

**YTÜ**  
**Bilgisayar Müh. Bölümü 2018-2019 Güz yy**  
**0114800 -Veri Madenciliğine Giriş**

**Proje Çalışması**

Ders kapsamında her öğrenciye ayrı ayrı çalışabileceği bir veriseti verilmiştir. Proje kapsamında aşağıda tanımlı işlemlerin gerçekleştirilmesi beklenmektedir. Verisetleri ile ilgili bilgiyi <http://archive.ics.uci.edu/ml/> adresinde bulabilirsiniz.

**1-WEKA ile çalışma:**

a-Veri Analizi: Veri setinin içeriği, kullanılan özelliklerin anlamı, sınıf sayısı ve isimleri, her sınıfa düşen örnek sayısı gibi bilgilerin çıkarılması gerekir. Veriseti içindeki özelliklerin sınıflamada ki ayırt ediciliklerine göre sıralanması ve eksik veri varsa giderilmesi yöntemlerinin araştırılması beklenmektedir. Ayrıca Aykırı Değer analizinin de yapılması gerekmektedir.

b-Sınıflama: Weka içinde yer alan sınıflama yöntemlerinden en başarılı olanların araştırılması ve 3 tanesinin seçilerek detaylı incelenmesi, sınıflama başarımlarının ve sınıf karışıklık matrislerinin karşılaştırılması gerekmektedir.

c-Kümeleme: Weka içinde yer alan kümeleme yöntemlerinden en başarılı olanların araştırılması ve 3 tanesinin seçilerek detaylı incelenmesi beklenmektedir. Kümeleme sonunda değerlendirme aşamasında örneklerin sınıf etiketlerinin kullanılarak çıkan kümelerin etiketlenmesi ve kümeleme başarımlarının karşılaştırılması gerekmektedir.

**2-Kodlama:**

Çalışma kapsamında bir sınıflama ve bir kümeleme yönteminin öğrenci tarafından kodlanması ve Weka ile karşılaştırılması yapılacaktır. Kodlamada MatLab veya C kullanabilirsiniz. Kodlanan yöntem ile ilgili hazır fonksiyonların kullanılmaması, yöntemin öğrenci tarafından anlaşılabilir şekilde koda dökülmesi gerekmektedir. Sınıflamada başarıyı ölçmek için kullanılacak test örneklerinin veriseti içinden “k-fold cross validation” için rasgele seçilerek bir altküme oluşturulması ve Weka ile analizde de aynı test örneklerinin kullanılması karşılaştırmayı mümkün kılacaktır.

**Değerlendirme:**

Proje kapsamında bir rapor hazırlanarak 18 Aralık 2018 Salı günü teslim edilmelidir. Öğrenci proje sunumları 18 Aralık Salı ve 25 Aralık Salı günleri yapılacaktır.

Doç.Dr.Songül Varlı  
12 Kasım 2018

## Konu Listesi

	Öğrenci No	Adı	Soyadı	Proje
SıraNo	Öğrenci No	Adı	Soyadı	Abalone
1	09011064	SELİM	ALKAN	Artificial Characters
2	10011067	YUNUS	YILDIRIM	Breast Cancer
3	11011064	BARTU	YAMAN	Breast cancer winconsin(orginal)
4	12011001	NİYAZİ	URAL	Adult
5	12011004	MURAD	KIZILAY	Annealing Data Set
6	12011015	FATİH	ÇOMAK	User Knowledge Modeling Data (Students' Knowledge Levels on DC Electrical Machines) Data Set
7	12011048	SALİH MERT	YİĞİT	Seismic-bumps
8	12011062	ECİR RECEP	GÜR	banknote authentication
9	12011079	OĞUZHAN	TİRYAKİ	Thoracic Surgery Data Data Set
10	13011014	TEVFİK EFE	TEKER	Qualitative_Bankruptcy Data Set
11	13011017	BARİŞ	CAN	Dermatology
12	13011052	FERHAT FİKRİ	MİNDER	Dexter Data Set
13	13011078	OĞUZHAN	ÖZTÜRK	Echocardiogram
14	13011118	SHARİFBEK	MUKHRANSKIY	Echoli Dataset
15	14011005	METİN MERT	AKÇAY	Credit Approval
16	14011012	TANSEL	GÖNÜL	Leaf
17	14011014	ÖYKÜ ZEYNEP	BAYRAMOĞLU	default of credit card clients Data Set
18	14011021	BETÜL	AKTAŞ	Haberman's Survival
19	14011024	BERKİN	İŞLEK	Forest type mapping Data Set
20	14011034	NERGİS	ARAFAL	Heart Disease
21	14011041	FERHAT	ATLİNAR	Hepatitis
22	14011045	YUSUF CEMAL	ÇELEBİ	Diabetic Retinopathy Debrecen Data Set
23	14011048	ENES	ERKOÇ	Urban Land Cover Data Set
24	14011058	MELİS	UYSAL	Detect Malicious Executable(AntiVirus) Data Set

25	14011059	EMRE	ÇETİN	Liver Disorders Dataset
26	14011065	ÇAĞATAY	BAY	Lung Cancer Data Set
27	14011070	SONER	PAYCI	Chronic_Kidney_Disease Data Set
28	14011073	MEHMET BURAK	İŞÇİ	Mammographic Mass
29	14011077	MELİH	KILIÇ	Smartphone-Based Recognition of Human Activities and Postural Transitions Data Set
30	14011078	EMRULLAH	YILMAZ	Mushroom
31	14011083	MİRAC FURKAN	GEMİCİ	Thyroid Dataset
32	14011085	ERKAN	KARABULUT	Polish companies bankruptcy data Data Set
33	14011090	ASLIHAN	ILGAZ	Drug consumption (quantified) Data Set
34	14011093	TUNAHAN	KOCATÜRK	Website Phishing Data Set
35	14011096	JASSER	DHAOUADI	Dota2 Games Results Data Set
36	14011501	ESİN	GEDİK	Gastrointestinal Lesions in Regular Colonoscopy Data Set
37	14011502	CEREN	KESKİN	Autistic Spectrum Disorder Screening Data for Children Data Set
38	14011904	MEHRİNİSO	MANGLIYEVA	Immunotherapy Dataset
39	14011907	MEKHTİ	GUMBATOV	Acute Inflammations Data Set
40	15011019	FURKAN	AKTAŞ	Student Academics Performance Data Set
41	15011022	HARUN	UZ	Abalone
42	15011026	MUSTAFA	BODUR	Artificial Characters
43	15011031	CEYDA	DEMİR	Breast Cancer
44	15011035	BURAK	TAHTACI	Breast cancer winconsin(orginal)
45	15011042	ÖMER FARUK	EKUKLU	Adult
46	15011056	GENCO	COŞGUN	Annealing Data Set
47	15011059	FURKAN	GERÇEK	User Knowledge Modeling Data (Students' Knowledge Levels on DC Electrical Machines) Data Set
48	15011083	YUSUF KÜRŞAD	KAYA	Seismic-bumps
49	15011086	BURAK	BAL	banknote authentication
50	15011087	GÖRKEM	ŞAHİN	Thoracic Surgery Data Data Set

51	15011088	MELİHA	ERGİNCAN	Qualitative_Bankruptcy Data Set
52	15011112	ABDURRAHİM	DERRIC	Dermatology
53	15011804	MUHAMMED YASİN	SAĞLAM	Dexter Data Set
54	15011901	NATIQ	AGHAYEV	Echocardiogram
55	16011060	FATMA BETÜL	ÇİFTÇİ	Echoli Dataset
56	16011601	BERK	SUDAN	Credit Approval
57	16011602	MİRAY ASENA	AYGÜZEL	Leaf
58	17011601	İPEK	UYANIK	default of credit card clients Data Set
59	17011603	ÇETİN	TEKİN	Haberman's Survival
60	17011604	HAVANUR	DERVİŞOĞLU	Forest type mapping Data Set
61	Erasmus1			Heart Disease
62	Erasmus2			Hepatitis