Veri Ambarlama ve Veri Madenciliği Dönem Projesi Raporu

**[Türkiye Öğrenci Değerlendirme]**

Öğrenci Adı – Soyadı: Recep Özkara

Öğrenci Numarası: 19yobi1020

**Problemin Tanımı**

[*Bu kısımda, proje kapsamında çözdüğünüz problemi açıklayınız (müşteriniz sizden ne istedi / hangi soruyu cevaplandırmanızı istedi?)*]

Başarılı-başarısızlık ölçütlerine dayanarak “Dersin amaç ve hedefleri dönemin başında açıkça belirtilmişti” özniteliğinin 3 veya 3 ün üzerinde değerlendirme alması yani başarılı olduğunun olasılığı kaçtır? problemini çözmeye çalıştım.

**Veri Edinimi ve Veri Seti Özellikleri**

[*Bu kısımda veri setini nasıl elde ettiğinizi açıklayınız, hangi kaynağı kullandığınız, veri setini hangi kriterlere göre seçtiniz, veri setinizin kısa tanıtımı, hangi konudadır, kaç veri bileşeni var, kaç öznitelik var, bu özniteliklerin anlamları nedir vb*]

Kullandığım veri setini UCI – Machine Learning Repository içerisinden buldum. Veri setimin ismi Türkiye öğrenci değerlendirme veri seti. Kullandığım veri setinin konusu ise öğrecilerin başarıları ve öğrenci aktiviteleri ile ilgili. Bunun yanında öğrencinin aldığı dersin öğretim görevlisinin nasıl işlediği, dersi nasıl hazırladığı, ders ile ilgili yorumlar bulunmaktadır. Veri setini seçerken baz aldığım kriterler başında veri setinin içerisinde 5820 adet verinin olmasıydı. Benim için veri bileşeni ne kadar fazla olursa anlamlı bilgiler çıkarabilmek o kadar kolay olacağından seçtim. Bunun yanında veri setimin içerisinde 33 tane öznitelik bulunmaktadır. Bu öznitelikler ise şu şekildedir:

instr: Eğitmenin tanımlayıcısı; {1,2,3}

sınıfından alınan değerler : Ders kodu (tanımlayıcı); {1-13}

tekrarından alınan değerler : Öğrencinin bu dersi kaç kez aldığı; {0,1,2,3,...}

katılımdan alınan değerler : Katılım düzeyinin kodu; {0, 1, 2, 3, 4}

zorluktan değerler: Öğrencinin algıladığı şekliyle dersin zorluk seviyesi; {1,2,3,4,5}

Q1'den alınan değerler : Başlangıçta dönem ders içeriği, öğretim yöntemi ve değerlendirme sistemi verilmiştir.

Q2: Dersin amaç ve hedefleri dönemin başında açıkça belirtilmişti.

Q3: Kurs, kendisine verilen kredi kadardı.

Q4: Ders, dersin ilk günü açıklanan müfredata göre işlendi.

Q5: Sınıf tartışmaları, ödevler, uygulamalar ve çalışmalar tatmin ediciydi.

Q6: Ders kitabı ve diğer ders kaynakları yeterli ve günceldi.

Q7: Ders, alan çalışması, uygulama, laboratuvar, tartışma ve diğer çalışmalara izin vermektedir.

Q8: Kısa sınavlar, ödevler, projeler ve sınavlar öğrenmeye yardımcı oldu.

Q9: Dersten çok keyif aldım ve derslere aktif olarak katılmaya istekliydim.

Q10: Dönem veya yılın sonunda kursla ilgili ilk beklentilerim karşılandı.

Q11: Kurs, mesleki gelişimim için alakalı ve faydalıydı.

Q12: Kurs, hayata ve dünyaya yeni bir bakış açısıyla bakmama yardımcı oldu.

Q13: Eğitmenin bilgisi ilgili ve günceldi.

Q14: Eğitmen derslere hazırlıklı geldi.

Q15: Öğretim elemanı ilan edilen ders planına göre ders işledi.

Q16: Eğitmen kursa bağlıydı ve anlaşılırdı.

Q17: Eğitmen derslere zamanında geldi.

Q18: Eğitmenin akıcı ve takip etmesi kolay bir sunum/konuşması var.

Q19: Eğitmen ders saatlerini etkin kullandı.

Q20: Eğitmen dersi açıkladı ve öğrencilere yardımcı olmaya hevesliydi.

Q21: Eğitmen öğrencilere olumlu bir yaklaşım sergiledi.

Q22: Öğretim elemanı öğrencilerin dersle ilgili görüşlerine açık ve saygılı davranmıştır.

Q23: Eğitmen kursa katılımı teşvik etti.

Q24: Eğitmen ilgili ödevleri/projeleri verdi ve öğrencilere yardım etti/rehberlik etti.

Q25: Eğitmen, kurs içindeki ve dışındaki kursla ilgili soruları yanıtladı.

Q26: Öğretim üyesinin değerlendirme sistemi (ara sınav ve final soruları, projeler, ödevler vb.) ders hedeflerini etkin bir şekilde ölçmüştür.

Q27: Öğretim elemanı sınavlara çözümler sunmuş ve öğrencilerle tartışmıştır.

Q28: Öğretim elemanı tüm öğrencilere doğru ve tarafsız davrandı.

Q1-Q28'in tümü değerlendirme ölçütüdür. {1,2,3,4,5}'dan alınmıştır.

**Yöntemin Belirlenmesi**

[*Bu kısımda hangi veri madenciliği fonksiyonunu/fonksiyonlarını seçtiğinizi, bu fonksiyon/fonksiyonlar altında hangi yöntemi(algoritmayı)/yöntemleri(algoritmaları) belirlediğinizi, bunları belirlerken hangi yöntemi kullandığınızı, nasıl uyguladığınızı, uygulama öncesinde veri seti üzerinde bir ön işleme yapmanız gerektiyse bu ön işleme adımlarını anlatınız.*]

Veri seti için veri madenciliği fonksiyonlarından olan tanımlayıcı fonksiyon altındaki kümeleme yönetimini seçtim. Çünkü veri sayısı fazla olduğu zaman bu verileri kümelere ayırıp yorumlaması daha kolay. Kümeleme içinde SimpleKmeans yönetimini kullandım. Veri setim için herhangi bir ön işleme yapmadım. Bunun yanında sınıflandırma fonksiyonunuda kullandım. Sınıflandırma fonksiyonu için ise logistic yönetimini ele aldım. Sınıflandırma olarak değerlendirme ölçütleri kullanıldı. Değerlendirme ölçütleri 1 den 5 e kadar olan sayılardır.

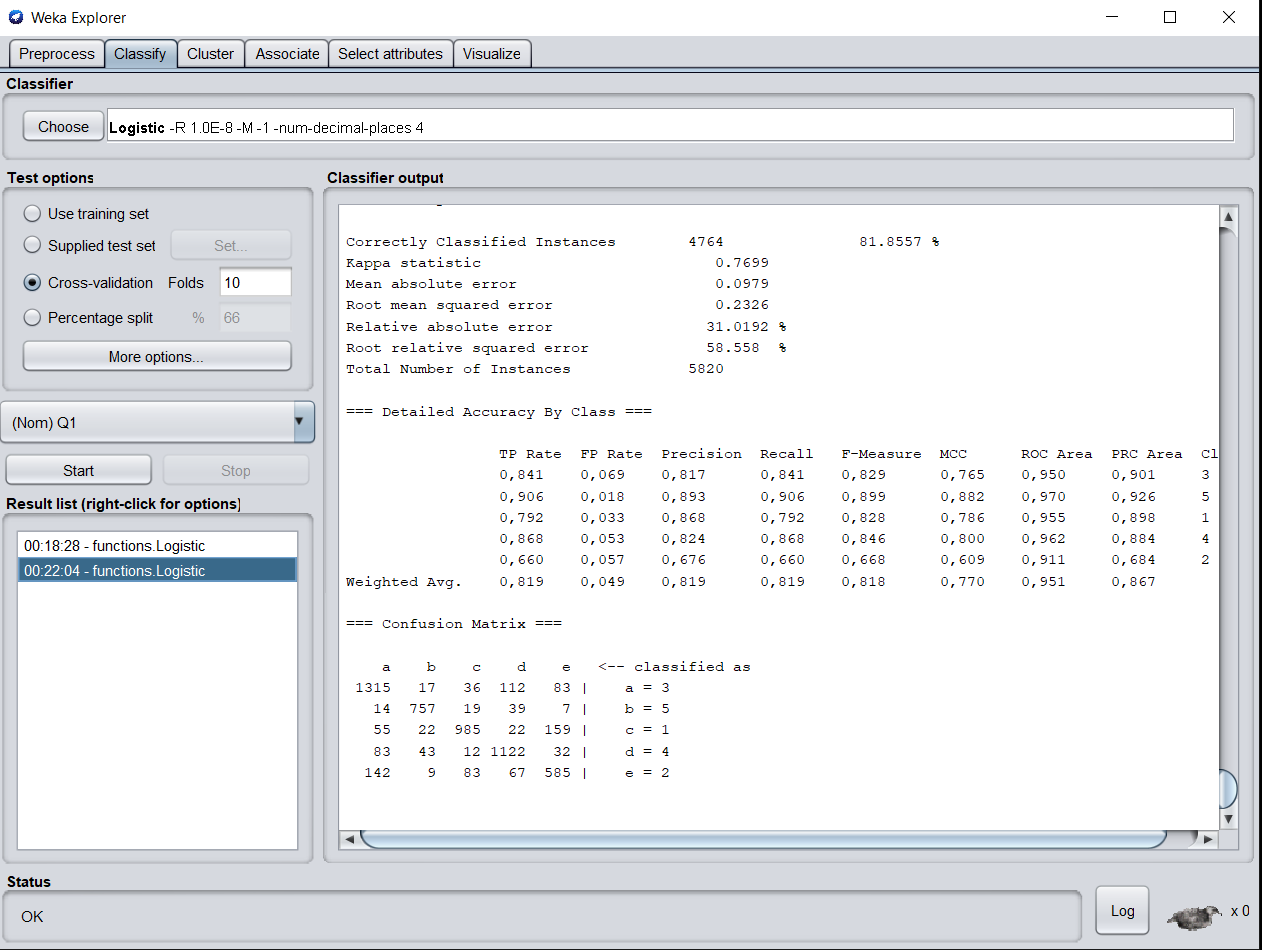
**Veri Madenciliği Fonksiyonunun Çalıştırılması**

[*Bu kısımda seçtiğiniz fonksiyon/fonksiyonlar altında yer alan algoritma/algoritmaları veri setine nasıl uyguladığınızı açıklayınız. Uygulama neticesinde elde ettiğiniz değerleri sununuz. Uygulama sürecinde karşılaştığınız sorunlar ve bu sorunlara getirdiğiniz çözümleri açıklayınız.*]

Öncelikle kullandığım kaynağın sitesinden veri setini indirirken csv dosyası olarak indirildi. İçerisindeki veriler düzensizdi, veriler hep virgül ile ayrılmıştı bunun için arff formatına çevrilmiyordu. Csv dosyasındaki verileri düzenlemem gerekti ve bilgisayar ayarlar kısmından ek tarih ve saat kısmından gerekli olan ayarlamaları yaptım. Ayarları yaptıktan sonra veriler düzenli bir şekilde birbirinden ayrıldı. Ondan sonra o csv uzantılı excel dosyasını arff ye dönüştürürken Excel2Arff Converter Programını kullandım. O Program içerisinde excel dosyasını kullanmam için excel dosyasını xls uzantılı olması gerekiyordu ve xls uzantısına çevirdim. Ondan sonra gerekli işlemleri yaparak excel dosyasını arff formatına dönüştürdüm. Ondan sonra Weka uygulamasını çalıştırdım Applcations kısmından explorer diyip sonra gelen pencereden sayfanın yukarısında bulunan menüden cluster bölümünden Choose butonuna tıklayarak orda bulunan SimpleKMeans algoritmasına tıklayarak gerekli olan işlemleri yaptım ve çıktıları elde ettim. Bunun yanında Classify bölümünden logistic yönetimini kullanarak değerleri elde ettim.

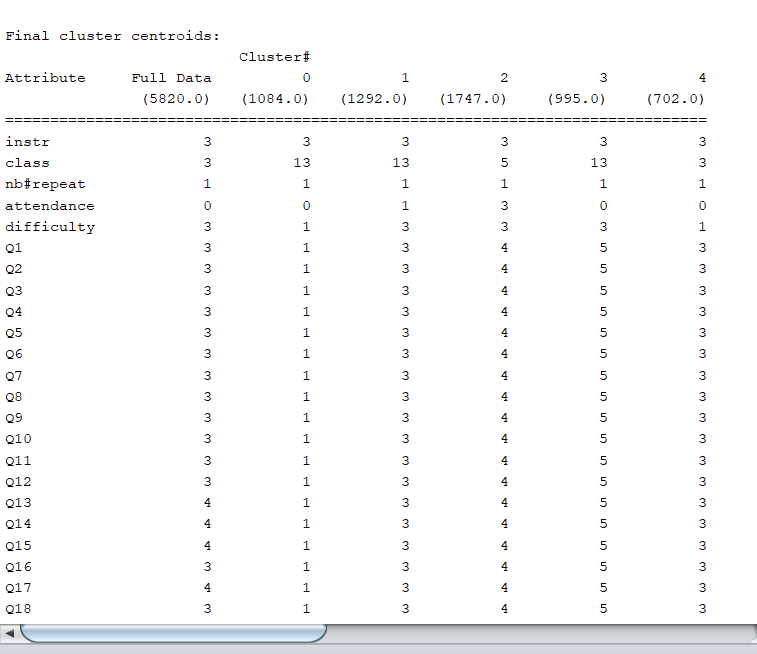
**LOGISTIC**

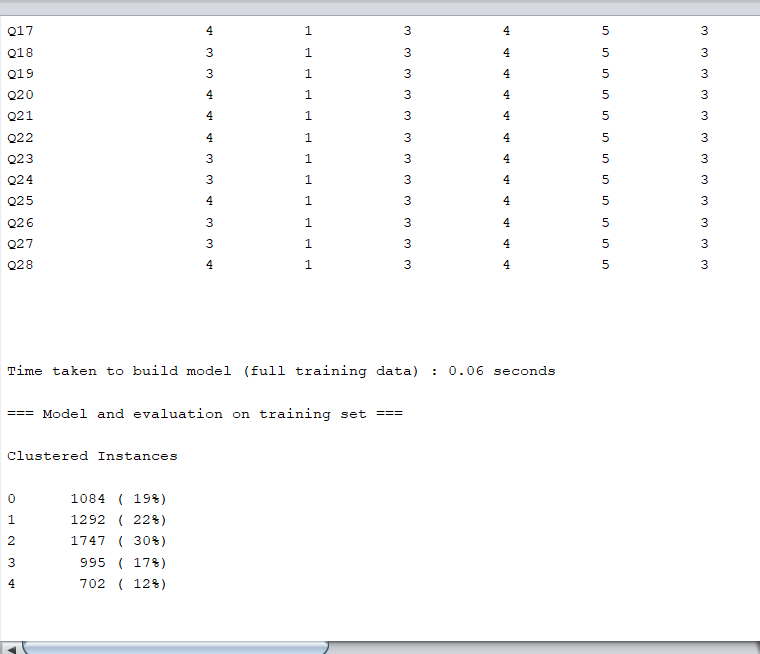
Logistic yönetimini kullanarak Q1’in cevabını verecek çıktıdır. Yani ders içeriğinin dönem başında öğrencilere verilmesi ile alakalı bir durum.



**SimpleKMeans**

Projem için kümele fonksiyonunuda kullandığımı belirtmiştim. Veri setimde küme sayımı 5 olarak ayarladım ve aşağıdaki çıktıları elde ettim. 5 olarak ayarlamamın sebebi ise verilerin daha anlamlı dağıldığını gördüm ve hata oranımında azaldığını gördüğüm için küme sayısını 5 yaptım. Burdan da görüleceği üzere 5 kümenin 4 ünde Q1 oranı 3 ve 3’ün üzerindedir sadece 1 tanesinde Q1 oranı 1 dir. Ortalama olarak ta zaten Q1 değeri 3 olarak gözükmektedir.

****

****

**Sonuçların Elde Edilmesi ve Yorumlama**

[*Bu kısımda yöntemin uygulanması neticesinde elde edilen sonuçları sayısal olarak sununuz ve kalite ölçümlerini yapınız. Kalite ölçümleri sonuçlarınızı yorumlayınız. Müşterinize sunulabilecek sonuçlardan örnekler veriniz.*]

Öncelikle snıfılandırma yaparken sınıflandırmayı değer ölçütüne göre yaptı. Bu değer ölçütleri 1 den 5 e kadar olan sayılardır. 1(Çok kötü), 2(kötü), 3(orta), 4(iyi), 5(çok iyi) anlamındadır.

Yani çok başarısız, başarısız, orta-başarılı, başarılı, çok başarılı şeklindedir.

A’nın orta değerini aldığını doğru söğleyen kişi sayısı 1315 dir. 248 kişi ise A’nın 3 olmadığını söylemektedir. B sınıfının çok başarılı olduğunu söyleyen kişi sayısı ise 757 olduğu görülmektedir. Geriye kalan 79 kişi ise çok başarılı yani 5 olmadığını söyler ve yanlış söylemiştir. C sınıfına baktığımızda 985 kişi çok kötü yani çok başarısız olduğunu doğru bilmiştir. Ama 258 kişi ise çok başarısız bulmamıştır. D sınıfının başarılı olduğunu söyleyen kişi sayısı 1122 dir ve doğru bilmiştir. 170 kişi ise yanlış bilmiştir. E sınıfının karşılığı 2 olan değeri doğru söyleyen kişi sayısı 585, yanlış söyleyen kişi sayısı ise 301 dir.

Bunları % olarak ifade edecek olursak, a sınıfı %81’I orta olduğunu doğru söylemiş, %19 ise yanlış olduğunu söylemiştir. Yani Orta-başarılı bulmuştur. B sınıfı %90’I çok başarılı demiştir.

C sınıfının doğru söyleyenlerin yüzdesi ise %74 tür yani %74 çok kötü, %26 ise çok kötü değil yorumunu yapmıştır. D sınıfına baktığımızda başarılı olduğunu doğru söyleyenlerin oranı %85 iken %15’ I yanlış bilmiştir. E sınıfında ise %49’u başarısız diyerek doğru bilmiştir ama geriye kalan %51 ise başarısız değil diyerek yanlış bilmiştir.

Tüm bunlara bakılarak A sınıfında bulunan 1609 kişiden %82 olasıklıkla 3 değerlendirmesini yaparak orta olduğunu söylemiştir.

B sınıfında toplam 848 kişiden %90’I çok başarılı yorumunu yapmıştır ve 5 değerini verdi.

C sınıfında toplam 1135 kişiden %88’i çok kötü yorumunda yani 1 değerini vermiştir.

D sınıfında toplam 1362 kişi bulunuyor ve bunlardan %82.5 olasılıkla başarılı bulup 4 değerini vermiştir.

E sınıfına baktığımızda ise 866 kişiden %67.5 olasılıkla kötü yorumunda bulunmuş ve 2 değerlendirmesini vermiştir.

Başarılı olan grupları başarısız olanlara karşı oranladığımız zaman ise %68 olasılıkla kişiler belirtilen problem için 3 veya 3’ün üzerinde değerlendirmesini yapıp, genellikle başarılı bulunmuştur.

SimpleKmeans çıktılarına baktığımız zaman ise veriyi 5 kümeye ayırmıştık. Elde edilen çıktılara göre 5 kümenin 4’ü 3 ve 3’ün üzerinde bir değerlendirme aldı. Zaten ortalama olarak Q1 satırına bakıcak olursak ortalama değerini 3 olarak gözükmektedir.