KAHRAMANMARAŞ SÜTÇÜ İMAM ÜNİVERSİTESİ
MÜHENDİSLİK VE MİMARLIK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
ÖRÜNTÜ TANIMA DERSİ

ÖĞRENCİ İŞ PROFİLİ SINIFLANDIRMA

ALPEREN USLU, 20110131048

Dr.Öğr.Üyesi FAHRİYE GEMCİ

OCAK 2024

1. Problem Hakkında Bilgiler.

Problemin tanımı

Bilgisayar bilimleri ve mühendislik öğrencilerinin teknik ve yumuşak becerileri üzerine kapsamlı bir veri seti mevcuttur. Ancak, bu veri setinin, öğrencilerin belirli profesyonel profillere ne kadar uygun olduğunu belirlemek için nasıl kullanılacağı konusunda bir belirsizlik vardır. Bu durum, eğitim kurumları, işe alımcılar ve araştırmacılar için bir sorun teşkil etmektedir.Bu veri seti, öğrencilerin becerilerini ve ilgi alanlarını analiz etmek için kullanılabilir, bu da onların UI/UX, geliştirme, veri bilimi vb. gibi farklı profesyonel profiller için uygunluğunu belirlemeye yardımcı olur.

Problemin amacı

Bu problemin amacı, bilgisayar bilimleri ve mühendislik öğrencilerinin teknik ve yumuşak becerilerini kapsayan bir veri setini kullanarak, öğrencilerin belirli profesyonel profillere ne kadar uygun olduğunu belirlemek için bir çerçeve oluşturmaktır. Bu çerçeve, öğrencilerin yeteneklerini ve ilgi alanlarını analiz ederek, onların UI/UX, geliştirme, veri bilimi vb. gibi farklı profesyonel profiller için uygunluğunu belirlemeye yardımcı olacaktır.

Projenin önemi

Bu tez, öğrencilerin teknik ve yumuşak becerilerini, hackathonlara katılımlarını ve belirli araçlar veya dillerdeki yeterliliklerini analiz ederek, onların belirli profesyonel profillere ne kadar uygun olduğunu belirlemek için bir çerçeve oluşturmayı amaçlamaktadır. Bu, eğitim kurumları, işe alımcılar ve araştırmacılar için büyük bir öneme sahiptir

Projenin Kapsamı

Teknoloji ve yazılım sektöründeki hızlı gelişmeler, öğrencilerin teknik ve yumuşak becerilerini sürekli olarak güncellemelerini gerektirir. Ancak, öğrencilerin bu becerileri nasıl ve ne sıklıkla geliştirdiklerini anlamak zordur. Bu durum, eğitim kurumları, işe alımcılar ve araştırmacılar için bir sorun oluşturur, çünkü öğrencilerin yeteneklerini ve ilgi alanlarını doğru bir şekilde değerlendiremezler. Bu nedenle, öğrencilerin teknik ve yumuşak becerilerini, hackathonlara katılımlarını ve belirli araçlar veya dillerdeki yeterliliklerini kapsayan ayrıntılı bir veri setinin analizi, öğrencilerin beceri eğilimlerini ve yerleştirme uygunluklarını daha doğru bir şekilde değerlendirmek için gereklidir.

2. Veri Toplama

Verilerin önemli bir kısmı öğrenciler tarafından doldurulan anketlerden elde edilmiştir. Bu anketler, çeşitli teknik ve sosyal beceriler, ilgi alanları ve hackathon'lara katılım gibi ders dışı etkinliklerde öz değerlendirme yeterliliği de dahil olmak üzere bir dizi bilgiyi toplamak için tasarlandı. Akademik Kayıtlar: Verilerin diğer kısmı öğrencinin akademik kayıtlarından elde edilmiştir. Bu, Veri Yapıları ve Algoritmalar (DSA), Veritabanı Yönetim Sistemleri (DBMS), İşletim Sistemleri (OS) ve Bilgisayar Ağları (CN) gibi kilit alanlardaki performanslarına ilişkin objektif ölçümler sağladı.Bu veri setindeki veriler, öğrencilerle yapılan anketler ve onların akademik kayıtlarının analizi yoluyla toplanmıştır. Anketler öğrencilerin kendi değerlendirdikleri beceri ve ilgi alanlarını tespit etmek için tasarlanmışken, akademik kayıtlar öğrencilerin çeşitli teknik konulardaki yeterliliklerine ilişkin nesnel ölçümler sağlıyordu.

Bu veriler kaggle sayfasından alınmıştır.[1]

3. Veri İşleme

3.1 Eksik Veri Doldurma

Veri setinde eksik değerler olmadığı için bu adımı geçtik.

3.2 Aykırı Değerlerin Ele Alınması

Aykırı değerler, IQR metodu kullanılarak tespit edilmiştir. Tespit edilen aykırı değerler, ilgili sütunun üst ve alt çeyrek değerleri arasına düşecek şekilde ayarlanmıştır.

3.3 Veri Normalleştirme

Veri setindeki tüm sayısal sütunlar, Standardizasyon işlemini StandardScaler ile ortalama değerin 0, standart sapmanın ise 1 değerini aldığı, dağılımın normale yaklaştığı bir metoddur. Formülü şu şekildedir, elimizdeki değerden ortalama değeri çıkartıyoruz, sonrasında varyans değerine bölüyoruz.
Normalleştirme olarak da Label Encoding kullanıyoruz.Bu kategorik niteliklerin her birine 0 ile n arasında ayrı bir rakam ataması yapar

Bu ön işleme adımları, veri setinin makine öğrenmesi modeli tarafından daha etkili bir şekilde kullanılabilmesini sağlamıştır.

4.Model Seçimi ve Model Eğitimi

4.1 Model Seçimi

Bu projede, çoklu sınıflandırma problemi için çeşitli makine öğrenmesi modellerini karşılaştırmak üzere bir dizi model seçtik.

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. XGBoost
5. LightGBM
6. CatBoost
7. Naive Bayes
8. KNN
9. GBM

4.2 Model Eğitimi

Model eğitimi, veri setinin %67'si eğitim verisi olarak ve %33'u test verisi olarak ayrıldıktan sonra gerçekleştirildi. Her model, eğitim verileri üzerinde eğitildi.

Model aşırı öğrenmeye yatkın olduğundan bir çok yol denendi ama yinede aşırı öğrenmeye bir tık engel olabildi. Veri çoğaltmaya çalışarak ve özellik sayısını azaltarak bu işlemler yapılmıştır.

Her modelin performansı, test verileri üzerinde değerlendirildi ve her modelin genel başarısı belirlendi. Bu süreç, her modelin gelecekteki veriler üzerinde ne kadar iyi performans göstereceğine dair bir tahmin sağlar.

5. Test ve Sonuçlar

Model Performansları

Her modelin performansı, test verileri üzerinde değerlendirildi. İşte her bir modelin performansı:

Tabii ki, aşağıdaki gibi değiştirebiliriz:

1. KNeighborsClassifier:
    - Doğruluk Oranı: 98.29%
    - Geri Çağırma Oranı: 98.29%
    - Hassasiyet Oranı: 98.32%
    - F1 Skoru: 98.29%
2. LogisticRegression:
    - Doğruluk Oranı: 78.63%
    - Geri Çağırma Oranı: 78.63%
    - Hassasiyet Oranı: 82.06%
    - F1 Skoru: 77.29%
3. SVC:
    - Doğruluk Oranı: 98.29%
    - Geri Çağırma Oranı: 98.29%
    - Hassasiyet Oranı: 98.32%
    - F1 Skoru: 98.29%
4. GaussianNB:
    - Doğruluk Oranı: 97.86%
    - Geri Çağırma Oranı: 97.86%
    - Hassasiyet Oranı: 97.93%
    - F1 Skoru: 97.86%
5. DecisionTreeClassifier:
    - Doğruluk Oranı: 99.14%
    - Geri Çağırma Oranı: 99.14%
    - Hassasiyet Oranı: 99.17%
    - F1 Skoru: 99.14%
6. RandomForestClassifier:
    - Doğruluk Oranı: 99.57%
    - Geri Çağırma Oranı: 99.57%
    - Hassasiyet Oranı: 99.58%
    - F1 Skoru: 99.57%
7. GradientBoostingClassifier:
    - Doğruluk Oranı: 99.14%
    - Geri Çağırma Oranı: 99.14%
    - Hassasiyet Oranı: 99.17%
    - F1 Skoru: 99.14%
8. CatBoostClassifier:
    - Doğruluk Oranı: 99.14%
    - Geri Çağırma Oranı: 99.14%
    - Hassasiyet Oranı: 99.17%

- F1 Skoru: 99.14%

9. LGBMClassifier:
    - Doğruluk Oranı: 99.14%
    - Geri Çağırma Oranı: 99.14%
    - Hassasiyet Oranı: 99.17%
    - F1 Skoru: 99.14%
10. XGBClassifier:
    - Doğruluk Oranı: 99.14%
    - Geri Çağırma Oranı: 99.14%
    - Hassasiyet Oranı: 99.17%
    - F1 Skoru: 99.14%

Sonuç olarak, RandomForest modeli, çoklu sınıflandırma problemi için en iyi model olarak belirlendi.

6. Gelişim

GridSearchCV kullanarak sonuçların daha iyi çıkması sağlanabilir ve gerekli hiperparametrelerle bu proje gerçek hayatta kullanılabilir.

KAYNAKLAR

1. https://www.kaggle.com/datasets/yuvjeetarora/student-job-profile


T.C.
KAHRAMANMARAŞ
SÜTÇÜ İMAM ÜNİVERSİTESİ
1992