Medical Research,; 20: 233–259

Orekkici Temel G., Erdoğan, S., Tanı Koyma Amaçlı Yapılan Tıbbi Çalışmalarda Jacknife, Bootstrap Ve Çapraz Geçerlilik Yöntemlerinin Kullanımı , Düzce Üniversitesi Sağlık Bilimleri Enstitüsü Dergisi 2011;1(3): 45-49 Issn: 2146-443x.

Orekkici Temel G., Erdoğan, S., Sınıflama Modelinin Performansını Değerlendirmede Yeniden Örnekleme Yöntemlerinin Kullanımı – Gülhan Orekkici Temel, Semra Erdoğan, Handan Ankaralı – Bilişim Teknolojileri Dergisi, Cilt:5, Sayı:3, Eylül 2012.

Takma,Ç., Atıl,H., 2006, "Bootstrap Metodu Ve Uygulanışı Üzerine Bir Çalışma 2. Güven Aralıkları, Hipotez Testi Ve Regresyon Analizinde Bootstrap Metodu", Ege Üniv. Ziraat Fak. Derg., , 43(2):63-72, Issn 1018-8851

Wehrens R, Putter H, Buydens Lmc. The Boostrap: A Tutorial.Chemometr Intell Lab 2000; 54: 35-52.

Onan, A., Şirket İflaslarının Tahmin Edilmesinde Karar Ağacı Algoritmalarının Karşılaştırmalı Başarım Analizi, Bilişim Teknolojileri Dergisi, Cilt: 8, Sayı: 1,2015;12

Özhan E., Yapay Zeka Yöntemleri İle Dünya Depremlerinin Modellene Bilirliği, Isıtes2014 Karabük – Turkey,2014; 1188

Zaman,T., Alakus, K., Bulut,H., American Journal Of Theoretical And Applied Statistics 2015; 58-63