Ödev 1

Soru 1

27.05.2021 de (bugün, 03.01.2023) github'da "fake news detection" adında veri madenciliği projesi paylaşılmış. Bu projede sahte haber tanıma sistemi yapmışlar:

- Veri İşlemesi
- Veri temizlemesi
- veri dönüşümü
- Makine öğrenmesi algoritmaları

teknikelerini kullanarak spChalk kullanıcı adlı bir araştırmacı tarafından Jupyter Notebook'da yapılmıştır. (Lojistic Regresyon, Naive Bayes, SVM ve Random Forests)

Soru 2

a) Veri madenciliği yaptığımız verilere genellikle bir veritabanından, veya hazırlanmış veri toplayan sensörler tarafından toplarız veya ulaşırız. Bu veri toplanan veya veriye ulaşma esnasındaki, sensör hataları, opsiyonel giriş olanları kullanıcı hataları, sistem hataları yüzünden veriler ve bazı nitelikleri eksik gelebilir.

Bir niteliğin durumu başka bir niteliğe bağlı olabilir.
Bir niteliğin uygunluğu diğer niteliğe bağlı olabilir. Örneğin çocukların Sabit geliri olmaz. İnsanlar tablosu oluşturulsa ve sabit gelir adında bir nitelik bulunsa. Çocuklar için bu alan boş bırakılmak zorunda olacak.

b) Eksik nitelik değerlerinin önüne geçmek için, **silme yöntemi**, **sabit değer tanımlama**, analiz sırasında **görmezden gelme** ve olabilecek bir diğer yöntemleri doldurma tekniklerini kullanabiliriz.

- Eksik değeri olan niteliğe sahip tüm verileri silebilirsiniz, elimine edebilirsiniz. Bunu ancak çok büyük veri setlerinde yapmak mantıklı olur çünkü elimizdeki veri setini daha da küçültmek istemeyiz.
- Eksik değer içeren niteliğe sahip veri parçasını, diğer verilere bakarak ortalamayı hesaplayarak bir veri girebilirsiniz. Bu durumda hem veri setiniz aynı boyutta kalır hem de ortalamasını aldığımız için analiz esnasında büyük, complex bir etki yaratmaz.
- Verileri görmezden gelebilecek kadar küçük katsayılar verirsiniz.

Soru 3

Çok fazla özniteliği olan bir veri setinde özniteliklerin azaltma adına şunlar yapılabilir:

Missing Value Ratio (Eksik değer oranı), belli bir sayıdan daha fazla eksik değer içeren sütunların, analiz sırasında büyük farklar yaratmayacağı bilinir. Bu nedenle bu sütunları silebilirsiniz.

Önceki metodda benzer olarak, hücrelerinde çok az değişim oranı olan sütunlar analiz sırasında büyük değişikliklere yol açmazlar. Bu nedenle bu sütunları da silebiliriz, görmezden gelebiliriz.

Diğer bir feature selection tekniği ise Backward Feature Elimination (Geriye dönük eleme) metodu. Bu metod önce tüm değerler ile işlem yapar. Ardından veri setinden sütunları silerek işlem yapmaya devam eder. Elde edilen değişikliğe göre sütunun silinmesini veya katma değerini hesaplar. Bu şekilde katma değeri düşük olan sütunların silinmesi mantıklıdır.