Veri Madenciliğine Giriş Dersi

Hazırlayanlar

Nisanur BULUT 152120121018

Metin VATANSEVER 152120131065

İçindekiler

Parkinson Hastalığı Hakkında Bilgi

- Parkinson hastalığı "motor sistem hastalıklar" adı verilen gruba ait, dopamin üreten beyin hücrelerinin kaybıyla ortaya çıkan bir hastalıktır.
- Kademeli olarak ilerler.
- Tedavisi yoktur ancak ilaçlı tedavi ile belirtileri azaltılabilir.
- Parkinson hastalığının konuşma ve ses üzerinde derin etkileri vardır.

Projenin Amacı

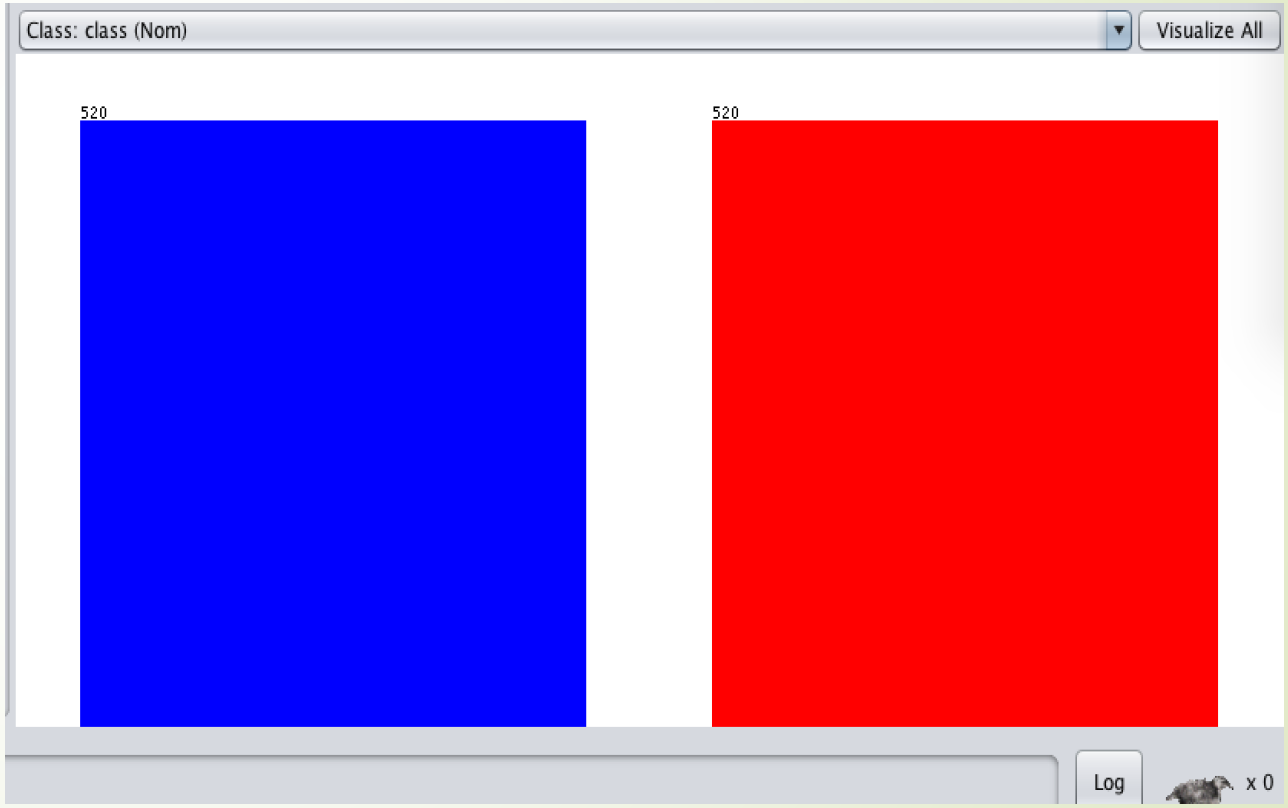
- Outlier'ların belirlenmesi
- Veri seti üzerinde uygulanabilecek farklı sınıflandırıcıların doğruluklarını belirlemek ve bunlar içerisinde en iyi olanı seçmek
- Özniteliklerin sınıflandırıcılara olan katkısının araştırılması
- Seçilen sınıflandırıcıya farklı parametreler vererek en yüksek doğruluk değerini belirlemek
- Veri seti üzerinde yapılan çalışmaları kıyaslama tablolarında göstermek

Veri Seti Bilgisi

- Kullanılan veriler İstanbul Üniversitesi Cerrahpaşa Tıp Fakültesi Nöroloji Anabilim Dalı'nda Uzm. Dr. Şakir Delil gözetiminde kliniğe gelen Parkinson hastaları ve sağlıklı bireylerden aynı ortamda toplanmıştır.
- Hasta örnekleri kapsamında alınan verilerden biri olan hastalığın teşhisinden itibaren geçen süre 0 ile 6 yıl arasında değişmektedir.
- Her hastadan sürekli sesler, sayılar, cümlecikler ve kelimelerden oluşan 26 ses örneği alınmıştır.

Veri Seti Üzerinde Yapılan İncelemeler

- Veri setimizde 29 Attribute 1040 instance bulunmaktadır.
- Class attributemizde 520 instance 520 instance 1 olarak görünmektedir.
- ID attribute alanı outliers kabul edilmiştir.
- Geriye 28 attribute kalmıştır.



Veri Seti Üzerinde Yapılan İncelemeler

- Veri seti incelemesi Naive Bayes algoritmasının kullanımıyla başlanmıştır. Bunun Sebebi:
- Decision Tree, Random Forest,Multiplayer Perceptron gibi clasffier'lar %100 accuracy yakalar.
- Bu nedenle attribute selection işleminin etkisi bu classfier'larda gözlenemez.

Veri Seti Üzerinde Yapılan İncelemeler

	TN	FP	FN	TP	Accuracy	Test Option
Naive Bayes	500	20	0	520	%98,07	Cross Validation (10Folds)
SVM	520	0	0	520	%100	Cross Validation (10Folds)
MLP	520	0	0	520	%100	Cross Validation (10Folds)
j48(Decision Tree)	520	0	0	520	%100	Cross Validation (10Folds)
Random Fores	520	0	0	520	%100	Cross Validation (10Folds)

Tablo 2 de de görüldüğü üzere veri seti üzerindeki değişimleri inclemek için Naive Bayes algoritması en idealidir.

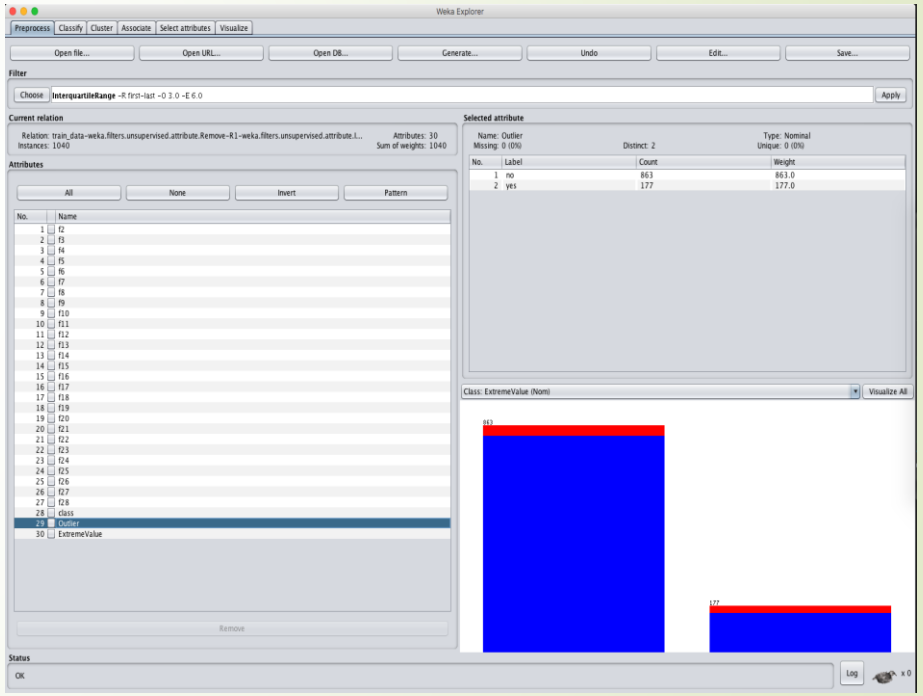
Veri Seti Üzerinde Yapılan İncelemeler

Test Option	Accuracy	Precision	TN	FP	FN	TP
Cros Validation 5 Folds	%97,88	%98	498	22	0	520
Cross Validation 10 Folds	%98,07	%98,1	500	20	0	520
Cros Validation 13 Folds	%98,26	%98,3	502	18	0	520
Cros Validation 15 Folds	%98,17	%98,2	502	18	1	520
Cros Validation 20 Folds	%98,17	%98,2	501	19	0	520
Percentage Split(%30) 312 İnstance	%97,4	%97,6	148	8	0	156
Percentage Split(%40) 416 İnstance	%97,83	%97,9	191	9	0	216
Percentage Split(%45) 458 İnstance	%98,03	%98,1	214	9	0	235
Percentage Split(%50) 520 İnstance	%97,69	%97,8	238	12	0	270

Tablo 3'de görüldüğü üzere Cross Validation 13 Folds işleminde accuracy %98,26 Percentage Split'te %45'ini alınca %98,03'lük bir accuracy yakalamaktadır.

Filtreleme ve Outlier Belirleme

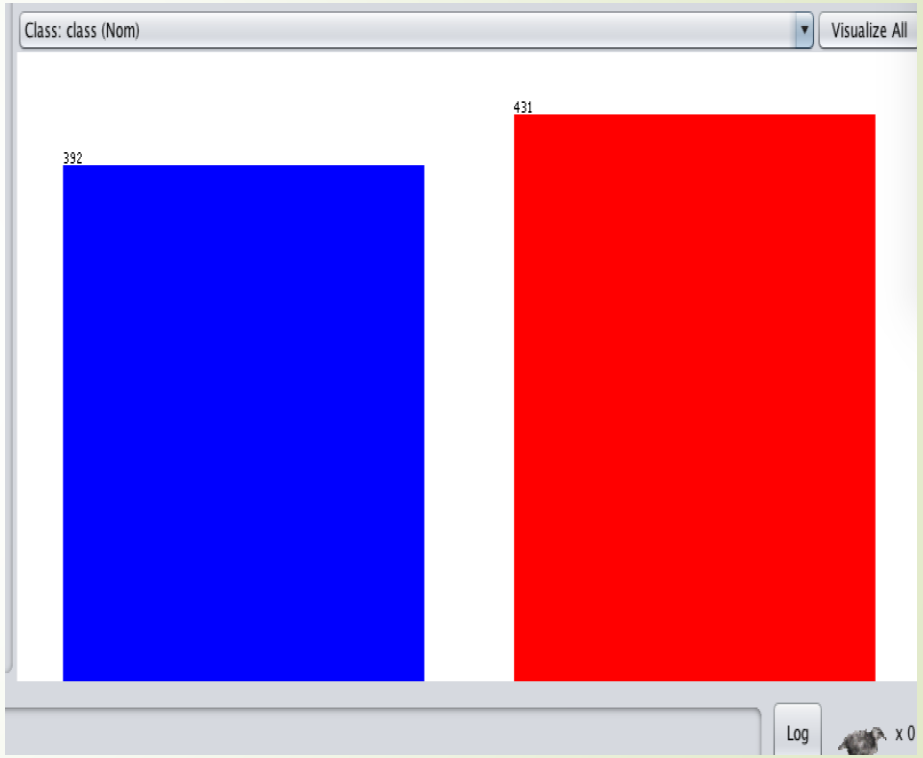
- Outlier olarak belirlenen ilk attribute değeri ID alanıdır.
- InterQuartile Range metodu kullanarak çıkarılan instance'ler Naive Bayes algoritmasınının ürettiği sonuçları arttırıp arttırmacağına deneme işlemi yapılmıştır.
- Interquartile Range sonucu Outlier ve Extreme Value adlı iki attribute oluşmuştur.
- Extreme Value ve Outlier içinde yes ve no'dan oluşan durumlar bulunur.Yes outlier olduğunu no outlier olmadığını belirtir.



Şekil 2

Filtreleme ve Outlier Belirleme

- Outlier ve Extreme Value'leri çıkarma işlemi yaptıktan sonra yandaki resimde gösterilen Class oluşmuştur.
- Şekil 3'de görüldüğü üzere veriler belirli sayı grupları aralarına dağıtılıp gruplandırılmıştır.
- Şekil 1 ve Şekil 3 arasındaki farklara göre class yapısında 0'lar 520 iken 392'ye düştü.1'ler 520 iken 431'e düşmüştür.1040 olan instance sayımız 823'e düşmüştür.



Şekil 3

Filtreleme ve Outlier Belirleme

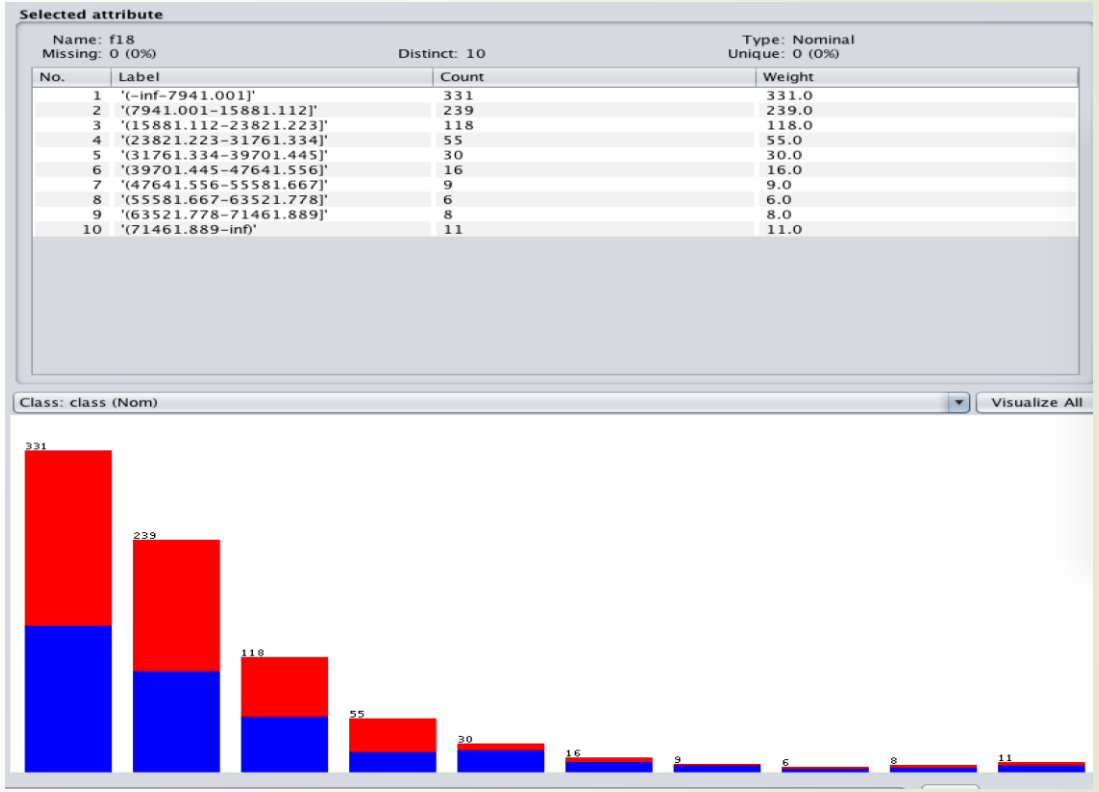
- Tablo 4'e göre İnterQuartile Range ile çıkarılan Extreme Value ve Outlier attributeleri accuracy'yi azaltmıştır.İnterquartileRange Filtrelemesi accuracy değeri azaltılmıştır.
- Dicretize Metodu attribute içindeki instanceleri belirli bir sayı grubuna oturtma işlemi yaparak verinin daha düzgün sonuçlanmasını sağlayabilir.
- Data setinde Discritize metodu kullanarak verinin daha da güçlendirip güçlendirilemeyeceği incelenmiştir.

	Classifier	Test Option	Accuracy
İnterQuartile Range Öncesi	Naive Bayes	10 Fold Cross Validation	%98,07
İnterQuartile Range Sonrası	Naive Bayes	10 Fold Cross Validation	%96,11

Tablo 4: İnterQuartile Range Öncesi ve Sonrası Accuracy Karşılaştırımı

Filtreleme ve Outlier Belirleme

- Discretize sonrası Accuracy %82'yi bulduğundan dolayı discretize işlemi verimizi daha az düşük bir orana sürüklemiştir.



Şekil 4: Discretize Metodu F18 Attribute Verileri

Attribute Selection Method ve İncelemeleri

- CFSSubsetEval Metodu ile Forward Backward yöntemleri ile en çok accuracy yakalayan attributeleri belirleme işlemi yapılmıştır.
- Forward yöntemi ilk başta tek feature ile başlayıp en çok oran yakalayan Feature grubunu bulma işlemi yapacaktır.
- Backward yöntemi ise bunun tam tersidir.
- Çoktan aza doğru ilerleyip en çok accuracy yakalayan feauture'ları grubunu yakalamaya çalışır.
- Forward ve Backward yöntemi ile seçilen Feature'lar F3,F18,F21,F28'dir.

	Option	TN	FP	FN	TP	Accuracy
Öncesi	Cross Validation 10 Folds	500	20	0	520	%98,07
F3,F18,F21,F28 Naive Bayes	Cross Validation 10 Folds	505	15	0	520	%98,55

Tablo 4: CFSSubset Eval ile Seçilen Feautre İncelemesi

WrapperSubsetEvaluator

- Tablo 5'e göre F28 attribute Naive Bayes'e büyük derecede accuracy kazandırmaktadır.
- Sadece F28 ile yapılan aramalar sürekli olarak yüzde 100 accuracy verdiğinden dolayı F28 ile Random Tree, Random Forest, MLP(Multiplayer Perceptron) Accuracy yükseltme işlemi yapılmasına gerek kalmamasıdır.

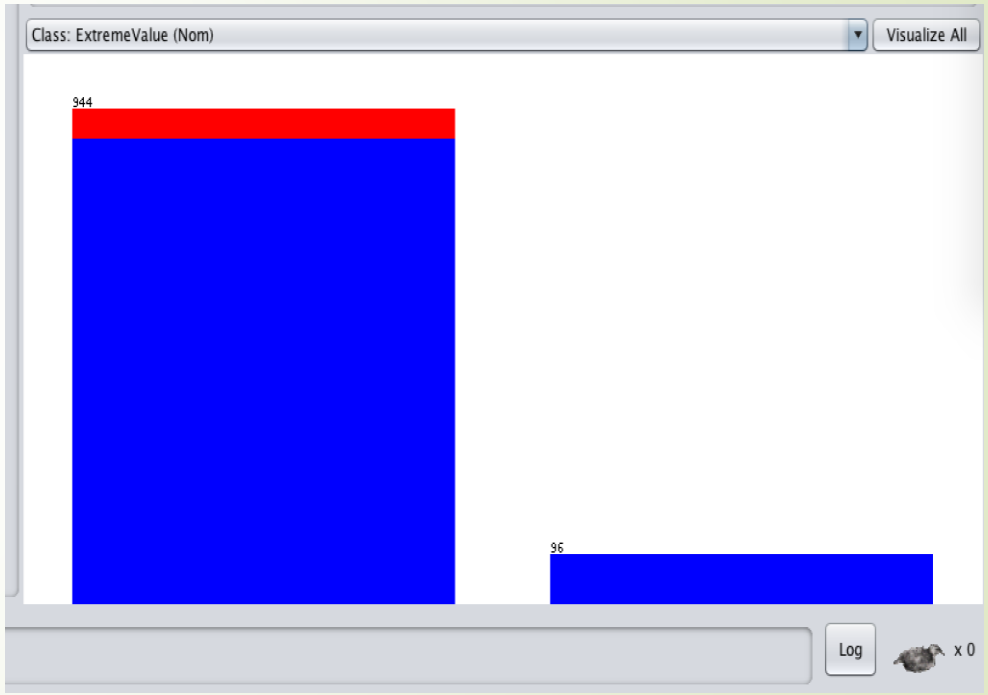
	Option	TN	FP	FN	TP	Accuracy
Öncesi	Cross Validation 10 Folds	500	20	0	520	%98,07
F28 Naive Bayes	Cross Validation 10 Folds	520	0	0	520	%98,55
F28 Silinmiş Naive Bayes	Cross Validation 10 Folds	246	224	171	349	%57,21

Tablo5:WrapperSubsetEvaluator ile Seçilen Feautre İncelemesi

CFSSubset ile Yapılan İlerlemeler

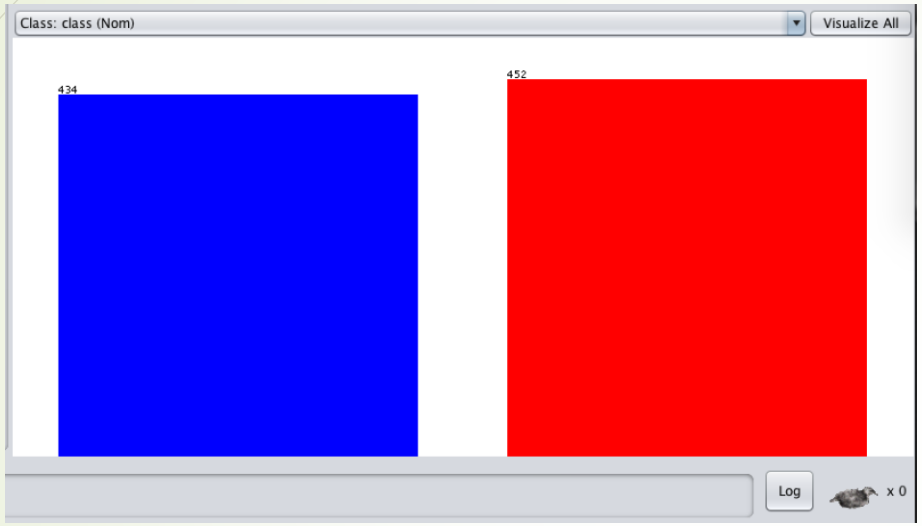
- CFSSubsetEvaluator sonucunda çıkan F13,F18,F21,F28 attributeleri üzerinde hangi işlemlerle accuracy değerini arttıracağımızı inceleme işlemi yapılmıştır.
- Şekil 5 ve Şekil 6'da görüldüğü üzere İnterquartile Range işlemi Feature Selection sonrası azalmıştır.Outlier 96(bknz. Şekil 5) Extreme Value 86(bknz. Şekil 6) olmuştur.

İnterQuartile Range ve Discretize

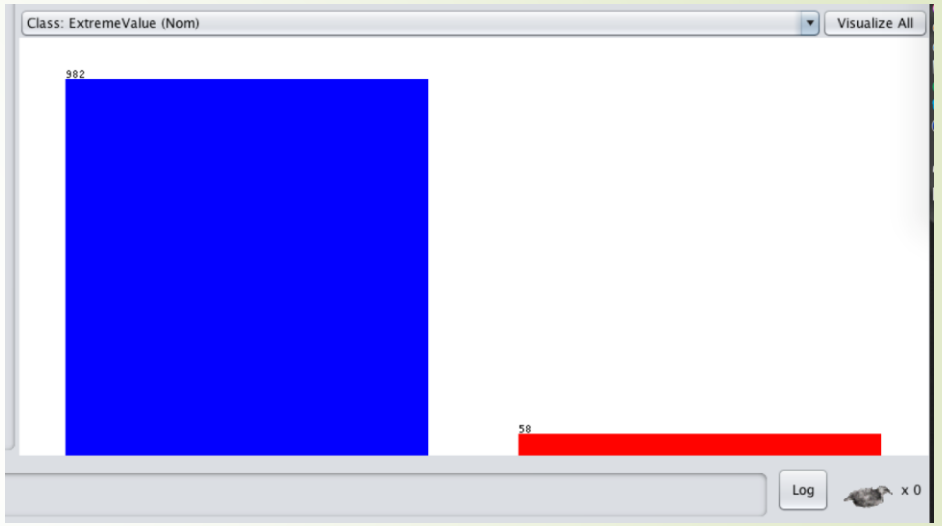


Şekil 5:Feature Selection Sonrası İnterQuartile Range Outliers

CFSSubset ile Yapılan İlerlemeler



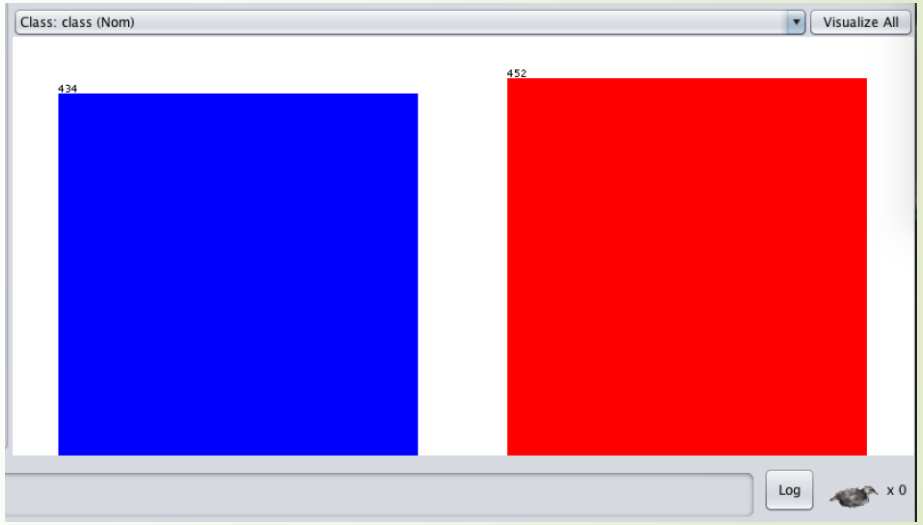
Şekil 7: İnterquartile Range Sonrası Class attribute



Şekil 6: InterQuartile Range Sonrası Extreme Values

CFSSubset ile Yapılan İlerlemeler

- Şekil 7'de görüldüğü üzere Feature Selection sonrası classlar daha düzgün duruma gelmiş bulunmaktadır.
- No 434,Yes 452'ye düşerken sınıflar arasındaki instance farkı 8 e düşmüştür.İnstance sayısı 1040'tan 886'ya düşmüştür.
- Sınıf arasındaki fark çok olmaması classify ederken daha verimli sayılara ulaşmasını sağlamasıdır.



Şekil 7: İnterquartile Range Sonrası Class attribute

CFSSubset ile Yapılan İlerlemeler

- Tablo 6'yı incelendiğinde İnterQuartile Range'in etkisi Accuracy'i %0,99 artırmaktadır.
- Tablo 7'de görüldüğü üzere Discretize filtremelesi uygulandığında accuracy büyük oranda düşmektedir.

	TN	FP	FN	TP	Accuracy
İnterQuartile Range Cross Validation (10 Folds)	432	2	2	450	%99,54
Öncesi Cross Validation (10 Folds)	505	15	0	520	%98,55

Tablo 6: Feature Selection ve Feture Selection Sonrası İnterQuartile Range İncelemesi

	TN	FP	FN	TP	Accuracy
Discritize Cross Validation (10 Folds)	518	2	78	442	%92,30
Öncesi Cross Validation (10 Folds)	505	15	0	520	%98,55

Tablo 7: Feauter Selection ve Feauture Selection Sonrası Discritize İncelemesi

Feature Selection ve İnterQuartile Range Sonrası Classify Optimizasyonu

- İnterquartile Range Filtreleme ile %99,54 ü bulunmuştur.
- Bununla birlikte Classify Options ile Cross Validation Sonucu Accuracy sonucu hiçbir şekilde değişmez.
- Percentage Split ile %30 işlemi ile %100 accuracy sonucuna ulaşılmıştır.

	TN	FP	FN	TP	Accuracy
Cross Validation (5 Folds)	432	2	2	450	%99,54
Cross Validation (10 Folds)	432	2	2	450	%99,54
Cross Validation (15 Folds)	432	2	2	450	%99,54
Cross Validation (20 Folds)	432	2	2	450	%99,54
Percentage Split (%30) 266 Instances	120	0	0	146	%100
Percentage Split (%40) 354 Instances	169	0	1	184	%99,71

Tablo 8: Naive Bayes Last Test Options