NECMETTİN ERBAKAN ÜNİVERSİTESİ

MENTAL SAĞLIK VERİLERİNDE EKSİK VERİ YAPISI VE TAHMİN MODELLERİ

AMAÇ: BU ÇALIŞMADA, BİR RUH SAĞLIĞI VERİ SETİ ÜZERİNDE EKSİK VERİ ANALİZLERİ YAPILARAK, TEDAVİ GÖRME DURUMUNUN TAHMİNİNE YÖNELİK MAKİNE ÖĞRENMESİ MODELLERİ DEĞERLENDİRİLMİŞTİR.

NECMETTİN ERBAKAN ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ

ÖZET

Bu çalışma, bireylerin ruh sağlığı tedavisi görme durumlarını eksik veri analizleri ve makine öğrenmesi modelleriyle tahmin etmeyi amaçlamaktadır. Mental Health veri seti üzerinde yapılan analizlerde, eksik verilerin yapısı istatistiksel testlerle incelenmiş ve eksikliklerin çoğunlukla MNAR (Missing Not At Random) türünde olduğu belirlenmiştir. Kategorik veriler sayısallaştırılarak Random Forest, Logistic Regression, KNN ve Gradient Boosting modelleri eğitilmiş ve test edilmiştir. Modeller arasında en yüksek doğruluk oranını %99.77 ile Random Forest göstermiştir. Ayrıca cinsiyet, sosyal zayıflık, ailede ruh sağlığı geçmişi ve evde geçirilen gün sayısı gibi değişkenlerin tedavi durumuyla istatistiksel olarak anlamlı ilişkiler taşıdığı tespit edilmiştir. Elde edilen sonuçlar, ruh sağlığı alanında erken teşhis ve müdahale mekanizmalarının geliştirilmesine katkı sağlayabilecek niteliktedir.

Giriş

Mental sağlık alanında erken teşhis, bireylerin yaşam kalitesini artırmada kritik bir rol oynamaktadır. Bu çalışmada, bir mental sağlık veri setindeki eksik verilerin yapısı incelenmiş ve ruh sağlığı tedavisi tahmini için makine öğrenmesi yöntemleri uygulanmıştır.

YÖNTEMLER

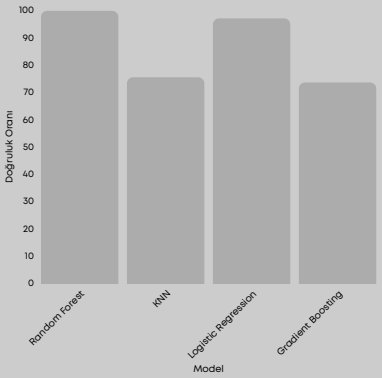
- Eksik Veri Analizi:
  1. Eksik veri yapısı ısı haritaları ile görselleştirilmiştir.
  2. Ki-Kare testi, t-testi ve lojistik regresyon yöntemleri kullanılarak eksik verilerin MAR, MCAR veya MNAR türünde olup olmadığı analiz edilmiştir.

- Veri Ön İşleme:
  1. Kategorik veriler LabelEncoder ve get_dummies() yöntemleriyle sayısallaştırılmıştır.
  2. Eksik veriler uygun stratejilerle doldurulmuş; bazı sütunlarda "Unknown" değeri atanmıştır.

- İstatistiksel Analiz:
  1. Cinsiyet, evde geçirilen gün sayısı, ailede ruh sağlığı geçmişi ve sosyal zayıflık gibi değişkenlerle tedavi görme durumu arasındaki ilişkiler chi-square testi ile değerlendirilmiştir.

- Makine Öğrenmesi Modelleri:
  1. Veri %80 eğitim ve %20 test olarak ayrılmıştır.
  2. Dört farklı model eğitilmiştir:
     - Random Forest
     - Logistic Regression
     - K-Nearest Neighbors (KNN)
     - Gradient Boosting
  3. Modellerin başarımı doğruluk (accuracy) skoru ve karışıklık matrisi (confusion matrix) ile değerlendirilmiştir.

MODEL KARŞILAŞTIRMASI GRAFİĞİ



Uygulanan dört farklı makine öğrenmesi modelinin doğruluk oranları karşılaştırılmıştır. En yüksek başarı oranı %99.77 ile Random Forest modeline ait olup, veri setindeki karmaşık ilişkileri en iyi şekilde öğrenebildiği gözlemlenmiştir. Logistic Regression ve KNN modelleri de yüksek doğruluk sunarken, Gradient Boosting görece daha düşük ancak dengeli bir performans göstermiştir. Bu karşılaştırma, sınıflandırma problemlerinde model seçiminde doğruluk oranlarının önemini vurgulamaktadır.

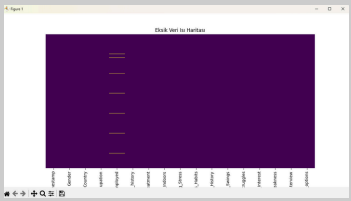
SONUÇLAR

- Eksik veri analizi sonucunda, eksik verilerin eksik olan değişkenin kendisiyle doğrudan ilişkili olduğu ve bu nedenle MNAR (Missing Not At Random) olabileceği belirlenmiştir.
- Cinsiyet ve ruh sağlığı tedavi durumu arasında güçlü bir ilişki bulunmuştur ($p < 0.001$).
- Evde geçirilen gün sayısı ile ailede ruh sağlığı geçmişi arasında istatistiksel olarak anlamlı bir ilişki vardır ($p \approx 0.025$).
- Stres seviyesi artışı ve çalışmaya olan ilgi arasında güçlü bir ilişki saptanmıştır ($p < 0.001$).
- Ailede ruh sağlığı geçmişi olan bireylerde tedavi alma oranı anlamlı şekilde yüksektir ($p < 0.001$).
- Çalışma durumu ve sosyal zayıflık değişkenlerinin stres veya tedavi ile ilişkisi istatistiksel olarak anlamlı bulunmamıştır ($p > 0.05$).
- Random Forest modeli %99.77 doğruluk oranı ile en iyi performansı göstermiştir.

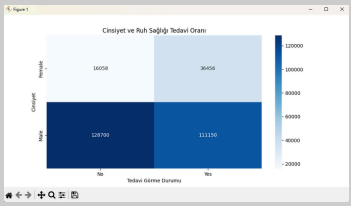
KAYNAKÇA / KULLANILAN KÜTÜPHANELER

- Veri Seti: Mental Health Dataset (Kaggle)
- Kütüphaneler: pandas, seaborn, matplotlib, scikit-learn, missingno, scipy

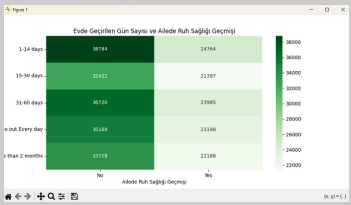
BULGULAR VE TARTIŞMA



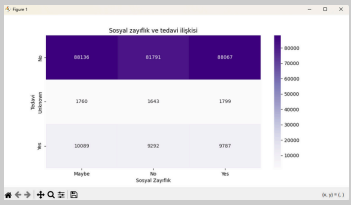
EKSİK VERİ ISI HARİTASI
- VERİSETİ İÇİNDEKİ EKSİK DEĞERLERİN GÖRSEL DAĞILIMINI GÖSTERİR.
1. SARI ÇİZGİLER EKSİK VERİ OLAN SATIRLARI TEMSİL EDER.
2. EKSİK VERİLER, ÖZELLİKLE SELF_EMPLOYED, WORK_INTERFERE, FAMİLY_HİSTORY GİBİ BAZI KATEGORİK SÜTUNLARDA YOĞUNLAŞMIŞTIR.
3. BU GÖRSELLEŞTİRME SAYESİNDE HANGİ DEĞİŞKENLERDE EKSİKLİK BULUNDUĞU KOLAYLIKLA TESPİT EDİLEBİLİR.



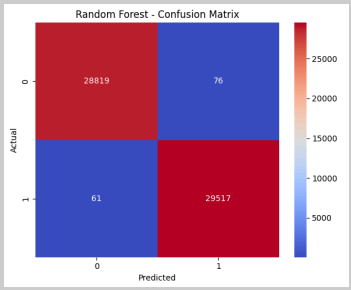
CİNSİYET VE RUH SAĞLIĞI TEDAVİ ORANI
- KADINLARIN RUH SAĞLIĞI TEDAVİSİ ALMA ORANI ERKEKLERE KIYASLA DAHA YÜKSEKTİR.
1. ERKEKLER (MALE): TEDAVİ ALMAYAN 128,700 → ALAN 111,150
2. KADINLAR (FEMALE): TEDAVİ ALMAYAN 16,058 → ALAN 36,456
3. BU DURUM, ERKEKLERİN YARDIM ARAMADA DAHA ÇEKİMSER OLDUĞUNU VEYA FARKINDALIĞIN KADINLARDA DAHA YÜKSEK OLDUĞUNU GÖSTEREBİLİR.



EVDE GEÇİRİLEN GÜN SAYISI VE AİLEDE RUH SAĞLIĞI GEÇMİŞİ
- EVDE UZUN SÜRE KALAN BİREYLERDE RUH SAĞLIĞI GEÇMİŞİ OLANLARIN SAYISI AZALMA EĞİLİMİNDEDİR.
1. ÖRNEĞİN, "1-14 DAYS" GRUBUNDA AİLE GEÇMİŞİ OLMAYANLAR 38,784, OLANLAR 24,764 KİŞİDİR
2. BU İLİŞKİ, SOSYAL ETKİLEŞİMİN VE DIŞ DÜNYA İLE BAĞLANTININ ÖNEMİNE İŞARET EDEBİLİR.



SOSYAL ZAYIFLIK VE TEDAVİ İLİŞKİSİ
- SOSYAL ZAYIFLIK, TEDAVİ ALIP ALMAMA ÜZERİNDE GÜÇLÜ BİR AYRIM YARATMIYOR.
1. TEDAVİ ALMAYAN KİŞİ SAYISI, SOSYAL ZAYIFLIK DURUMU NE OLURSA OLSUN NEREDEYSE AYNI. YANİ SOSYAL OLARAK ZAYIF OLAN DA, OLMAYAN DA, BELKİ OLAN DA BENZER ORANLARDA TEDAVİ ALMAMIŞ.
2. TEDAVİ ALANLARDA DA BENZER BİR DAĞILIM VAR. YANİ SOSYAL ZAYIFLIK DÜZEYİ DEĞİŞSE BİLE, TEDAVİ ALMA ORANI ÇOK FAZLA FARKLILAŞMIYOR.



RANDOM FOREST
- RANDOM FOREST MODELİNİN PERFORMANSINI GÖSTEREN KARIŞIKLIK MATRİSİ.
1. TRUE NEGATİVE (0,0): 28,819 — MODEL "HAYIR" TAHMİN ETTİ VE DOĞRUYDU.
2. FALSE POSİTİVE (0,1): 76 — MODEL "EVET" DEDİ AMA GERÇEK "HAYIR"DI.
3. FALSE NEGATİVE (1,0): 61 — MODEL "HAYIR" DEDİ AMA GERÇEK "EVET"Tİ.
4. TRUE POSİTİVE (1,1): 29,517 — MODEL "EVET" TAHMİN ETTİ VE DOĞRUYDU.
5. BU SONUÇ, MODELİN HEM DUYARLILIK HEM ÖZGÜLLÜK AÇISINDAN OLDUKÇA BAŞARILI OLDUĞUNU GÖSTERMEKTEDİR.

Elde edilen bulgular, ruh sağlığı tedavisi görme durumunun cinsiyet, sosyal zayıflık, ailede ruh sağlığı geçmişi gibi değişkenlerle anlamlı ilişkiler taşıdığını göstermektedir. Özellikle kadın bireylerin tedaviye daha fazla yönelmesi, yardım arama davranışlarında toplumsal cinsiyet rollerinin etkili olabileceğine işaret etmektedir. Sosyal zayıflığı olan bireylerde tedavi oranının yüksek olması ise sosyal destek eksikliğinin psikolojik yardım ihtiyacını artırdığını düşündürmektedir.
Modelleme sonuçlarına göre Random Forest algoritması oldukça yüksek doğruluk oranıyla öne çıkmıştır. Bu durum, kompleks ve çok boyutlu kategorik verilerle çalışırken bu modelin tercih edilmesi gerektiğini göstermektedir. Ancak, eksik verilerin MNAR türünde olması, bazı değişkenlerin sistematik şekilde eksik kalabileceğini ve model başarısını bu yönde etkileyebileceğini ortaya koymaktadır. Bu nedenle, eksik veri yönetimi aşamasında dikkatli ve veri yapısına özel stratejiler geliştirilmesi önemlidir.

HAZIRLAYAN: Ali Rıza ŞAHİN - 22100011066
İLETİŞİM: sahinaliriza888@gmail.com