

**INTRODUCTION TO DATA MINING  
PROJECT REPORT**

**STUDENT  
FATİH ÇOMAK  
12011015**

**LECTURER  
DOÇ. DR. SONGÜL VARLI**

**DATASET  
ADULT**

**DUE  
10 DECEMBER 2019**

1. WEKA

a) Data Analysis

Adult veri seti, yetişkin veri seti olarak da nitelendirebileceğimiz nüfus sayımı verilerinin toplandığı veri kümesidir.

Bu veri setindeki özniteliklerin bize verdiği değer, yıllık gelirin 50K $’ı geçip geçmediğini sunar. Bu veri seti, ayrıca nüfus sayımı geliri veri seti olarak da bilinir.

Kayıp değerler vardır, bunların çözülmesi için diğer değerlerin ortalamasının bu missing values’a aktarılması düşünülmüştür.

adult.data isimli data dosyamızda **32561** adet kişi bilgisi bulunur.

Ayrıca test yapmak için veri seti adult.test isimli **16281** kişilik bir veri seti sunmuştur. Bu data dosyaları weka’da çalıştırılabilmesi için .arff dosya formatına dönüştürülmüştür.

15 adet öznitelik bulunur:

1. **Age(yaş)** :sayısal bir değer olarak yaş bilgisini verir.
2. **Workclass**: kategorik olarak çalışma şekli bilgisi verir.
3. **Fnlwgt**: sayısal değer verir.
4. **Education**: kategorik şekilde eğitim durumunu belirtir.
5. **education-num**: sayısal değer olarak eğitim durumunu verir.
6. **marital-status**: medeni hal(kategorik).
7. **occupation**:meslek dalı(kategorik).
8. **relationship**:ilişkisi(kategorik).
9. **race**:ırk(kategorik).
10. **sex**: cinsiyet(kategorik).
11. **capital-gain**: sermaye kazancı(sayısal).
12. **capital-loss**:sermaye kaybı(sayısal).
13. **hours-per-week**: haftalık çalışma saati(sayısal).
14. **native-country**: ülke bilgisi verir(kategorik).
15. **class**: 50K$’dan büyük mü küçük mü şeklinde 2 kategori verir.

Attribute açıklamaları da şöyle:

age: continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: continuous.  
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

education-num: continuous.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.  
sex: Female, Male.

capital-gain: continuous.

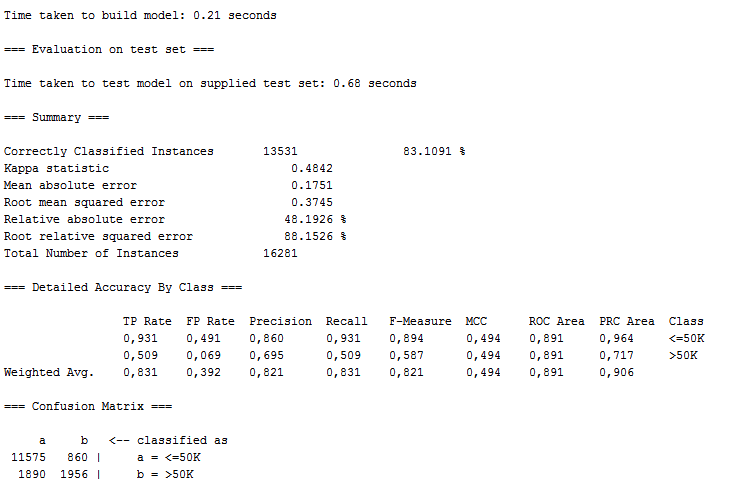
capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

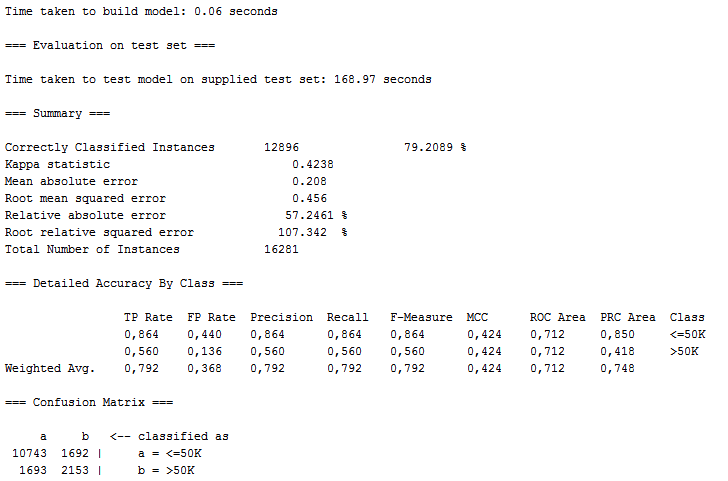
Class: >50K, <=50K.

b) Classifying

1)NaiveBayes

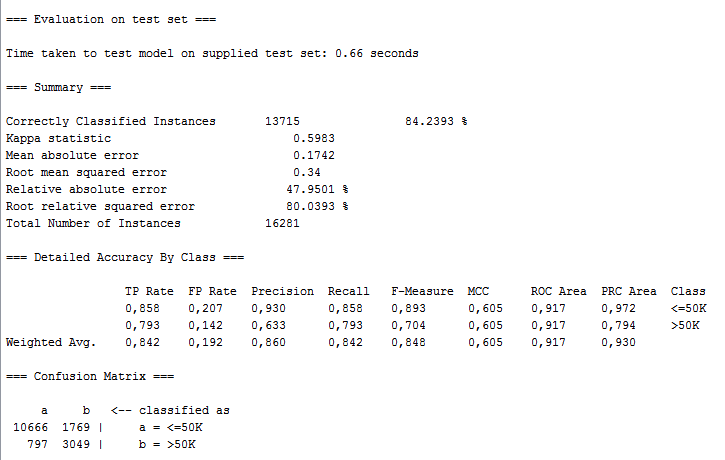
83%lük bir başarı elde edildi.

2)lazy.IBk



79%’luk bir başarı elde edildi.

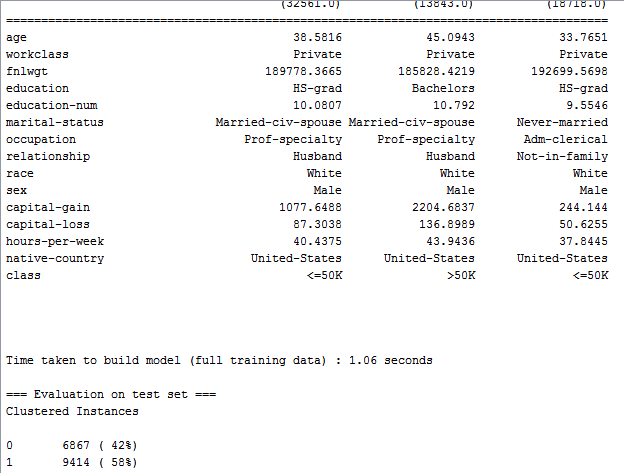
3) BayesNet



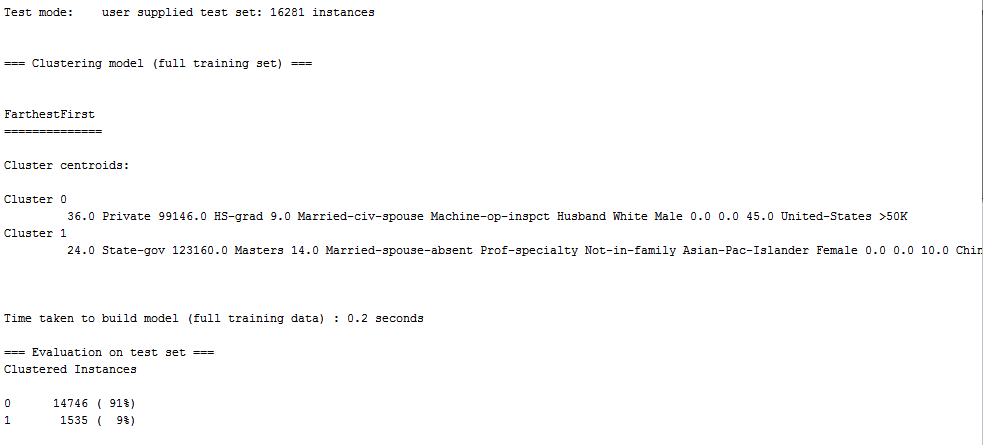
84%’lük bir başarı elde edildi.

b) Clustering

1) SimpleKMeans



2) FarthestFirst



3) Canopy

Time taken to build model (full training data) : 1.29 seconds

=== Evaluation on test set ===

Clustered Instances

0 817 ( 5%)

1 416 ( 3%)

2 224 ( 1%)

3 144 ( 1%)

4 337 ( 2%)

5 887 ( 5%)

6 745 ( 5%)

7 437 ( 3%)

8 540 ( 3%)

9 131 ( 1%)

10 385 ( 2%)

11 279 ( 2%)

12 192 ( 1%)

13 35 ( 0%)

14 126 ( 1%)

15 84 ( 1%)

16 115 ( 1%)

17 198 ( 1%)

18 371 ( 2%)

19 291 ( 2%)

20 468 ( 3%)

21 152 ( 1%)

22 389 ( 2%)

23 156 ( 1%)

24 19 ( 0%)

25 547 ( 3%)

26 396 ( 2%)

27 18 ( 0%)

28 597 ( 4%)

29 54 ( 0%)

30 94 ( 1%)

31 69 ( 0%)

32 25 ( 0%)

33 60 ( 0%)

34 32 ( 0%)

35 9 ( 0%)

36 56 ( 0%)

37 104 ( 1%)

38 11 ( 0%)

39 121 ( 1%)

40 110 ( 1%)

41 62 ( 0%)

42 76 ( 0%)

43 46 ( 0%)

44 103 ( 1%)

45 60 ( 0%)

46 388 ( 2%)

47 159 ( 1%)

48 39 ( 0%)

49 142 ( 1%)

50 76 ( 0%)

51 222 ( 1%)

52 153 ( 1%)

53 23 ( 0%)

54 40 ( 0%)

55 104 ( 1%)

56 55 ( 0%)

57 138 ( 1%)

58 5 ( 0%)

59 254 ( 2%)

60 108 ( 1%)

61 197 ( 1%)

62 285 ( 2%)

63 79 ( 0%)

64 19 ( 0%)

65 86 ( 1%)

66 71 ( 0%)

67 9 ( 0%)

68 134 ( 1%)

69 141 ( 1%)

70 111 ( 1%)

71 28 ( 0%)

72 71 ( 0%)

73 45 ( 0%)

74 159 ( 1%)

75 499 ( 3%)

76 173 ( 1%)

77 39 ( 0%)

78 60 ( 0%)

79 39 ( 0%)

80 112 ( 1%)

81 188 ( 1%)

82 31 ( 0%)

83 78 ( 0%)

84 23 ( 0%)

85 53 ( 0%)

86 34 ( 0%)

87 93 ( 1%)

88 38 ( 0%)

89 70 ( 0%)

90 17 ( 0%)

91 124 ( 1%)

92 89 ( 1%)

93 51 ( 0%)

94 49 ( 0%)

95 177 ( 1%)

96 96 ( 1%)

97 22 ( 0%)

98 87 ( 1%)

99 110 ( 1%)