BURSA TEKNİK ÜNİVERSİTESİ

BİLGİSAYAR MÜHENDİSLİĞİ

VERİ MADENCİLİĞİNE GİRİŞ PROJE RAPORU

LÜTFÜ BEDEL

21360859030

Özet :

Bu proje, çevresel hava kalitesi göstergeleri ile bireylerin duygusal durumları arasındaki ilişkiyi analiz ederek, gerçek zamanlı duygu sınıflandırması yapabilen model geliştirmeyi amaçlamaktadır. Çalışmada, gerçek dünyadan elde edilmiş sensör verileri (PM, NO_2, CO, NH_3, gürültü seviyesi vb.) ve bireylerin öz bildirim yoluyla ifade ettikleri duygusal durumlar kullanılmıştır. Bu veriler, düşük maliyetli taşınabilir cihazlar aracılığıyla toplanmış ve Random Forest algoritması kullanılarak sınıflandırılmıştır. Geliştirilen model, %97 doğruluk oranı ve yüksek AUC değerleri ile çevresel verilerden duygu durumu tahmini yapılabileceğini göstermiştir. Elde edilen sonuçlar, hava kalitesi gibi çevresel faktörlerin bireylerin psikolojik durumlarını etkileyebileceğini ortaya koymakta ve uç cihazlar üzerinde çalışan yapay zekâ uygulamaları için umut vadeden bir çözüm sunmaktadır.

Giriş :

Gelişen çevresel sensör teknolojileri ve yapay zekâ algoritmaları, bireylerin duygusal durumlarının gerçek zamanlı olarak izlenmesini ve sınıflandırılmasını mümkün kılmaktadır. Özellikle hava kalitesi gibi çevresel faktörlerin insan psikolojisi üzerindeki etkileri, son yıllarda daha fazla dikkat çekmeye başlamış ve duygudurum tahmini gibi uygulamalarda alternatif veri kaynakları olarak değerlendirilmeye başlanmıştır.

Bu çalışma, hava kirliliği göstergeleri (PM, NO_2, CO, NH_3, gürültü seviyesi vb.) ile bireylerin duygusal durumları arasındaki ilişkiyi modellemek, ve bu amaçla Random Forest algoritması kullanarak gerçek zamanlı sınıflandırma yapabilen bir sistem geliştirmek üzerine odaklanmaktadır. Klasik duygu durum tespiti yöntemleri genellikle yüksek donanımsal kaynaklara ihtiyaç duyarken, bu projede düşük maliyetli ve taşınabilir bir donanım üzerinde çalışan bir çözüm önerilmektedir. Böylece, uç (edge) cihazlar üzerinde çalışan yapay zekâ modelleri ile hem veri gizliliği sağlanmakta hem de çevrimdışı ortamlarda çalışabilen uygulamalar mümkün hale gelmektedir.

Projede kullanılan veriler, gerçek dünya koşullarında elde edilmiş hava kalitesi sensör ölçümleri ile eş zamanlı olarak alınan bireysel duygu etiketlerinden oluşmaktadır. Bu veri seti, çevresel değişkenlerin duygusal durumlar üzerindeki etkisini incelemek açısından oldukça değerli bir kaynak sunmaktadır. Random Forest algoritması, birden fazla karar ağacının bir arada çalıştığı bir topluluk öğrenme yöntemi olup; özellikle çok boyutlu ve düzensiz verilerde yüksek başarı gösterdiğinden dolayı tercih edilmiştir.

Veri Seti Tanıtımı :

Bu çalışmada kullanılan veri seti, DigitalExposome adlı önceki bir araştırmadan elde edilen gerçek dünya verilerine dayanmaktadır. Söz konusu çalışma, bireylerin çevresel hava kalitesi değişkenlerine maruz kalma düzeyleri ile kendi beyan ettikleri duygusal durumlar arasındaki ilişkiyi incelemeye yönelik olarak tasarlanmıştır.

Veri toplama sürecinde her katılımcıya iki farklı cihaz verilmiştir:

- **Enviro-IoT:** Bu cihaz, çevresel hava kalitesini sürekli izleyen taşınabilir bir sensör sistemidir. Cihaz; Partikül Madde (PM1, PM2.5, PM10), Azot Dioksit (NO_2), Karbon Monoksit (CO), Amonyak (NH_3) ve Gürültü (dB) gibi temel hava kirliliği parametrelerini yüksek frekansla ölçerek kaydetmektedir.
- **EnvBodySens Uygulaması:** Samsung marka akıllı telefonlara yüklenen bu mobil uygulama aracılığıyla katılımcılar, belirli zaman aralıklarında duygusal durumlarını 1 (çok mutsuz) ile 5 (çok mutlu) aralığında ifade eden emojiler aracılığıyla öz bildirimde bulunmuşlardır.

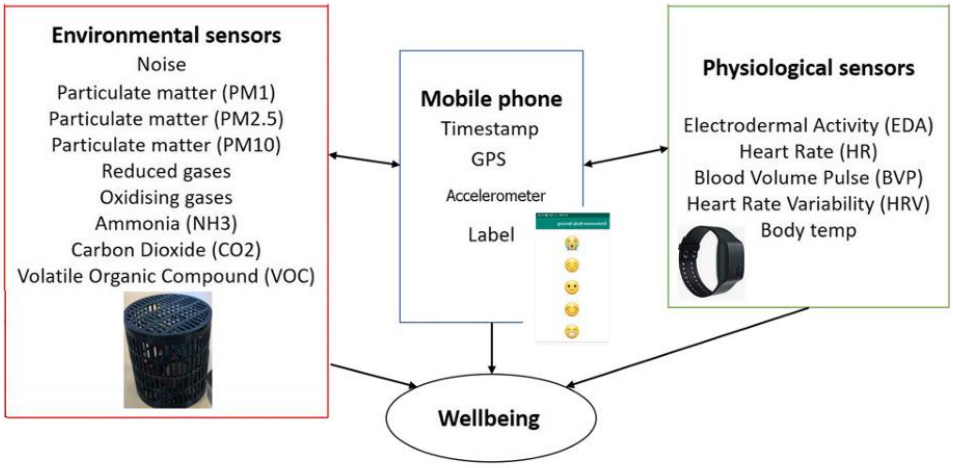


Fig. 4 List of the fused variables collected by each device

Toplanan veriler, her bir katılımcının bulunduğu çevredeki hava kalitesi ölçümleri ile aynı anda bildirilen duygusal durumların eşleştirilmesiyle oluşturulmuştur. Böylece her gözlem birimi, ilgili çevresel ölçümler ile birlikte, katılımcının o anda hissettiği duygu durumunu da içeren etiketli bir kayıt haline getirilmiştir.

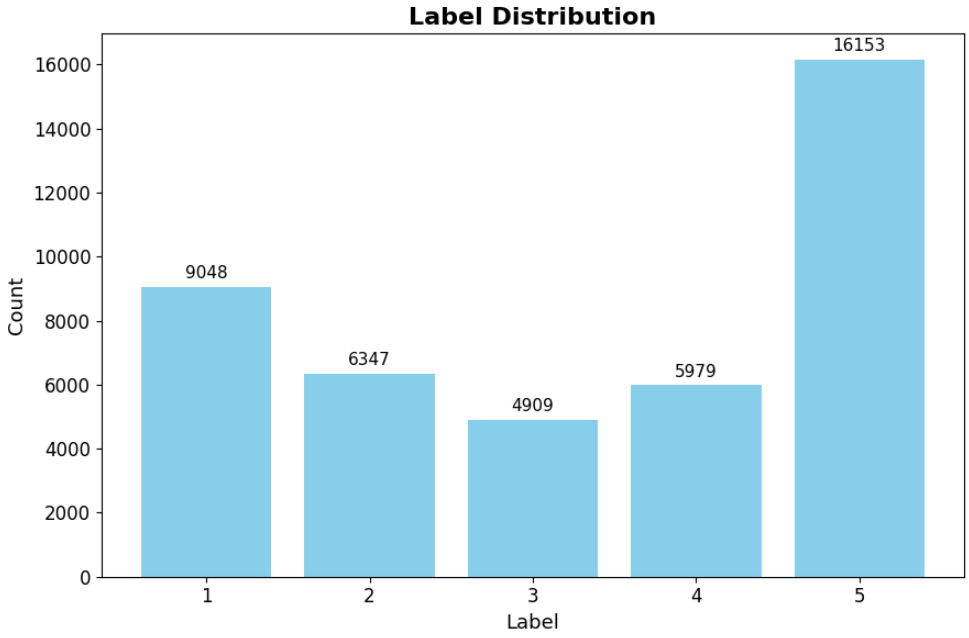
Çevresel (Environmental) Veriler:	
NO2 (Nitrogen Dioxide)	Azot dioksit gazı konsantrasyonu (hava kirliliği göstergesi)
NOISE	Ortamdaki ses seviyesi (desibel cinsinden, dB)
NH3 (Ammonia)	Amonyak gazı seviyesi (ppm - parts per million)
PM10	10 mikrometreye kadar partikül madde
CO (Carbon Monoxide)	Karbon monoksit gazı konsantrasyonu (ppm)
PM25 (PM2.5)	2.5 mikrometreye kadar ince partikül madde
PM1	1 mikrometreye kadar partikül madde

Fizyolojik (Physiological) Veriler :	
EDA (Electrodermal Activity)	Deri iletkenliği, stres seviyesiyle ilişkilidir
BVP (Blood Volume Pulse)	Kan hacmi dalgalanması
HR (Heart Rate)	Kalp atış hızı (bpm - beats per minute)
IBI (Inter-Beat Interval)	Kalp atımları arasındaki süre (ms cinsinden)

Sonuç olarak elde edilen veri seti; 12 sütun (özellik) ve 42.437 örnek (satır) içermektedir. Bu veri yapısı, makine öğrenmesi algoritmaları ile duygusal durumların sınıflandırılması amacıyla kullanılmak üzere oldukça uygun ve zengin bir içerik sunmaktadır.

IBI	HR	NO2	Noise	NH3	PM10	CO	PM25	Label	PM1	EDA	BVP
0.0	0.377574	0.0	0.511358	0.003018	0.003091	0.871758	0.000000	5	0.000000	0.0	0.0
0.0	0.196398	0.0	0.490903	0.003018	0.003091	0.876848	0.003091	5	0.001854	0.0	0.0
0.0	0.454163	0.0	0.470449	0.006036	0.006181	0.881939	0.006181	5	0.003709	0.0	0.0
0.0	0.322451	0.0	0.449995	0.009055	0.009272	0.887030	0.009272	5	0.005563	0.0	0.0
0.0	0.237595	0.0	0.429540	0.012073	0.012362	0.892121	0.012362	5	0.007417	0.0	0.0

(Veriseti ilk 5 satır)



(Veri Setinin Label Değerlerine Göre Dağılımı)

Modelleme: Random Forest Classifier :

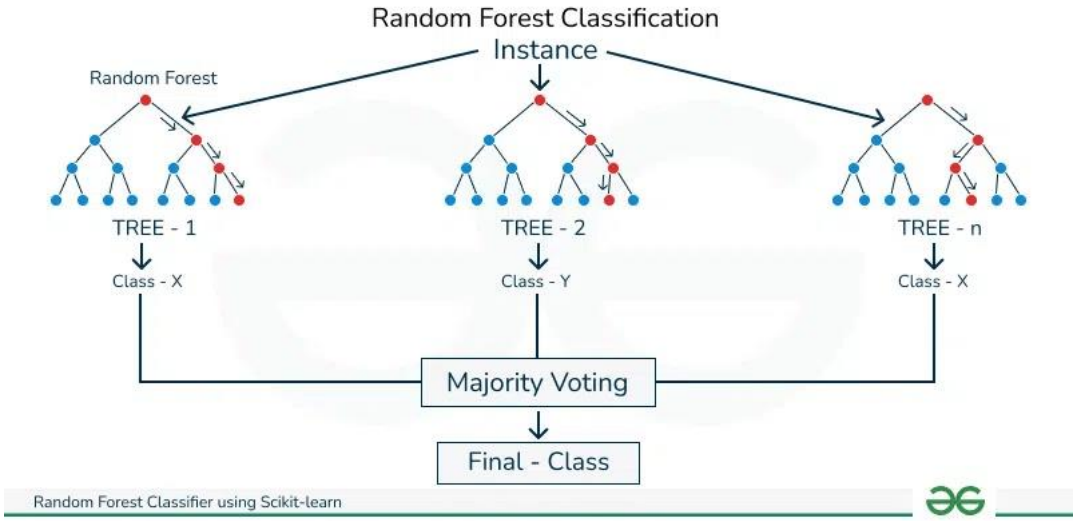
Bu çalışmada sınıflandırma problemini çözmek amacıyla Random Forest algoritması tercih edilmiştir. Bu algoritma; yüksek doğruluk oranı, aşırı öğrenmeye (overfitting) karşı dayanıklılığı ve karmaşık, çok boyutlu veri yapıları üzerinde gösterdiği başarılı performans nedeniyle veri madenciliği uygulamalarında yaygın olarak kullanılmaktadır.

Random Forest, temel olarak birden fazla karar ağacının bir araya gelerek oluşturduğu bir topluluk (ensemble) öğrenme yöntemidir. Sınıflandırma görevlerinde, her bir ağaç bağımsız olarak bir tahmin üretir ve nihai çıktı, tüm ağaçların oy çokluğuna dayalı olarak belirlenir. Bu yaklaşım, modelin genelleme yeteneğini artırmakta ve tek bir karar ağacına kıyasla daha dengeli sonuçlar elde edilmesini sağlamaktadır.

Random Forest Classifier Avantajları :

- Büyük hacimli ve yüksek boyutlu veri kümeleri ile etkin şekilde çalışabilir.
- Birden fazla karar ağacından elde edilen tahminlerin birleştirilmesi sayesinde, tek bir ağaç modeline kıyasla aşırı uyum (overfitting) riski önemli ölçüde azalır.
- Gürültülü verilere karşı dayanıklıdır ve hem sayısal hem de kategorik değişkenler ile etkili şekilde çalışabilir.

Bu nedenlerle Random Forest, çevresel sensör verileri gibi düzensiz ve gürültülü olabilen veri kümeleri üzerinde duygusal durum sınıflandırması için uygun bir yöntem olarak değerlendirilmiştir.



Performans Değerlendirme :

Bu projede geliştirilen Random Forest modeli, çevresel sensör verileri ve fizyolojik parametreler yardımıyla bireylerin duygusal durumlarını sınıflandırmada yüksek başarı göstermiştir. Modelin başarımını değerlendirmek için çeşitli sınıflandırma metrikleri uygulanmıştır: Precision (Kesinlik), Recall (Duyarlılık), F1-Score, Confusion Matrix ve ROC Eğrileri üzerinden AUC (Area Under Curve) analizleri gerçekleştirilmiştir.

Precision, Recall ve F1-Score Analizi:

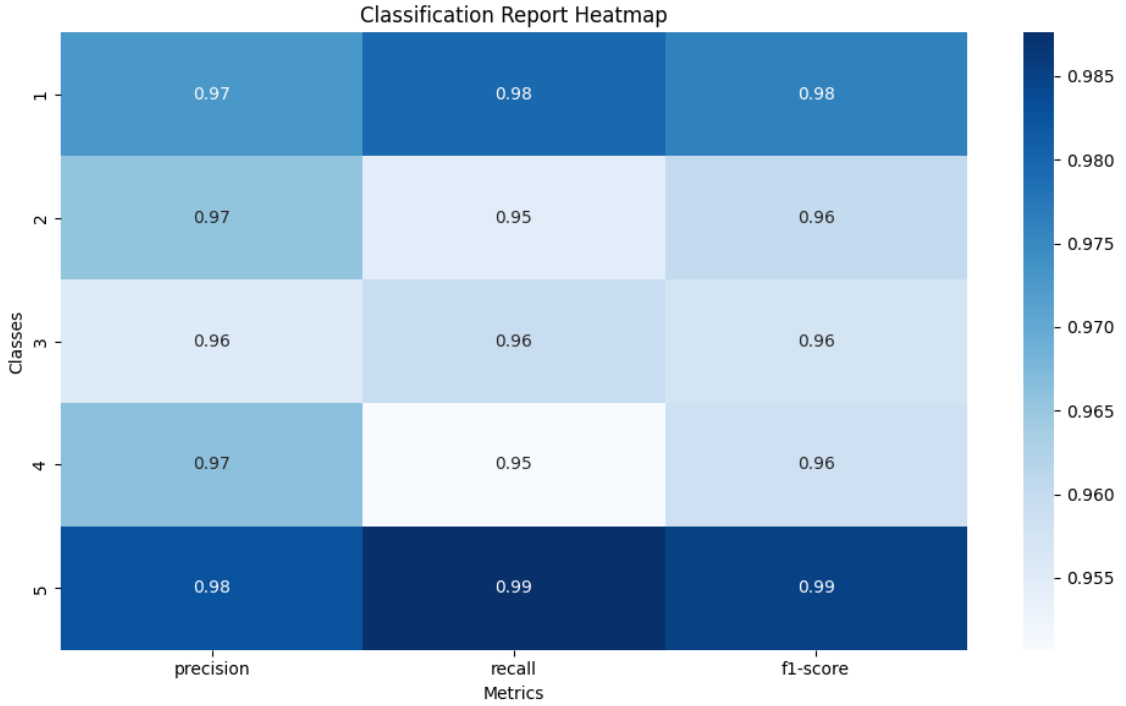
Modelin Label sütununda yer alan duygusal durumları (1 – Çok mutsuz ile 5 – Çok mutlu arası) doğru şekilde tahmin edip edemediği, sınıf bazlı şu metriklerle değerlendirilmiştir:

- **Precision**: Tahmin edilen pozitif sınıfların ne kadarının gerçekten doğru olduğunu gösterir. Düşük precision, modelin çok sayıda yanlış pozitif tahmin yaptığını gösterir.

- **Recall**: Gerçek pozitif sınıfların ne kadarının doğru tahmin edildiğini gösterir. Düşük recall, modelin bazı pozitif örnekleri gözden kaçırdığını gösterir.
- **F1-Score**: Precision ve Recall'un harmonik ortalamasıdır. Özellikle dengesiz veri setlerinde daha güvenilir bir ölçüttür.

Modelin tüm sınıflar için precision, recall ve f1-score değerleri oldukça yüksektir (%95-99 arası). Bu durum, veri setinin dengeli olduğunu ve modelin duygusal durumları güvenilir şekilde ayırt edebildiğini göstermektedir.

Modelin genel başarı oranını ifade eden accuracy (doğruluk) metriği de %97 olarak hesaplanmıştır. Bu değer, modelin tüm örnekler arasında %97 oranında doğru tahmin yaptığını gösterir.

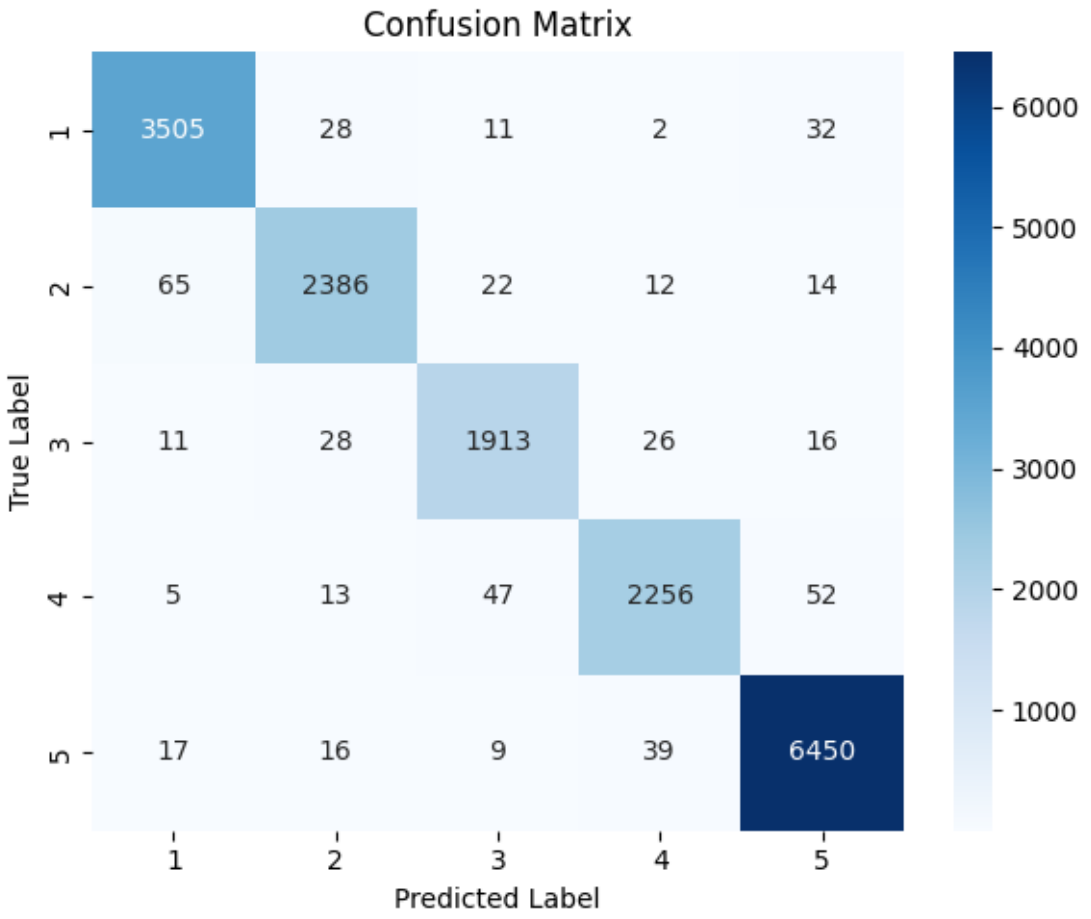


Confusion Matrix (Karmaşıklık Matrisi):

Confusion Matrix yardımıyla, modelin hangi sınıflarda doğru tahmin yaptığı, hangi sınıflarda karışıklık yaşadığı analiz edilmiştir.

Görüldüğü üzere, model özellikle 5. sınıf (çok mutlu) için oldukça yüksek bir doğruluk sergilemiştir. Diğer sınıflarda da hata oranı son derece düşüktür. Modelin, birbirine yakın sınıflar (örneğin 3 ve 4) arasında zaman zaman karışıklık yaşadığı gözlemlenmekle birlikte, genel anlamda sınıflar arası ayrım performansı oldukça yüksektir.

Bu durum, modelin duygusal durumları çevresel verilere dayanarak güvenilir bir şekilde tahmin edebildiğini ve sınıflar arasında net ayrımlar yapabildiğini göstermektedir.

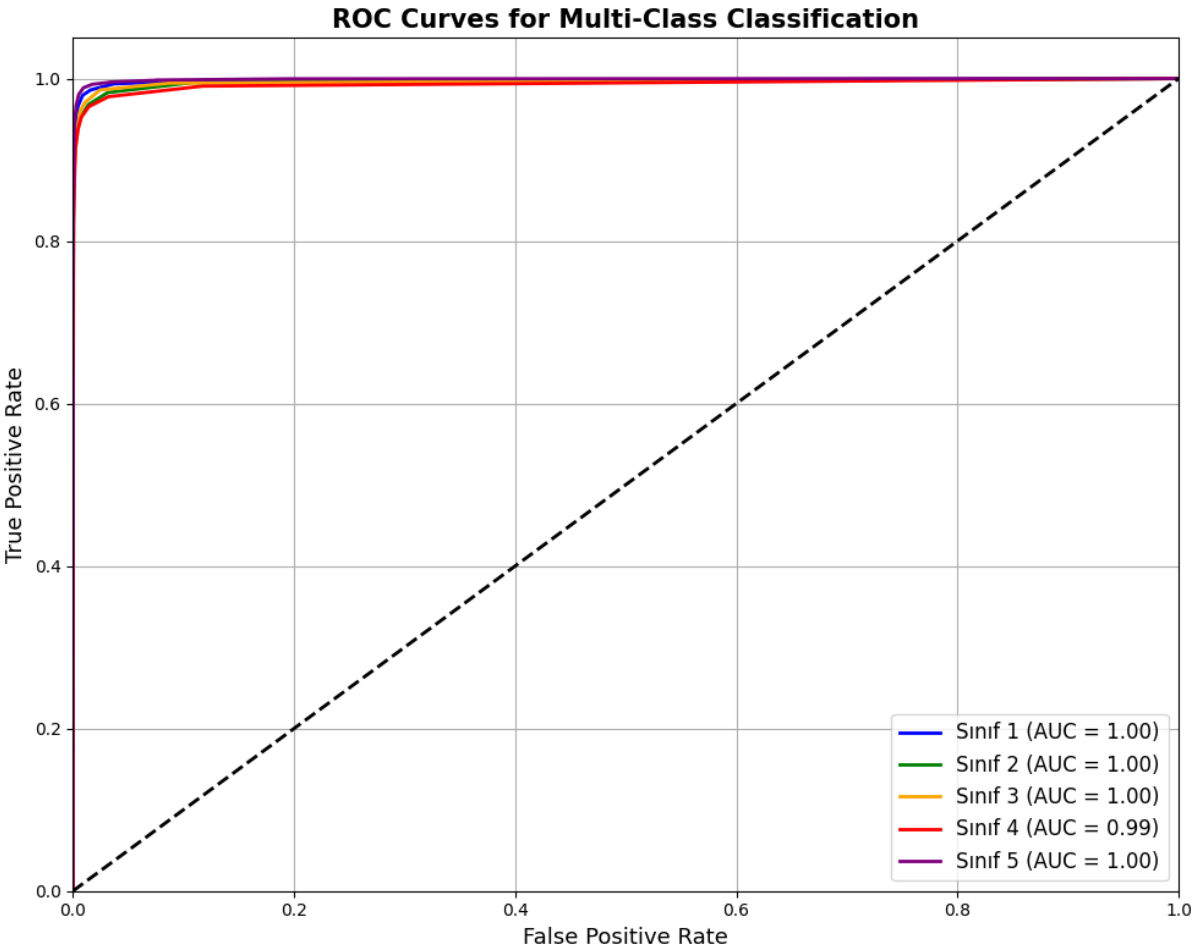


ROC Eğrileri ve AUC Değerleri:

Her bir sınıf için ROC eğrileri çizilmiş ve modelin sınıfları ayırt etme performansı değerlendirilmiştir. AUC (Area Under Curve) değeri, 1'e ne kadar yakınsa modelin ayrım gücü o kadar yüksek demektir.

- Sınıf 1, Sınıf 2, Sınıf 3, ve Sınıf 5 için AUC değeri 1.00 olarak elde edilmiştir. Bu, modelin bu sınıflar için neredeyse kusursuz bir ayrım gücüne sahip olduğunu göstermektedir.
- Sınıf 4 için AUC değeri 0.99 olup, yine oldukça yüksek bir başarıyı temsil etmektedir.

Sonuç olarak, ROC analizleri modelin genel sınıflandırma başarısının çok yüksek olduğunu ve tüm sınıflar için güvenilir tahminler üretebildiğini göstermektedir.



İlgili Çalışmalar :

Bu çalışmada kullanılan **DigitalExposome veri seti**, daha önce farklı projelerde de kullanılmış ve çevresel hava kalitesi ile bireylerin duygusal durumları arasındaki ilişki çeşitli yaklaşımlarla incelenmiştir.

Bu alandaki en dikkat çekici çalışmalardan biri olan **"Emotion on the Edge: Air Quality Sensors Decoded as a Real-World Emotion Indicator"** başlıklı makalede, çevresel hava kirliliği verileri kullanılarak gerçek zamanlı duygu sınıflandırması yapılmış ve model, taşınabilir bir cihaz üzerinde çalışacak şekilde tasarlanmıştır. Çalışmada kullanılan sistem, düşük maliyetli bir Raspberry Pi cihazına entegre edilmiş ve yalnızca hava kalitesi sensör verilerine dayanarak duygusal durumların sınıflandırılması hedeflenmiştir.

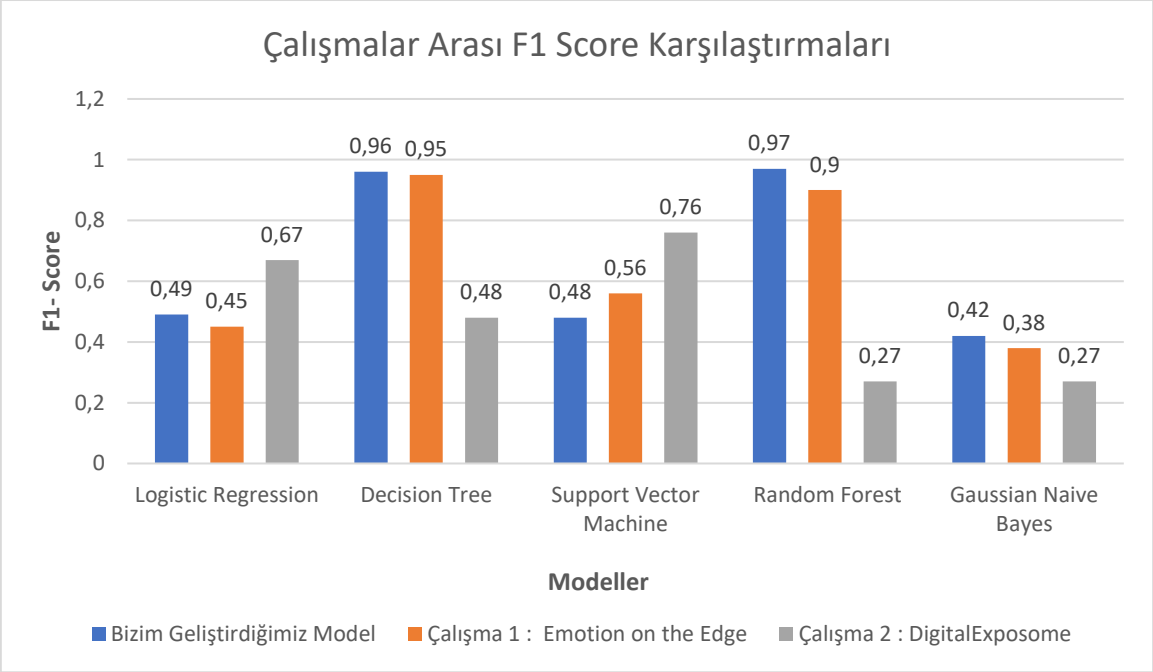
Makalenin temel katkıları şunlardır:

- Çeşitli makine öğrenmesi modelleri test edilmiş; bunlar arasında en başarılı sonuç **Decision Tree** modeli ile elde edilmiştir (**%95 doğruluk** ve **0.95 F1-score**).
- **Random Forest** modeli de değerlendirilmiş ve bu modelle **0.90 F1-score** elde edilmiştir. Bu sonuç, bizim çalışmamızdaki Random Forest yaklaşımı ile elde edilen başarımların benzer düzeyde olduğunu göstermektedir.

Bir diğer önemli çalışma, Thomas Johnson, Eiman Kanjo ve Kieran Woodward tarafından 2023 yılında gerçekleştirilen "**DigitalExposome: quantifying impact of urban environment on wellbeing using sensor fusion and deep learning**" başlıklı makaledir. Bu çalışmada, hava kirliliği (PM1, PM2.5, PM10, NH_3, NO_2 vb.) ile fizyolojik veriler (EDA, HR, HRV) birleştirilerek bireylerin anlık ruh hali sınıflandırılmıştır.

Veriler taşınabilir cihazlarla gerçek zamanlı toplanmış, makine öğrenmesi ve derin öğrenme yöntemleriyle analiz edilmiştir. En yüksek başarı, CNN ile çıkarılan özelliklerle eğitilen Random Forest modeliyle elde edilmiştir (%76 F1-score). Yalnızca çevresel verilerle %67, yalnızca fizyolojik verilerle ise %61 F1-score elde edilmiştir.

Bu sonuçlar, çevresel verilerin bireylerin duygusal durumları üzerinde önemli bir etkisi olduğunu ve tek başına da yüksek doğrulukla tahmin yapılabildiğini göstermektedir.



Sonuçlar ve Tartışma :

Bu çalışma kapsamında, çevresel sensör verileri ile bireylerin duygusal durumları arasındaki ilişkiyi modellemek amacıyla Random Forest algoritması tabanlı bir sınıflandırma sistemi geliştirilmiştir. Yapılan analizler sonucunda, geliştirilen modelin oldukça yüksek doğruluk oranları ile çalıştığı ve duygusal durumları doğru bir şekilde tahmin edebildiği gözlemlenmiştir.

Modelin sınıflandırma performansı %97 gibi yüksek bir doğruluk oranı ile değerlendirilmiş; precision, recall ve F1-score gibi metriklerde ise tüm sınıflar için %95'in üzerinde başarı sağlanmıştır. Ayrıca ROC eğrileri ve AUC değerleri de modelin sınıflar arasında güçlü bir ayrım yapabildiğini kanıtlamıştır. Bu sonuçlar, çevresel değişkenlerin bireylerin duygusal durumları üzerinde doğrudan veya dolaylı etkileri olduğunu göstermekte ve bu tür verilerin duygudurum tahmini gibi uygulamalarda kullanılabileceğini desteklemektedir.

Projede kullanılan veri setinin gerçek dünya koşullarında elde edilmesi, modelin pratikte uygulanabilirliğini ve genellenebilirliğini artıran önemli bir avantaj olmuştur. Ayrıca, sistemin düşük maliyetli, taşınabilir ve uç cihazlar üzerinde çalışabilir bir yapıda tasarlanması; veri gizliliği, enerji verimliliği ve çevrimdışı kullanım gibi önemli gereksinimlere yanıt vermektedir. Böylelikle, geleneksel

yüksek donanım gerektiren sistemlerin aksine, daha erişilebilir ve ölçeklenebilir bir çözüm sunulmuştur.

Sonuç olarak, bu proje, çevresel faktörlerin bireylerin duygusal durumları üzerindeki etkilerini anlamaya yönelik önemli bir adım niteliğindedir. Random Forest algoritması ile desteklenen bu yaklaşım, duygudurum tahmini alanında hem akademik hem de endüstriyel uygulamalara katkı sağlayabilecek nitelikte güçlü bir temel sunmaktadır.

İlgili Linkler :

Github : https://github.com/lutfubedel/Veri_Madenciligi_Donem_Projesi

Youtube : https://youtu.be/q-vq2YO5kOk

Kaynakça :

DigitalExposome: A dataset for wellbeing classification using environmental air quality and human physiological data : https://www.sciencedirect.com/science/article/pii/S235234092500174X

DigitalExposome: quantifying impact of urban environment on wellbeing using sensor fusion and deep learning : https://link.springer.com/article/10.1007/s43762-023-00088-9#Sec2

Sensor Fusion and The City: Visualisation and Aggregation of Environmental & Wellbeing Data : https://ieeexplore.ieee.org/document/9562852

Emotion on the Edge: Air Quality Sensors Decoded as a Real-World Emotion Indicator : https://ieeexplore.ieee.org/document/10502563

Scikit-learn RandomForestClassifier : https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

IBM What is random forest? : https://www.ibm.com/think/topics/random-forest

GeeksforGeeks Random Forest Classifier using Scikit-learn : https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/