WEKA ile Veri Madenciliği

VERİ HAZIRLAMA

Dr.Günay TEMÜR

Dosya Formatı

WEKA'da işlemler yapabilmemiz için desteklenen dosya formatlarını kullanmamız gerekir.

Genelde internet üzerinden edinilen veriler bir excel tablosunda tutuluyor olabilir. Bu aşamda işimizi kolaylaştırma adına 2 aşamalı çalışma yapmamız gerekecektir.

1. Verileri öncelikle bir .csv formatına dönüştürmek

2. WEKA için .arff formatına dönüştürmek

.csv Dosya Formatı

Excel programınında desteklediği virgülle ayrılmış değerler dosyası, değerleri ayırmak için virgül kullanan sınırlandırılmış bir metin dosyasıdır. Dosyanın her satırı bir veri kaydıdır. Her kayıt, virgülle ayrılmış bir veya daha fazla alandan oluşur. Alan ayırıcı olarak virgül kullanılması, bu dosya biçiminin adının kaynağıdır. Bu format excel veya basit olarak note defteri ile açılabilmektedir. Bu sebple elde edilen bir xls dosyasını öncelikle cvs formatına çevirmek işimizi oldukça kolaylaştıracaktır.

.xls to .csv

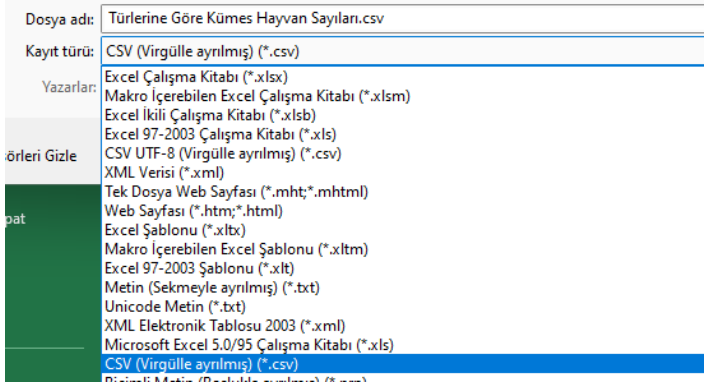
XLS

	A	B	D	F	H	J
1	**Türlerine göre kümes hayvanları sayısı**					
2	Number of poultry animals by types					
3	(**Türkiye** - Turkey)					
4		**Yumurta tavuğu**	**Et tavuğu**	**Hindi**	**Kaz**	**Ördek**
5	**Yıl**	Laying hens	Broilers	Turkeys	Geese	Ducks
6	Year	(**adet** - number)	(**adet** - number)	(**adet** - number)	(**adet** - number)	(**adet** - number)
7	1991	50 826 656	88 379 548	3 132 676	1 599 831	1 112 015
8	1992	52 224 952	100 305 100	3 332 794	1 752 495	1 154 743
9	1993	58 179 047	120 080 935	3 340 241	1 687 596	1 171 961
10	1994	57 842 034	125 842 269	3 441 995	1 719 833	1 186 891
11	1995	57 324 654	71 689 773	3 291 000	1 745 163	1 199 925
12	1996	53 883 070	99 073 900	3 063 540	1 641 915	1 093 860
13	1997	61 401 783	104 870 702	5 327 501	1 794 610	1 828 792
14	1998	69 722 271	167 275 380	3 805 345	1 771 327	1 339 468
15	1999	71 885 207	167 862 730	3 762 516	1 670 916	1 294 824
16	2000	64 709 040	193 459 280	3 681 558	1 496 604	1 104 176
17	2001	55 675 750	161 899 442	3 254 018	1 397 560	913 748
18	2002	57 139 257	188 637 066	3 092 408	1 400 136	832 091
19	2003	60 399 520	217 133 076	3 994 093	1 336 775	810 910
20	2004	58 774 172	238 101 895	3 902 346	1 250 634	770 436
21	2005	60 275 674	257 221 440	3 697 103	1 066 581	656 409
22	2006	58 698 485	286 121 360	3 226 941	830 081	525 250
23	2007	64 286 383	205 082 159	2 675 407	1 022 711	481 829
24	2008	63 364 818	180 915 558	3 230 318	1 062 887	470 158
25	2009	66 500 461	163 468 942	2 755 349	944 731	412 723

CSV

	A	B	C
1	YIL	TÜR	SAYI
2	1991	Yumurta tavuğu	50 826 656
3	1991	Et tavuğu	88 379 548
4	1991	Hindi	3 132 676
5	1991	Kaz	1 599 831
6	1991	Ördek	1 112 015
7	1992	Yumurta tavuğu	52 224 952
8	1992	Et tavuğu	100 305 100
9	1992	Hindi	3 332 794
10	1992	Kaz	1 752 495
11	1992	Ördek	1 154 743
12	1993	Yumurta tavuğu	58 179 047
13	1993	Et tavuğu	120 080 935
14	1993	Hindi	3 340 241
15	1993	Kaz	1 687 596
16	1993	Ördek	1 171 961
17	1994	Yumurta tavuğu	57 842 034
18	1994	Et tavuğu	125 842 269
19	1994	Hindi	3 441 995
20	1994	Kaz	1 719 833

Bu aşama için tablolarda ufak bir düzenleme gerçekleştirip «farklı kaydet» özelliğinden tür olarak .csv özelliği seçilmesi yeterlidir.



.xls to .csv

XLS

	A	B	C	D	E	F	G	H	I	J
1	**Türlerine göre kümes hayvanları sayısı**									
2	Number of poultry animals by types									
3	**(Türkiye** - Turkey)									
4		**Yumurta tavuğu**		**Et tavuğu**		**Hindi**		**Kaz**		**Ördek**
5	**Yıl**	Laying hens		Broilers		Turkeys		Geese		Ducks
6	Year	**(adet** - number)		**(adet** - number)		**(adet** - number)		**(adet** - number)		**(adet** - number)
7	1991	50 826 656		88 379 548		3 132 676		1 599 831		1 112 015
8	1992	52 224 952		100 305 100		3 332 794		1 752 495		1 154 743
9	1993	58 179 047		120 080 935		3 340 241		1 687 596		1 171 961
10	1994	57 842 034		125 842 269		3 441 995		1 719 833		1 186 891
11	1995	57 324 654		71 689 773		3 291 000		1 745 163		1 199 925
12	1996	53 883 070		99 073 900		3 063 540		1 641 915		1 093 860
13	1997	61 401 783		104 870 702		5 327 501		1 794 610		1 828 792
14	1998	69 722 271		167 275 380		3 805 345		1 771 327		1 339 468
15	1999	71 885 207		167 862 730		3 762 516		1 670 916		1 294 824
16	2000	64 709 040		193 459 280		3 681 558		1 496 604		1 104 176
17	2001	55 675 750		161 899 442		3 254 018		1 397 560		913 748
18	2002	57 139 257		188 637 066		3 092 408		1 400 136		832 091
19	2003	60 399 520		217 133 076		3 994 093		1 336 775		810 910
20	2004	58 774 172		238 101 895		3 902 346		1 250 634		770 436
21	2005	60 275 674		257 221 440		3 697 103		1 066 581		656 409
22	2006	58 698 485		286 121 360		3 226 941		830 081		525 250
23	2007	64 286 383		205 082 159		2 675 407		1 022 711		481 829
24	2008	63 364 818		180 915 558		3 230 318		1 062 887		470 158
25	2009	66 500 461		163 468 942		2 755 349		944 731		412 723

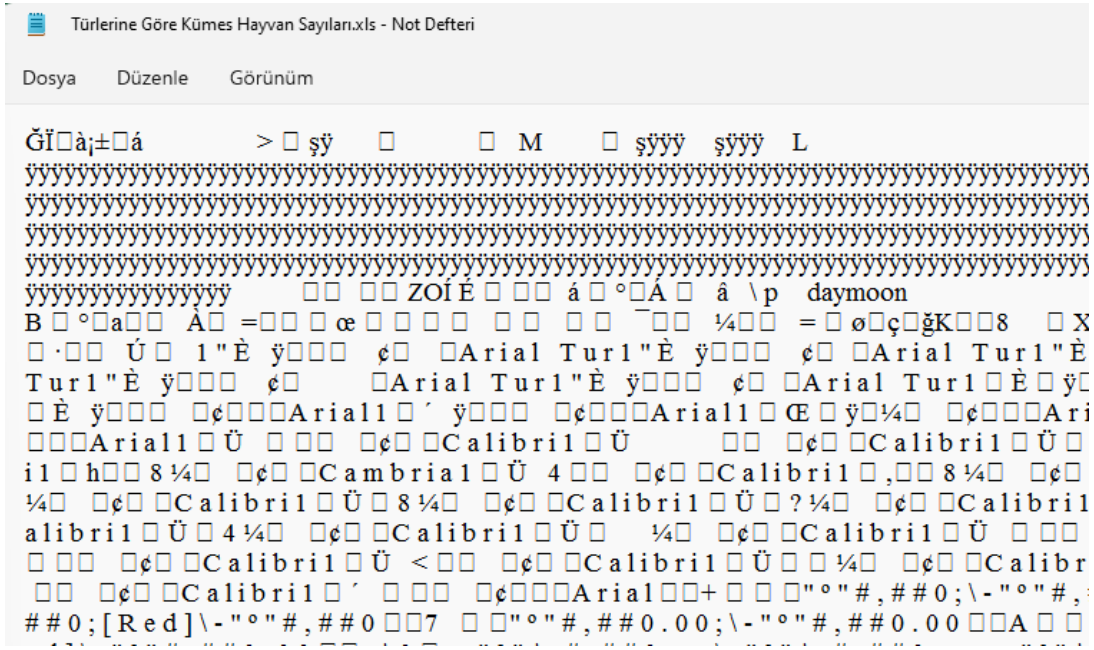
CSV

	A	B	C
1	YIL	TÜR	SAYI
2	1991	Yumurta tavuğu	50 826 656
3	1991	Et tavuğu	88 379 548
4	1991	Hindi	3 132 676
5	1991	Kaz	1 599 831
6	1991	Ördek	1 112 015
7	1992	Yumurta tavuğu	52 224 952
8	1992	Et tavuğu	100 305 100
9	1992	Hindi	3 332 794
10	1992	Kaz	1 752 495
11	1992	Ördek	1 154 743
12	1993	Yumurta tavuğu	58 179 047
13	1993	Et tavuğu	120 080 935
14	1993	Hindi	3 340 241
15	1993	Kaz	1 687 596
16	1993	Ördek	1 171 961
17	1994	Yumurta tavuğu	57 842 034
18	1994	Et tavuğu	125 842 269
19	1994	Hindi	3 441 995
20	1994	Kaz	1 719 833

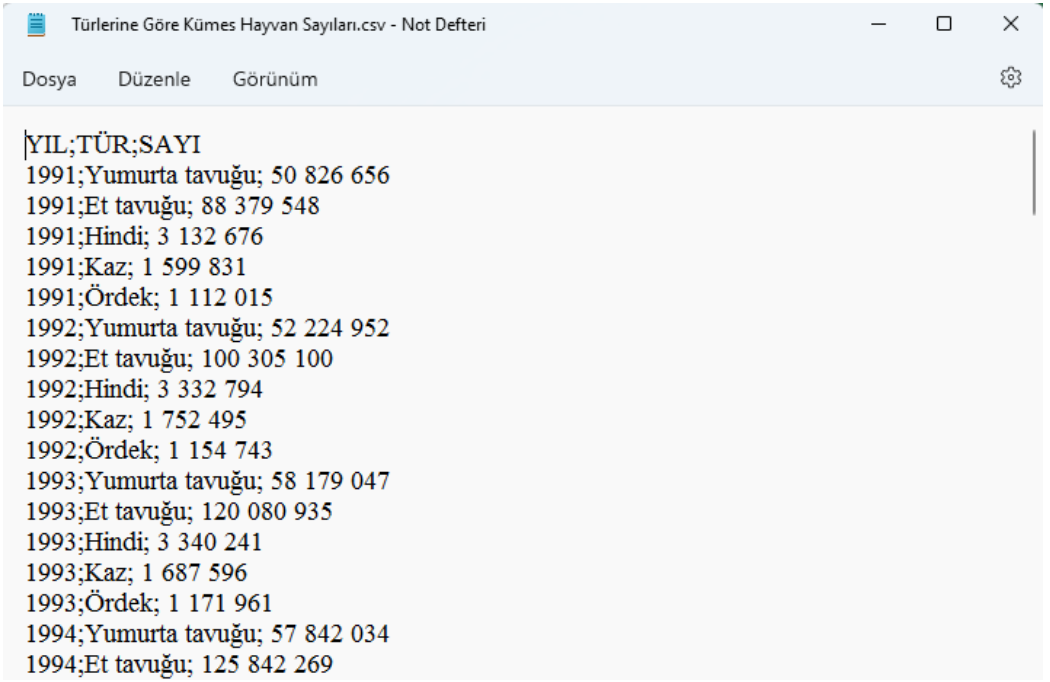
Görüldüğü üzere her iki dosya türü arasındaki farkı excel üzerinde anlamak mümkün değildir. Fakat iki dosya bir note defteri yardımı ile açılmak istendiğinde fark anlaşılabilmektedir.

Notepad İle Açma

XLS TO NOTEPAD

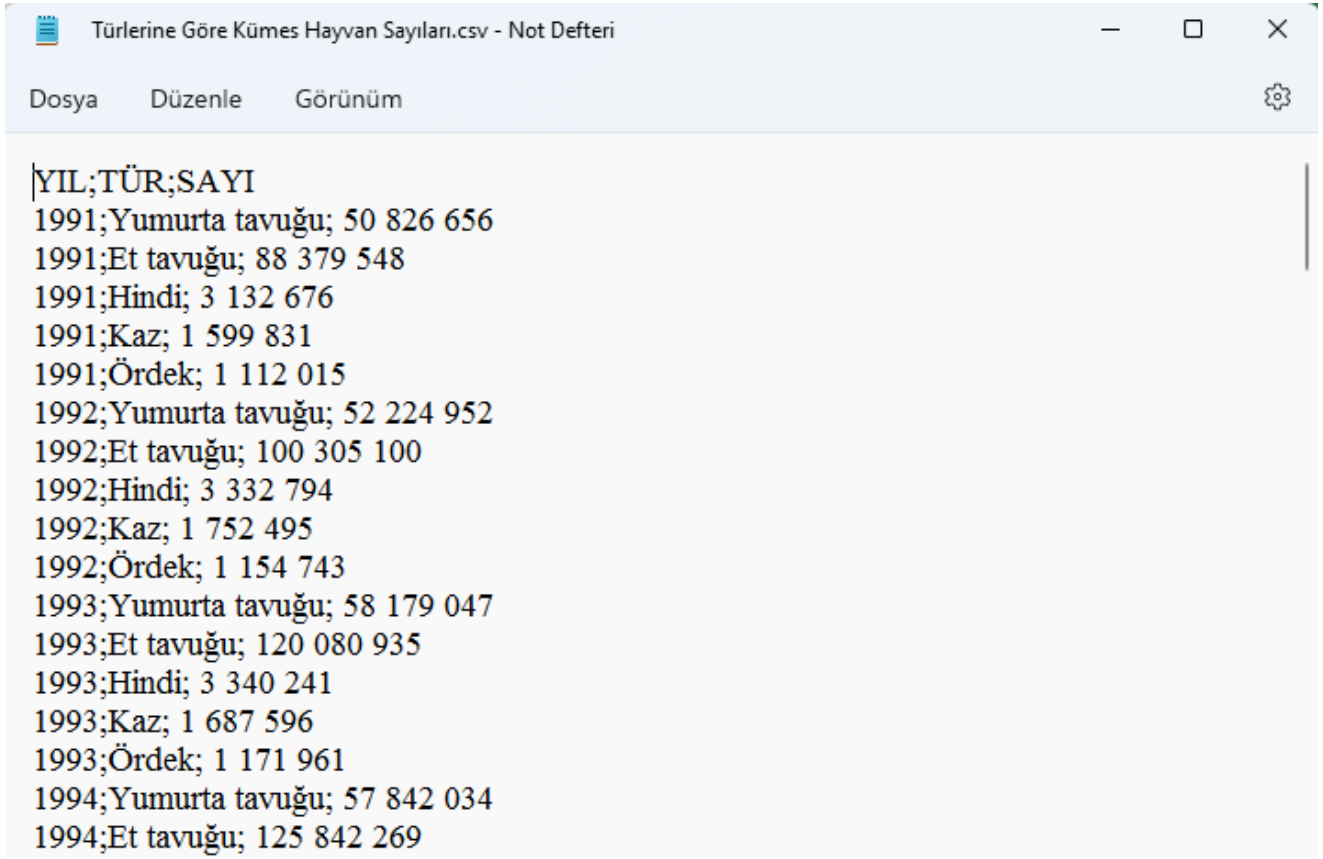


CSV TO NOTEPAD



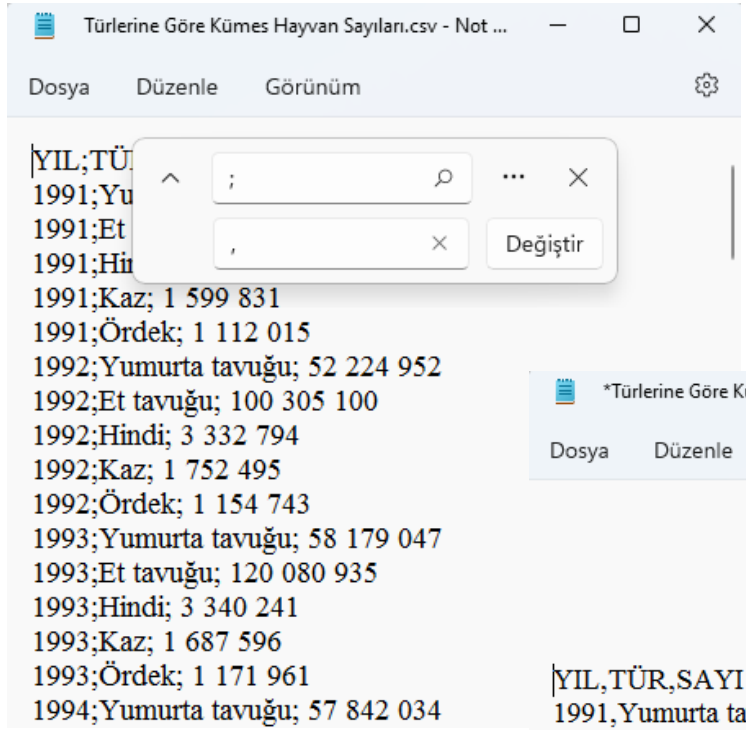
Csv to notepad

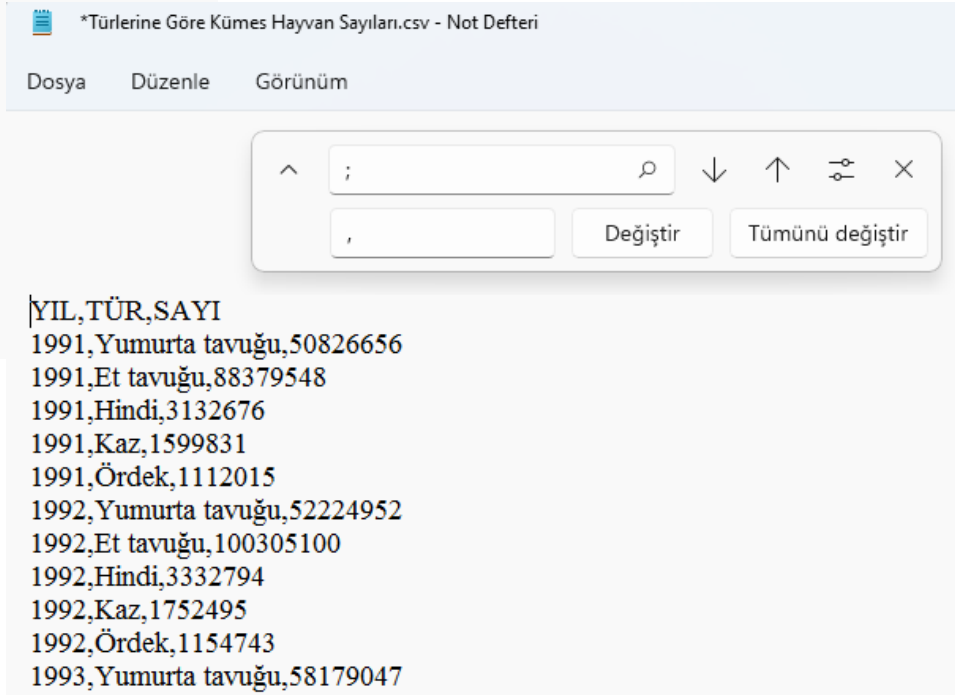
.csv olarak kaydettiğimiz dosyayı bir metin editörü (notdefteri,wordpad vb.) ile açtığımızda karşımıza yandaki gibi bir görüntü gelecektir. Bu aşamadan sonra gerçekleştireceğimiz ufak değişiklikler ile hedeflenen dosya özelliğine ulaşabileceğiz.



Düzenleme

Öncelikle bütün noktalı virgülleri virgüllere çeviriyoruz. (WEKA'da değerler , ile birbirinden ayrılmaktadır) Bunların değişimi için tüm programlardaki bul-değiştir kardeşliğinden faydalanacağız. Kullandığımız metin editöründe düzenle-değiştir-aranan değere ";" yeni değere ise "," yazıp tümünü değiştir dediğimizde bütün noktalı virgüller virgüle çevrilecektir. Dikkat edilmesi gerek diğer nokta sayısal veriler ondalık ise nokta ile ayrılmış olması tam değer ise aralarında boşlukların olmamasıdır.





Düzenleme

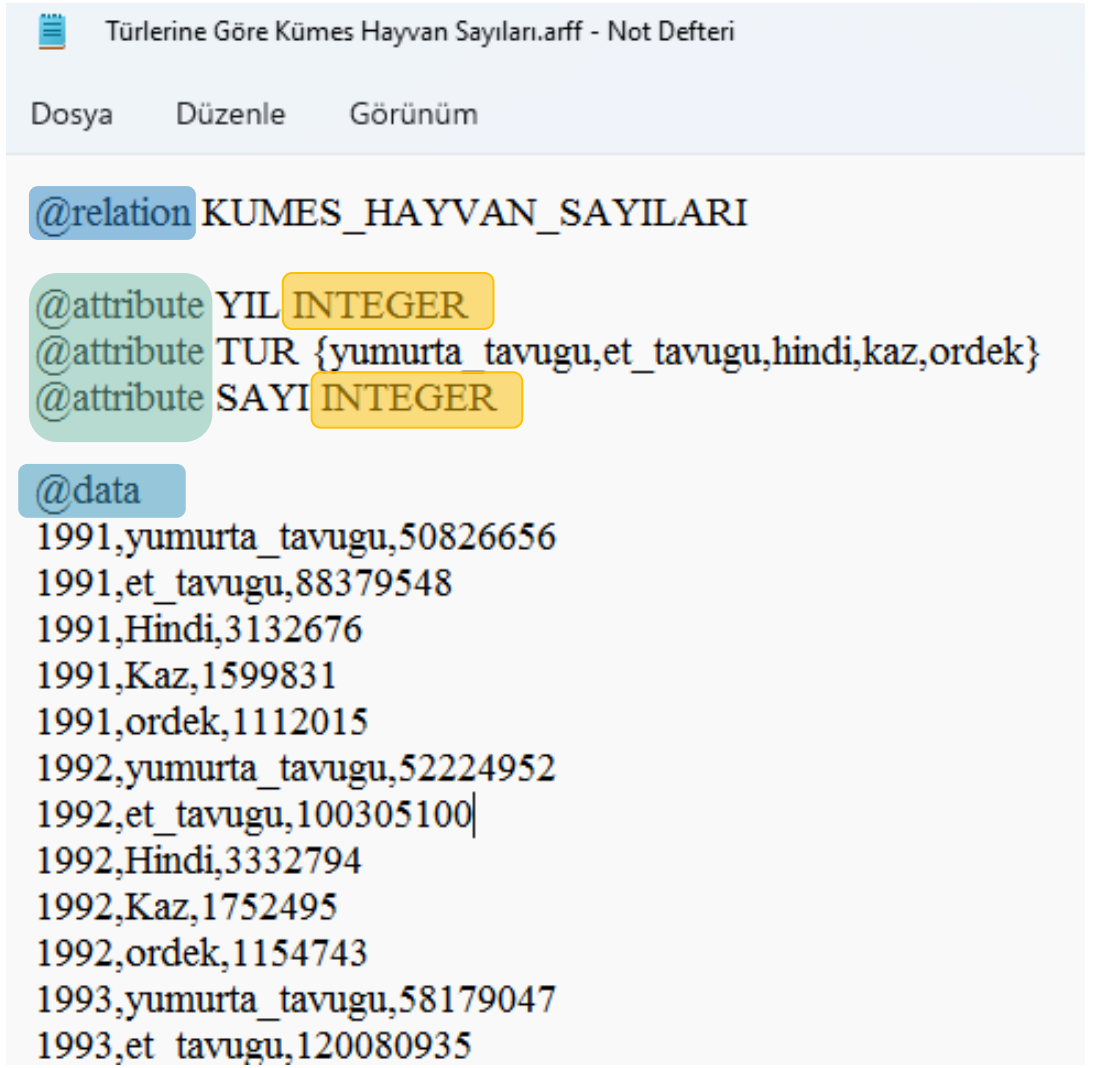
Dosyayı .arff formatına çevirmek için ilk satıra **@RELATİON** tagı ile birlikte **veri setimizin başlığını** yazıyoruz. HİÇ BİR ALANDA TÜRKÇE KARAKTER ve BOŞLUK KULLANMIYORUZ.

Sütun isimlerimizi (yıl, tür, sayı) ise verileri için **@ATTRİBUTE** tagı kullanıp, yanlarına veri tiplerini ekliyoruz.

•Eğer sütun ondalık sayısal ifadeler içeriyorsa **REAL** tam sayı ise **İNTEGER** olarak etiketliyoruz,

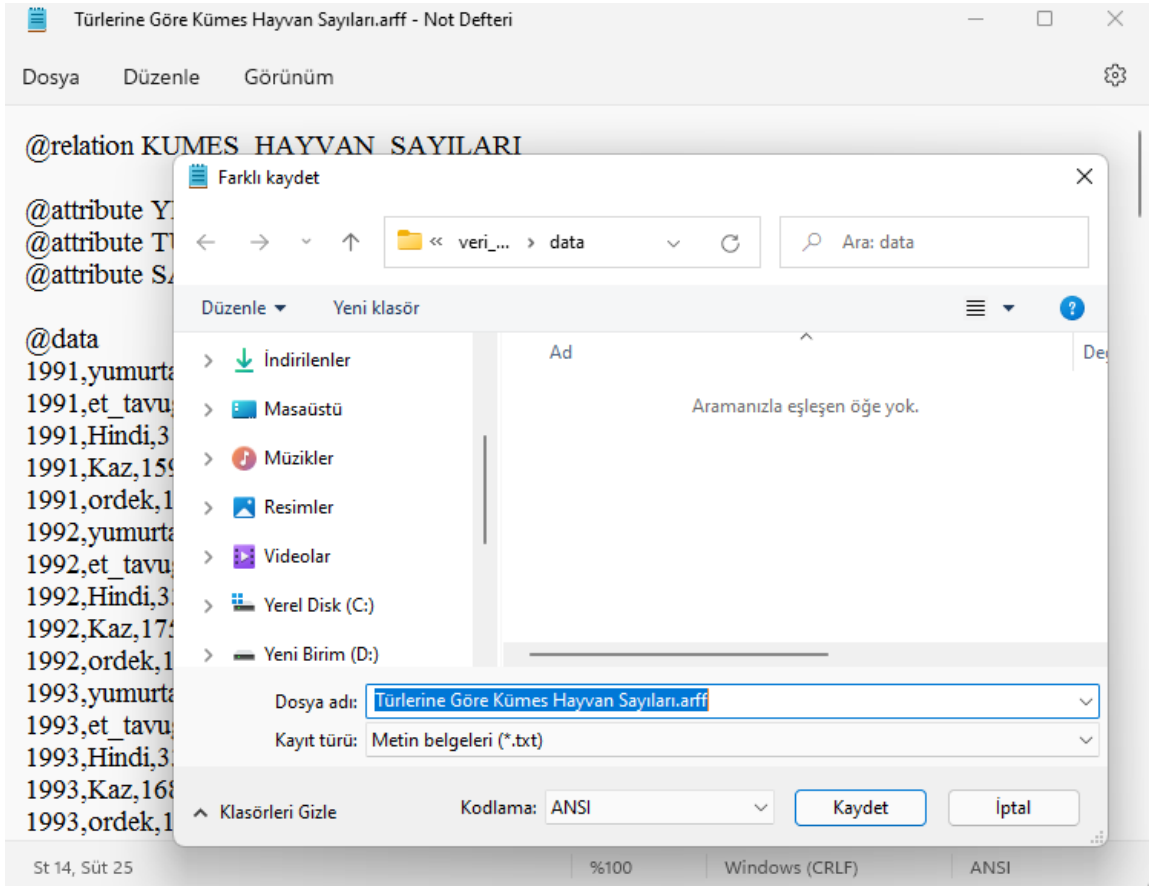
Metinsel ifadeler için **parantez içinde içerebileceği değerleri** belirtiyoruz.

Son olarak ise verimizin başladığı satırdan önce **@DATA** tagı ekliyoruz devamında verilere yer veriyoruz.



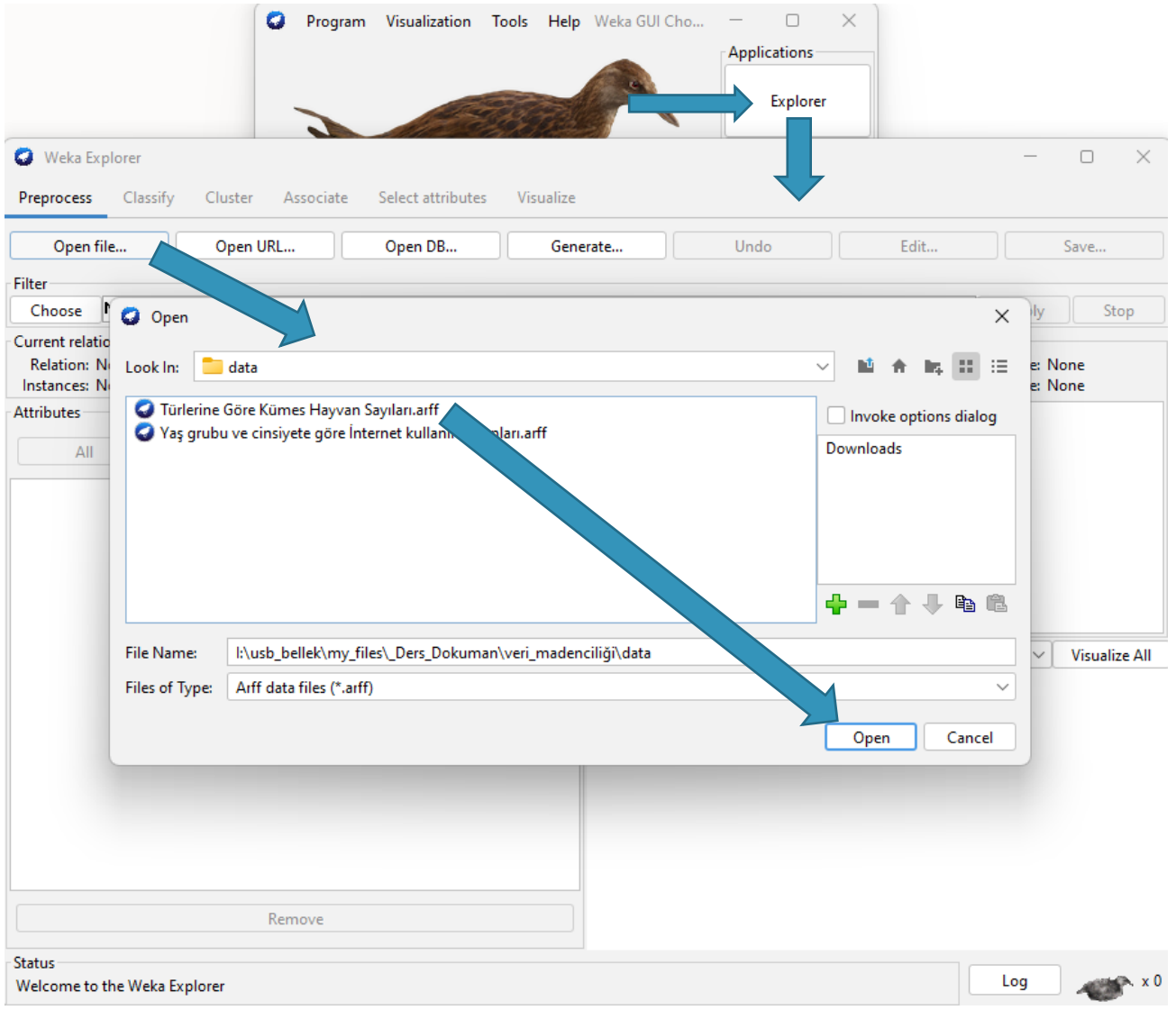
Kaydetme

Bu işlemler sonrasında Weka da çalışabilmek için farklı kaydet seçeneğinden dosyamızı .arff olarak kaydediyoruz.



Kaydetme

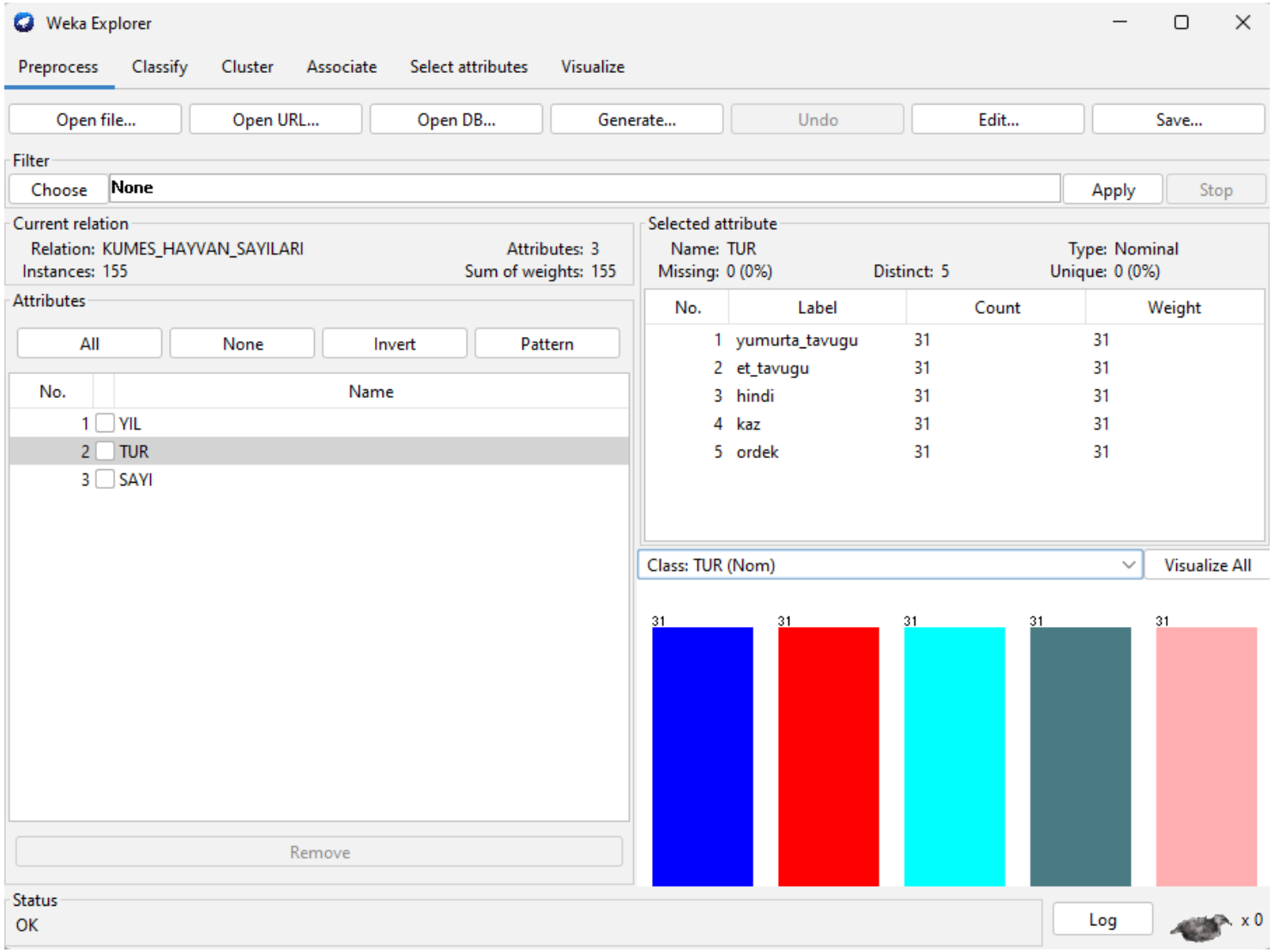
Weka yı çalıştırıp Explorer bölümüne tıkladıktan sonra Preprocess bölümünden open file diyerek .arff uzantılı dosyamızı açıyoruz.



Önizleme

Karşımıza gelen yandaki pencere; .arff uzantılı dosyamıza eklediğimiz sütun isimlerinin, sınıfların (class) ve özelliklerin (attribute) WEKA tarafından sağlıklı bir şekilde okunabildiğini göstermektedir. Bu aşamanın ardından gerekli çalışmaları yürütebiliriz.

Preproces bölümü bize dosyamız hakkında bilgi verir. Sütun isimleri, sınıflar (class), özellikler (attribute) vb.



DOSYA HAZIR;

Kaynaklar:
https://bilgisayarkavramlari.com
https://www.veribilimiokulu.com
https://zeynepozturkk.wordpress.com
https://kubracosar.blogspot.com
https://tr.myservername.com

Dr.Günay TEMÜR