

Weka ile Veri Madenciliđi

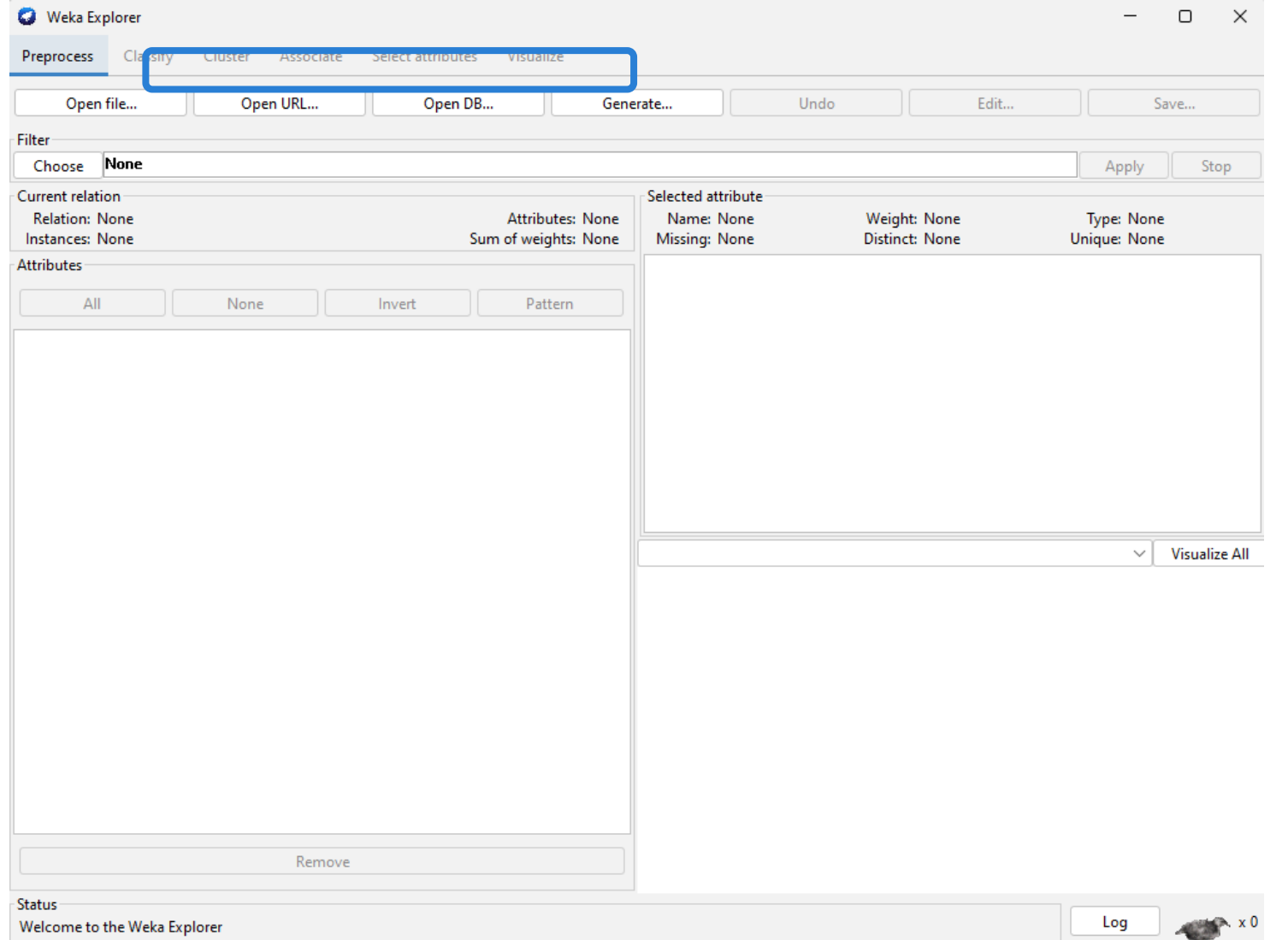
VERİ ÖNİŐLEME

Veri Ön İşleme

Veri madenciliği (Data Mining) uygulamasının ilk aşaması veri ön işleme aşamasıdır. Bu aşama; işlenecek veriyi ön eleme işlemlerinden geçirerek analiz için en uygun veriyi elde etme aşamasıdır. Weka'da, veri madenciliği yöntemleri olan sınıflandırma, kümeleme ve birliktelik analizi işlemlerinin sonucunda başarılı geri dönüşler alabilmek için eldeki veri ön işlemeden geçirilmelidir.

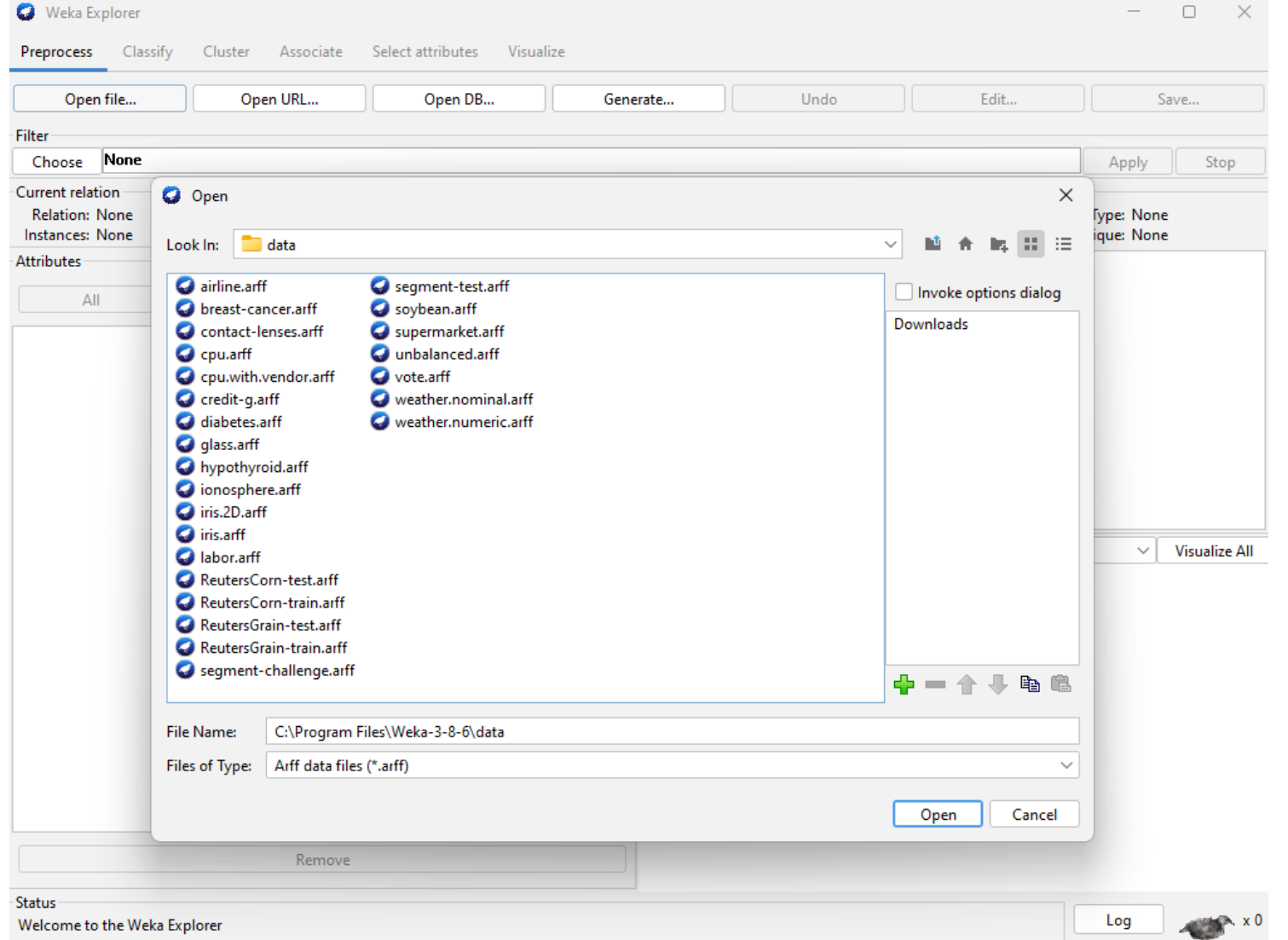
Veri Yükleme

Weka programı açıldığında öncelikle Preprocess haricindeki sekmeler tıklanamaz durumdadır yani ilk olarak veri analizinde kullanılacak veri kümemizi seçip önışleme işlemini yapmamız gereklidir. Veri kümemizi seçmek için "Open file..." butonuna tıklayıp analiz edeceğiniz veri kümesini seçilir. Buradan Weka'nın desteklediği dosya türleri olan ARFF dosya türündeki veriler sisteme yüklenebilir.



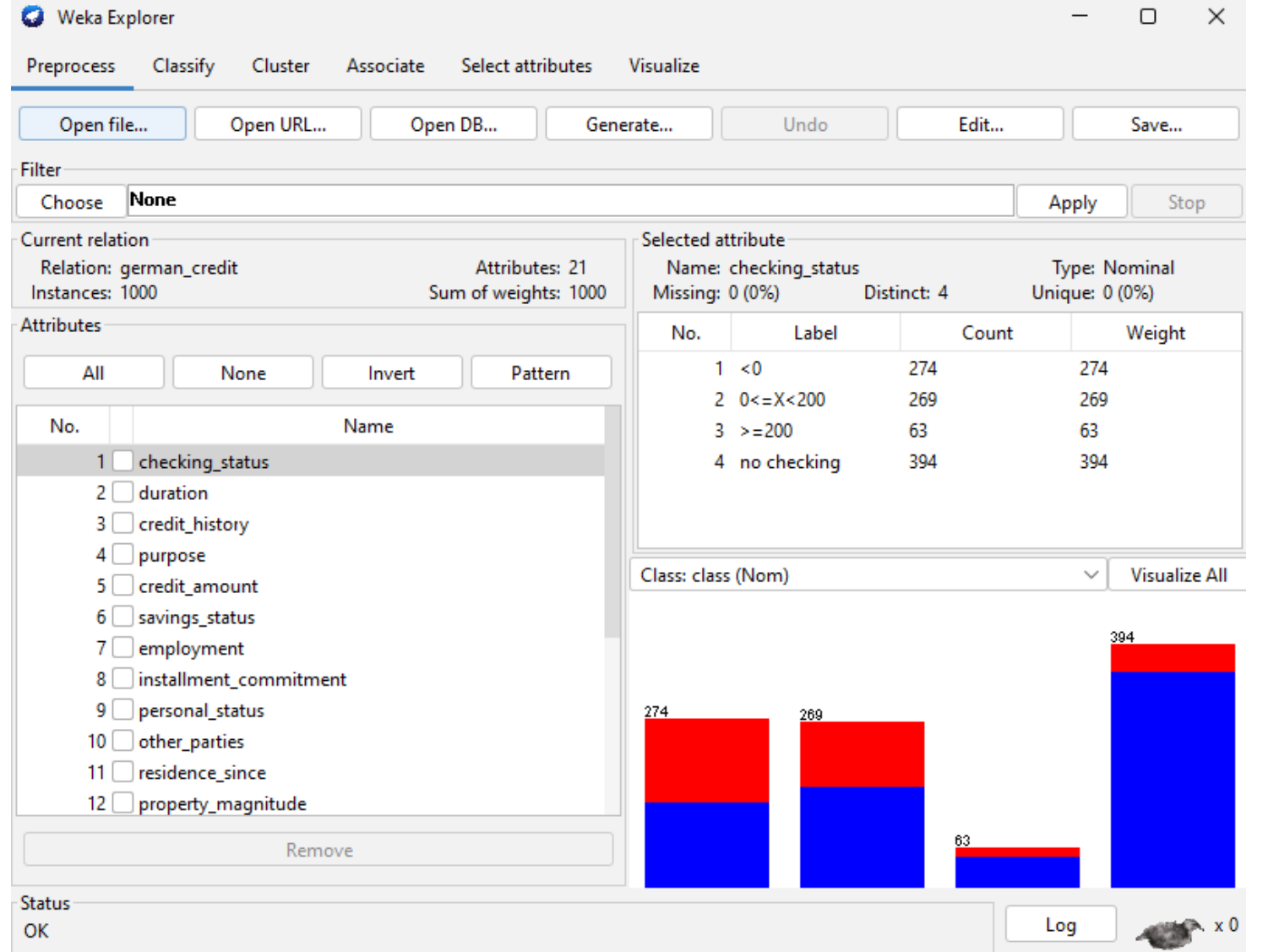
Hazır Veri Setleri

Bu aşamada Hazır veri setleri kullanılmak istenirse C'de Program Files altında Weka klasörünün altındaki data klasöründe (C:\Program Files\Weka-3-8-6\data) yer alan veri setlerinden biri kullanılabilir. Ayrıca "Open URL" butonuna tıklayarak verilecek URL'den veri kümesini ve "Open DB" butonuna tıklayarak veri tabanı bağlantısı kurarak çekilecek veri kümesi Weka çalışma ortamına yüklenebilir.



Hazır Veri Setleri

Görselde de görüldüğü gibi credit-g.arff isimli veri seti seçilerek Weka çalışma ortamına yüklendi. Gelen ekranda veri setine ait görüntüler yer alır. Veri setindeki bilgilerin ne anlama geldiğini bilirse veri madenciliği tekniklerinin uygulanması hızlanır ve analiz sonucunda doğru sonuçlar elde edilir.



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose None Apply Stop

Current relation: Relation: german_credit Attributes: 21 Instances: 1000 Sum of weights: 1000

Attributes: All None Invert Pattern

No.	Name
<input checked="" type="checkbox"/>	checking_status
<input type="checkbox"/>	duration
<input type="checkbox"/>	credit_history
<input type="checkbox"/>	purpose
<input type="checkbox"/>	credit_amount
<input type="checkbox"/>	savings_status
<input type="checkbox"/>	employment
<input type="checkbox"/>	installment_commitment
<input type="checkbox"/>	personal_status
<input type="checkbox"/>	other_parties
<input type="checkbox"/>	residence_since
<input type="checkbox"/>	property_magnitude

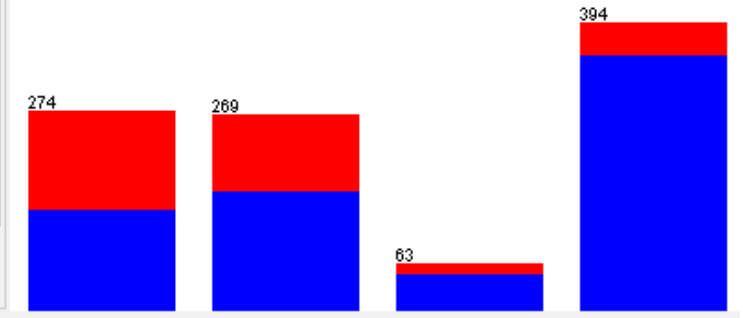
Remove

Status: OK Log x 0

Selected attribute: Name: checking_status Type: Nominal Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

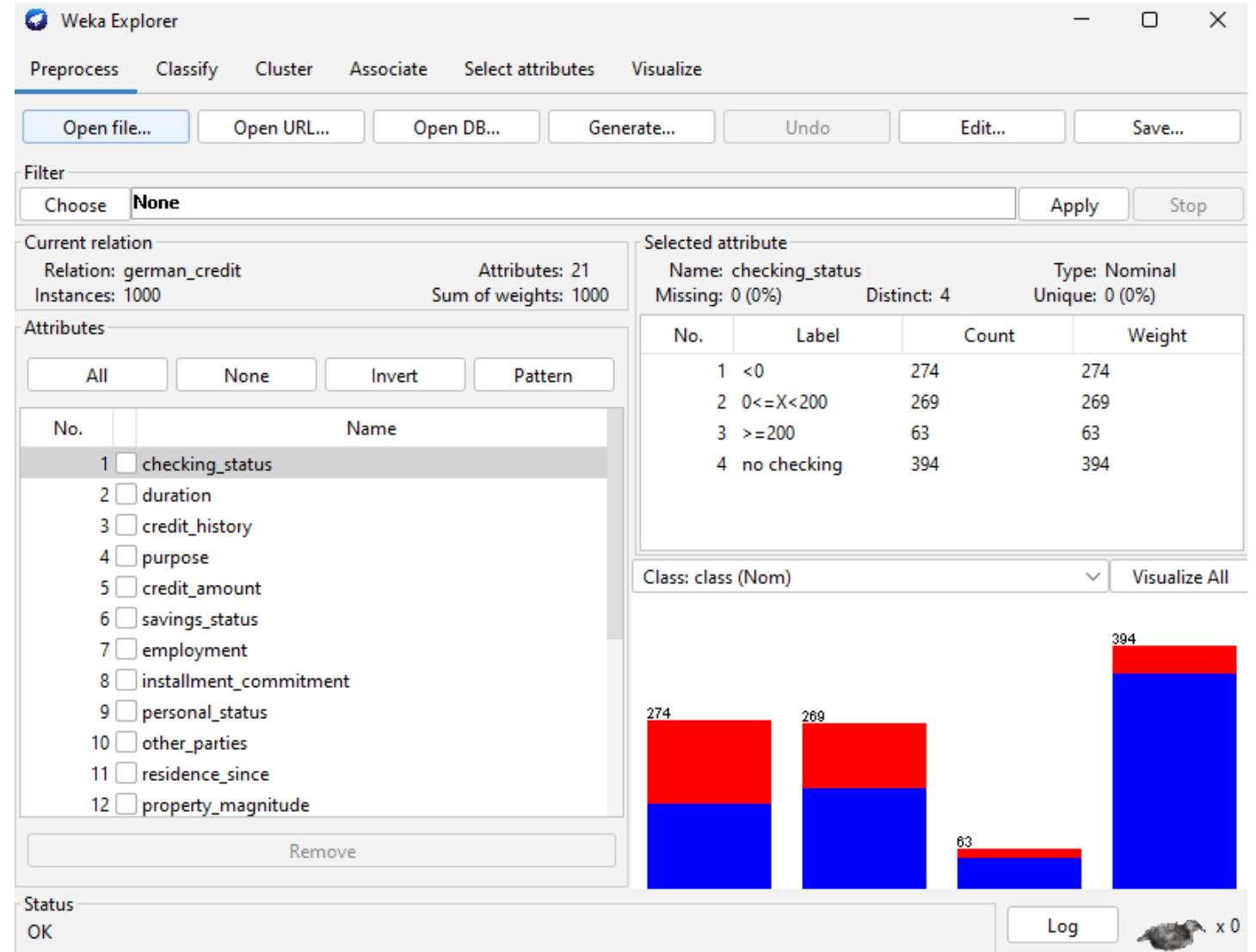
No.	Label	Count	Weight
1	<0	274	274
2	0<=X<200	269	269
3	>=200	63	63
4	no checking	394	394

Class: class (Nom) Visualize All



Hazır Veri Setleri

Üzerinde çalışılacak veri kümesi german-c Alman Kredi Verisi olarak adlandırılan bir veri kümesidir. Bu veri kümesinde 21 tane alan (attribute - nitelik) ve 1000 kayıt (instances) bulunmaktadır. Her bir niteliğin özelliklerini görmek için üzerine tıklayabilirsiniz. Selected attribute kısmından seçilen alanın değerleri kontrol edinilebilir.



The screenshot shows the Weka Explorer interface. The 'Selected attribute' section is active, displaying the following information:

Name: checking_status
Type: Nominal
Missing: 0 (0%)
Distinct: 4
Unique: 0 (0%)

No.	Label	Count	Weight
1	<0	274	274
2	0<=X<200	269	269
3	>=200	63	63
4	no checking	394	394

Below the table, there is a bar chart showing the distribution of the 'checking_status' attribute. The bars are stacked with blue and red colors. The counts for each bar are 274, 269, 63, and 394, corresponding to the labels in the table above.

Class: class (Nom) Visualize All

Nitelik Filtreleme

Veri madenciliğinde veri setinde bulunan bütün niteliklerin kullanılması zorunlu değildir. Hangi konu üzerinde çalışılacaksa ve hangi alanlar yarar sağlayacaksa onlar seçilmelidir. Bu nitelik seçme işlemi Nitelik Filtreleme olarak adlandırılır. Weka'da Filtreleme yapmak için Weka çalışma ortamında yani Weka Explorer 'da Filter panelindeki Choose butonuna tıklanır. "unsupervised" kategorisi altındaki attribute klasörü altındaki "remove" seçilir. Filtre uygulanmadan önce çalışılmayacak nitelikler Attributes panelinden yanındaki kutucuk işaretlenerek seçilir. Nitelikler seçildikten sonra Filter Panelinde yer alan Apply (Uygula) butonuna tıklanarak seçilen alanlar(nitelikler) çalışma ortamından çıkarılır. Veri setinden istenmeyen alanlar çıkarıldıktan sonra veri setinin son halini kayıt etmek için "Save" butonuna tıklanır.

Dr.Günay TEMÜR

The screenshot shows the Weka Explorer interface. The 'Filter' panel is active, and the 'Remove' button is selected. The 'Current relation' is 'german_credit' with 1000 instances and 21 attributes. The 'Attributes' panel shows a list of attributes with checkboxes. The 'own_telephone' attribute is selected. The 'Selected attribute' panel shows the distribution of the 'own_telephone' attribute: 'none' (596 instances) and 'yes' (404 instances). A bar chart below the 'Selected attribute' panel visualizes this distribution, with the 'none' bar being red and the 'yes' bar being blue.

No.	Label	Count	Weight
1	none	596	596
2	yes	404	404

Veri Parçalama

Bir diğer önışleme yöntemi veriyi parçalara ayırma işlemidir. Veri ayırma işlemi işimize yarayacak? Bu özellik veriyi belli bir formata çekme işlemidir. Örneğin veri setimizde yaş alanı 19 dan başlayıp 75'e kadar ulaşmaktadır. Bu alanla çalışırken her yaşı bir değer olarak almak yerine belirli alanı belirli aralıklarla parçalayarak veri üzerinde çalışmayı kolaylaştırabiliriz. Filter panelinde yer alan Discretize filtresi bu işlem için kullanılır.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **Discretize** -B 10 -M -1.0 -R first-last -precision 6 Apply Stop

Current relation: Relation: german_credit Instances: 1000 Attributes: 21 Sum of weights: 1000

Attributes: All None Invert Pattern

No.	Name
1	<input type="checkbox"/> checking_status
2	<input type="checkbox"/> duration
3	<input type="checkbox"/> credit_history
4	<input type="checkbox"/> purpose
5	<input checked="" type="checkbox"/> credit_amount
6	<input type="checkbox"/> savings_status
7	<input type="checkbox"/> employment
8	<input type="checkbox"/> installment_commitment
9	<input type="checkbox"/> personal_status
10	<input type="checkbox"/> other_parties
11	<input type="checkbox"/> residence_since
12	<input type="checkbox"/> property_magnitude
13	<input type="checkbox"/> age
14	<input type="checkbox"/> other_payment_plans
15	<input type="checkbox"/> housing
16	<input type="checkbox"/> existing_credits

Remove

Status: OK

Log

Class: class (Nom) Visualize All

Statistic	Value
Minimum	250
Maximum	18424
Mean	3271.258
StdDev	2822.737

250 198 290 171 115 52 39 40 31 17 11 9 8 6 4 4 0 1 18424

Veri Parçalama

Weka - Filter Paneli - Choose - unsupervised - attribute - Discretize yolunu izleyerek Discretize filtresine ulaşabilirsiniz. Ben örnek olarak age ve credit_amount niteliklerine bu filteri uygularım. Bu konu hakkında bir ipucu vermek gerekirse bu filtreleme için min ve max değer içeren nitelikler seçilmesi uygundur. Bu niteliği kullanmak isterseniz minimum - maximum içermesine dikkat edin.

Dr.Günay TEMÜR

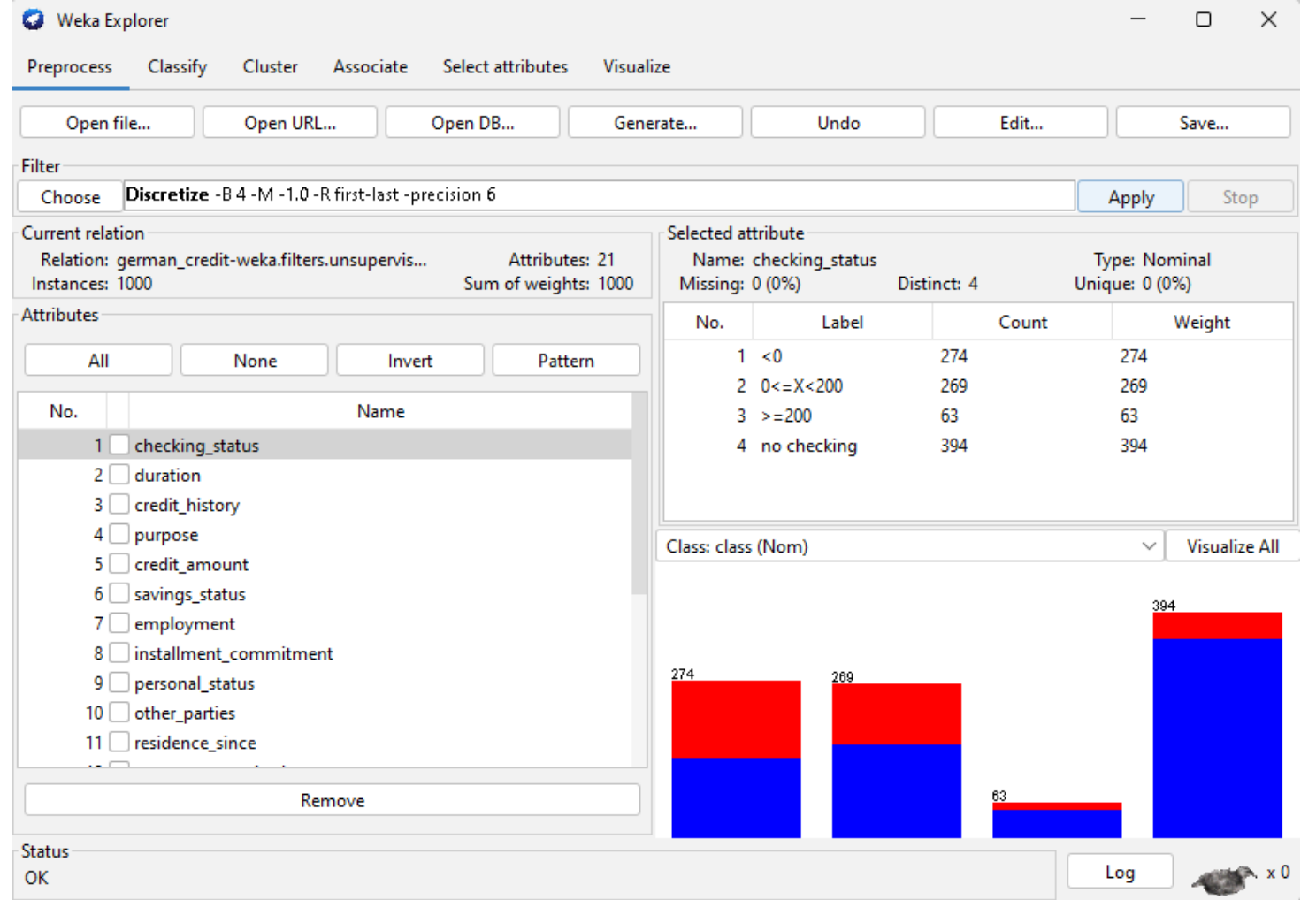
The screenshot shows the Weka Explorer interface with the Discretize filter applied to the 'credit_amount' attribute. The filter configuration is as follows:

- Filter: Discretize -B 4 -M -1.0 -R first-last -precision 6
- Current relation: german_credit (Instances: 1000)
- Selected attribute: credit_amount (Type: Numeric, Unique: 847 (85%))
- Attributes list: credit_amount is selected (checked).
- Filter configuration dialog (weka.gui.GenericObjectEditor):
 - attributeIndices: first-last
 - binRangePrecision: 6
 - bins: 4
 - debug: False
 - desiredWeightOfInstancesPerInterval: -1.0
 - doNotCheckCapabilities: False
 - findNumBins: False

Numbered callouts (1-5) highlight the 'Open' button, the filter name, the 'Apply' button, the 'OK' button in the dialog, and the 'Save' button in the dialog respectively.

Veri Parçalama

Yukarıdaki görseldeki adımlar izlenerek Binning metodu da denilen veri parçalama işlemi kolaylıkla yapılabilir. 3. adımda bins değerine 4 verilmesi o niteliği yandaki gibi 4 parçaya bölünmesi anlamına gelmektedir.



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **Discretize** -B 4 -M -1.0 -R first-last -precision 6 Apply Stop

Current relation: Relation: german_credit-weka.filters.unsupervis... Attributes: 21
Instances: 1000 Sum of weights: 1000

Attributes: All None Invert Pattern

No.	Name
<input checked="" type="checkbox"/>	checking_status
<input type="checkbox"/>	duration
<input type="checkbox"/>	credit_history
<input type="checkbox"/>	purpose
<input type="checkbox"/>	credit_amount
<input type="checkbox"/>	savings_status
<input type="checkbox"/>	employment
<input type="checkbox"/>	installment_commitment
<input type="checkbox"/>	personal_status
<input type="checkbox"/>	other_parties
<input type="checkbox"/>	residence_since

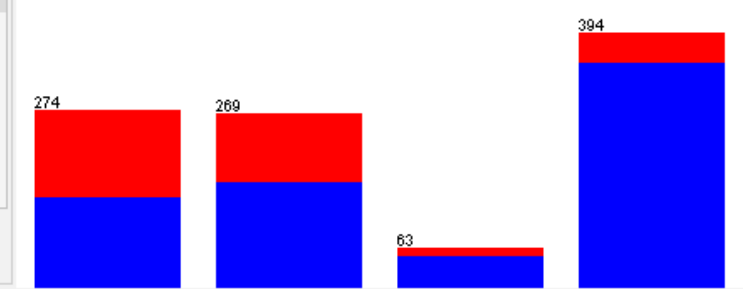
Remove

Status: OK Log x 0

Selected attribute: Name: checking_status Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count	Weight
1	<0	274	274
2	0<=X<200	269	269
3	>=200	63	63
4	no checking	394	394

Class: class (Nom) Visualize All



Veri Önişleme Tamamlandı.

Kaynaklar:

<https://bilgisayarkavramlari.com>

<https://www.veribilimiokulu.com>

<https://zeynepozturkk.wordpress.com>

<https://kubracosar.blogspot.com>

<https://tr.myservername.com>

Dr.Günay TEMÜR