**Dermatolojik Hastalıkların Teşhisinde Ensemble Öğrenme Yaklaşımları**

Ayşenur Eskin, Büşra Mermer, Zeynep Özcan, Muhammet Demir, Eslem Oluk

Anahtar kelimeleri (dermatoloji, sınıflandırma, ensemble yöntemleri) öne çıkarmak

**Özet**

Bu çalışmada, dermatolojik hastalıkların sınıflandırılmasında farklı ensemble yöntemlerinin performansı incelenmiştir. Kaggle üzerinden temin edilen "Dermatology Dataset Classification" veri seti kullanılarak, Random Forest, AdaBoost ve Stacking gibi makine öğrenmesi algoritmaları uygulanmış ve modellerin başarısı doğruluk (accuracy), hassasiyet (precision), duyarlılık (recall) ve F1 skoru gibi metriklerle değerlendirilmiştir.

Veri setinin normal dağılıma uymadığı Kolmogorov-Smirnov testi ile belirlenmiş ve parametrik olmayan yöntemler tercih edilmiştir. Özellik seçimi aşamasında Kruskal-Wallis H Testi kullanılarak sınıflandırma için en önemli özellikler belirlenmiştir. Yapılan analizler sonucunda, Random Forest modelinin %97.2 doğruluk oranı ve 0.97 F1 skoru ile en yüksek performansı gösterdiği tespit edilmiştir. AdaBoost modeli ise diğer yöntemlere kıyasla daha düşük başarı sergilerken, Stacking yöntemi yüksek doğruluk sağlamasına rağmen daha uzun bir eğitim süresi gerektirmiştir.

Bu çalışma, dermatolojik hastalıkların teşhisinde ensemble yöntemlerinin etkinliğini ortaya koymuş ve özellikle Random Forest gibi modellerin bu alandaki potansiyelini vurgulamıştır. Gelecekte, daha geniş ve çeşitli veri setleri kullanılarak modelin genellenebilirliğinin artırılması, derin öğrenme tabanlı yaklaşımların denenmesi ve klinik uygulamalarda entegrasyonun araştırılması önerilmektedir. Bu tür çalışmalar, dermatolojik tanı süreçlerinin doğruluğunu ve hızını artırarak tıbbi karar destek sistemlerine katkı sağlayabilir.

**1. Giriş**

Dermatoloji (Cildiye), cilt, saç, tırnak ve mukozaları etkileyen hastalıkların tanı, tedavi ve önlenmesiyle ilgilenen tıp dalıdır [1]. Dermatologlar, akne, egzama, sedef hastalığı, saç dökülmesi, cilt kanseri gibi çok çeşitli cilt sorunlarıyla ilgilenirler [1]. Aynı zamanda dermatoloji, kozmetik dermatoloji alanında da hizmetler sunarak cilt sağlığını ve estetik görünümünü iyileştirmeyi amaçlar [1]. Dermatolojinin önemi, cildin vücudun en büyük organı olarak çevresel tehditlere karşı koruyucu bir bariyer işlevi görmesi, sıcaklık düzenlemesi yapması ve duyusal algıyı kolaylaştırmasıyla vurgulanmaktadır [2]. Dermatoloji, cilt sağlığını korumada, cilt hastalıklarını teşhis etmede ve kişinin yaşam kalitesini ciddi şekilde etkileyebilecek hastalıkları tedavi etmede önemli bir rol oynamaktadır [3].

Birçok kişi kendilerinde bulunan deri değişikliklerini hastalık olarak değerlendirmezler fakat cilt rahatsızlıkları oldukça önemlidir [3]. Cilt hastalıkları, dermatolojik sorunlar arasında önemli bir yer tutmakta ve bireylerin fiziksel görünümünün yanı sıra psikolojik sağlıklarını da etkileyen yaygın bir durumdur [4]. Toplumda, cilt hastalıklarının yaygınlığı her geçen gün artmakta; sosyal yaşantıyı, iş hayatını ve bireylerin özgüvenini olumsuz yönde etkileyen önemli bir faktör haline dönüşmektedir [3]. Bazı cilt hastalıkları, cilt kanserlerinde olduğu gibi erken tanı konulmazsa ölüme dahi neden olabilir [3]. Bu nedenle dermatoloji uzmanlarının yapacağı muayene hem cilt hem de genel sağlık açısından çok önemlidir [5]. Cilt hastalıkları dünyadaki en ciddi sağlık sorunudur [5].

Dermatoloji, karmaşıklığı nedeniyle teşhis edilmesi en öngörülemez ve zor alanlardan biridir. Yaygın olmasına rağmen, teşhisi son derece zordur ve alanda kapsamlı deneyim gerektirir [6]. Cilt hastalıkları, çeşitli nedenlere bağlı olarak farklı belirtiler gösterebilen durumlar olarak kabul edilir [7]. Bu hastalıkların tanısında, hastaların deneyimlediği semptomlar büyük bir öneme sahiptir [8]. Dermatolojik hastalıkları teşhis etmeye yönelik geleneksel yaklaşım öncelikle görsel muayeneye ve klinik uzmanlığa dayanmaktadır [8].

**2. Materyal ve Metot**

**2.1. Kullanılan Veri Seti**

Bu çalışmada, dermatolojik hastalıkların sınıflandırılmasına yönelik olarak [Kaggle](https://www.kaggle.com/datasets/olcaybolat1/dermatology-dataset-classification) platformunda yer alan "Dermatology Dataset Classification" adlı açık kaynak veri seti kullanılmıştır [9]. Veri seti, farklı dermatolojik hastalıklara sahip bireylerden toplanan sayısal tıbbi verileri içermekte olup, 34 özellik ve 1 hedef sınıf (class) olmak üzere toplam 35 sütundan oluşmaktadır. Hedef sınıf, altı farklı dermatolojik hastalık kategorisinden birini temsil etmektedir. Veri setinde 366 adet örnek bulunmaktadır [9].

Veri setinde yer alan özellikler; klinik bulgular, laboratuvar test sonuçları ve semptom derecelendirmeleri gibi çeşitli tıbbi parametreleri kapsamaktadır. Eksik veriler minimal düzeyde olup, eksik değerler uygun imputation teknikleri ile doldurulmuştur.

**2.2. Normal Dağılıma Uygunluk Testi ve Özellik Çıkarımı**

**2.2.1.** **Normal Dağılıma Uygunluk Testi**

Veri analizi sürecinde, kullanılacak istatistiksel testlerin belirlenmesi açısından verilerin normal dağılıma uygunluğunun değerlendirilmesi kritik bir adımdır. Bu değerlendirme, parametrik testlerin varsayımlarının sağlanıp sağlanmadığını anlamak için gereklidir [10]. Normal dağılıma uygunluğu test etmek amacıyla çeşitli normallik testleri mevcuttur. Bu çalışmada, veri setinin normal dağılıma uyup uymadığını belirlemek için Kolmogorov-Smirnov (K-S) testi tercih edilmiştir [11].

K-S testi, özellikle örneklem büyüklüğü 50 ve üzeri olan veri setlerinde normallik değerlendirmesi için yaygın olarak kullanılmaktadır [12]. Bu test, örneklem dağılımının teorik normal dağılım ile uyumunu değerlendirir ve sonuç olarak bir p-değeri üretir. Elde edilen p-değeri, genellikle 0.05 anlamlılık düzeyi ile karşılaştırılır. Eğer p-değeri 0.05'ten küçükse, bu durum veri setinin normal dağılıma uymadığı şeklinde yorumlanır [13].

Bu çalışmada analiz edilen veri seti 366 gözlemden oluşmaktadır. Bu nedenle, örneklem büyüklüğü dikkate alınarak K-S testi uygulanmıştır. Test sonucunda elde edilen p-değeri, veri setinin normal dağılıma uymadığını göstermiştir. Bu bulgu doğrultusunda, parametrik testlerin varsayımları sağlanmadığı için parametrik olmayan (non-parametrik) testlerin kullanılması gerektiği sonucuna varılmıştır.

**2.2.2. Özellik Çıkarımı**

Özellik seçimi, sınıflandırma modelinin başarısı üzerinde önemli bir etkiye sahiptir. İyi seçilmiş özellikler, modelin doğruluğunu artırırken, gereksiz özelliklerin dahil edilmesi modelin karmaşıklığını artırabilir ve aşırı öğrenmeye (overfitting) yol açabilir [14]. Bu çalışmada, veri setindeki 33 sürekli özelliğin hedef değişkenle ilişkisi değerlendirilmiştir.

Veri setinde yalnızca sürekli değişkenler bulunduğundan, kategorik değişkenlerin analizi yapılmamıştır. Sürekli değişkenlerin, hedef değişken ile olan ilişkisini incelemek için Kruskal-Wallis H Testi uygulanmıştır. Kruskal-Wallis testi, normal dağılım göstermeyen bağımsız grupların ortalamalarını karşılaştırmak için kullanılan bir parametrik olmayan testtir [15].

Bu test, özellikle birden fazla sınıf içeren kategorik değişkenlerle sınıflandırma yapan veri setlerinde kullanılır. Bu çalışmada, etiket sütununun nominal olduğu ve birden fazla sınıfa sahip olduğu göz önünde bulundurularak Kruskal-Wallis H testi kullanılmıştır. Test sonucunda elde edilen p-değerleri, hangi özelliklerin sınıflandırma için önemli olduğunu belirlemeye yardımcı olmuştur.

Sonuç olarak, normal dağılıma uymayan veri seti için parametrik olmayan testlerin uygun olduğu ve Kruskal-Wallis H Testi’nin kullanılarak sürekli değişkenler arasındaki farkların değerlendirildiği anlaşılmıştır. Bu yaklaşım, modelin performansını optimize etmek amacıyla doğru özelliklerin seçilmesini sağlamaktadır.

**2.3. Ensemble Yöntemleri**

Makine öğreniminde model performansını artırmak ve genelleme yeteneğini güçlendirmek amacıyla kullanılan en etkili tekniklerden biri ensemble yöntemleridir. Ensemble yöntemler, birden fazla zayıf veya orta düzeyde güçlü modelin çıktısını birleştirerek daha başarılı ve kararlı sonuçlar üretir [16]. Bu çalışmada, üç temel ensemble yöntemi olan Bagging, Boosting ve Stacking yaklaşımları uygulanmış ve elde edilen sonuçlar karşılaştırılmıştır.

**2.3.1. Bagging (Bootstrap Aggregating)**

Bagging, özellikle yüksek varyansa sahip modellerin (örneğin karar ağaçlarının) performansını artırmak için kullanılan bir ensemble yöntemidir [17]. Temel prensibi, eğitim verisinin farklı alt örneklemlerine (bootstrap örnekleri) aynı algoritmayı uygulayarak birden fazla model oluşturmak ve bu modellerin tahminlerini ortalamak (regresyon) veya çoğunluk oyu ile birleştirmektir (sınıflandırma).

#### **Random Forest**

Bu çalışmada bagging yaklaşımı kapsamında random forest yöntemi kullanılmıştır. Random Forest, rastgele seçilen özellik alt kümeleri ile çok sayıda karar ağacı eğitilerek oluşturulan güçlü bir topluluk modelidir. Ağaçlar birbirinden bağımsız olarak eğitildiği için aşırı öğrenme (overfitting) riski azaltılmış olur [17].

**2.3.2. Boosting**

Boosting, zayıf öğrenicileri art arda eğiterek her bir yeni modelin, önceki modellerin hatalarını düzeltmeye çalıştığı bir ensemble yöntemidir [18]. Bu çalışmada üç farklı boosting algoritması uygulanmıştır:

* **AdaBoost (Adaptive Boosting)**

İlk boosting algoritmalarından biridir ve yanlış sınıflandırılan örneklere daha fazla ağırlık vererek model performansını artırır [18].

### **2.3.3. Stacking (Stacked Generalization)**

Stacking, birden fazla sınıflandırıcı modelinin birleşiminden faydalanan bir meta-öğrenme yöntemidir [19]. Bu çalışmada, Seviye-0 (Base) ve Seviye-1 (Meta) modeller kullanılarak stacking uygulanmıştır.

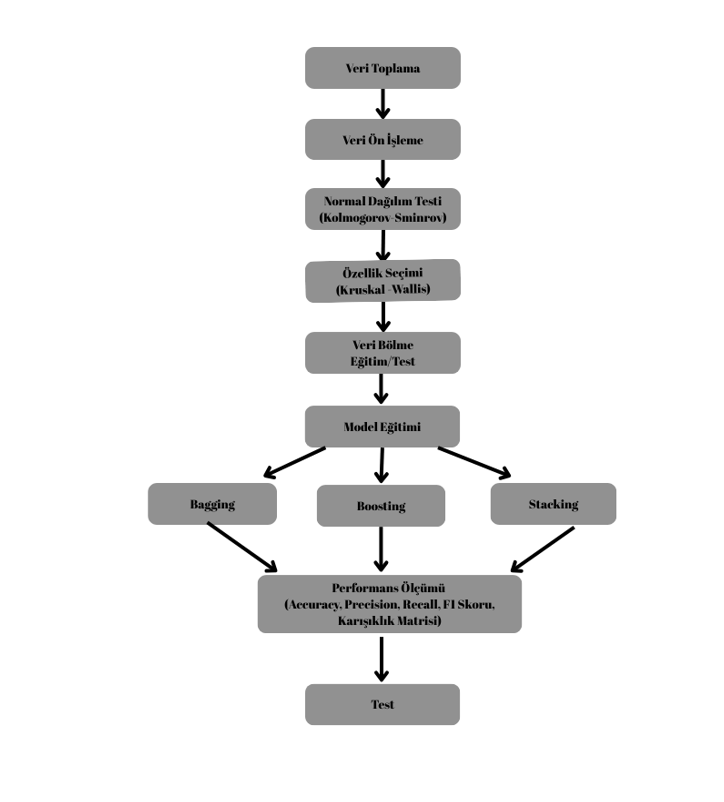
#### **Seviye-0 (Base) Modeller**

* RandomForestClassifier (rf)
* DecisionTreeClassifier (dt)
* KNeighborsClassifier (knn)
* LogisticRegression (lr)
* GaussianNB (gnb)
* XGBClassifier (xgb)
* LGBMClassifier (lgbm)
* GradientBoostingClassifier (gb)

#### **Seviye-1 (Meta) Model**

Seviye-1 model olarak Logistic Regression kullanılmıştır. Bu model, base modellerin tahminlerini girdiler olarak kullanır ve nihai tahmini yapar.

**3. Yöntemin Uygulanması**

Bu diyagram, makine öğrenmesi süreçlerinin sistematik bir şekilde nasıl uygulandığını göstermektedir. Süreç; veri toplama, ön işleme ve istatistiksel analiz adımlarıyla başlayıp, anlamlı değişkenlerin seçimi ve eğitim-test ayrımıyla devam etmektedir. Model eğitimi aşamasında farklı topluluk öğrenme yöntemleri (Bagging, Boosting, Stacking) kullanılarak performans karşılaştırması yapılmaktadır. Elde edilen modeller, çeşitli değerlendirme metrikleri aracılığıyla analiz edilerek en uygun yöntem belirlenmektedir. Bu yapılandırılmış yaklaşım, modelin doğruluğunu ve genellenebilirliğini artırmayı amaçlamaktadır.

**4. Performans Değerlendirme**

Tüm modeller Python programlama dili kullanılarak geliştirilmiş ve **random forest,** **adaboost**, **stacking** gibi popüler makine öğrenme kütüphaneleri kullanılmıştır. Veri setinin %80’i eğitim, %20’u test verisi olarak ayrılmıştır. Eğitim öncesi tüm değişkenler standartlaştırılmış ve eksik veriler ortalama değere göre doldurulmuştur.

Modellerin başarımı aşağıdaki sınıflandırma metrikleriyle değerlendirilmiştir:

**4.1. Doğruluk (Accuracy)**

· Doğruluk, modelin doğru tahmin ettiği örneklerin toplam örnek sayısına oranıdır. En yaygın kullanılan performans metriğidir. Ancak sınıf dengesizliğinin yüksek olduğu veri setlerinde yanıltıcı sonuçlar verebilir [20].

Burada:

· TP (True Positive): Doğru pozitif tahminler

· TN (True Negative): Doğru negatif tahminler

· FP (False Positive): Yanlış pozitif tahminler

· FN (False Negative): Yanlış negatif tahminler

**4.2. Hassasiyet (Precision)**

· Hassasiyet, modelin pozitif olarak tahmin ettiği örneklerin gerçekten pozitif olma oranıdır. Yanlış pozitif (False Positive) sayısının düşük olması gereken uygulamalarda önemlidir. Örneğin yanlış bir hastalık teşhisinin maliyetli olduğu senaryolarda kritik bir metriktir [21].

**4.3. Duyarlılık (Recall)**

· Duyarlılık, gerçek pozitif örneklerin model tarafından doğru şekilde yakalanma oranını ifade eder. Yanlış negatiflerin (False Negatives) düşük tutulması gereken durumlarda önceliklidir. Örneğin, kanserli hastaların atlanmaması gereken bir tanı sisteminde yüksek recall hedeflenir [22].

**4.4. F1 Skoru**

· F1 skoru, hassasiyet (precision) ve duyarlılığın (recall) harmonik ortalamasıdır. Özellikle dengesiz veri setlerinde daha güvenilir bir başarı ölçüsüdür. Yüksek F1 skoru, hem az hatalı hem de kapsayıcı bir model anlamına gelir [23].

**4.5. Karışıklık Matrisi (Confusion Matrix)**

· Karışıklık matrisi, sınıflandırma modelinin tüm tahminlerini görselleştirmeye yardımcı olan bir tablodur. Gerçek sınıflar ile tahmin edilen sınıflar arasındaki ilişkileri satır ve sütunlarda göstererek modelin nerelerde hata yaptığını detaylıca analiz etmeyi sağlar [24]. Her hücre, modelin kaç adet örneği doğru veya yanlış sınıflandırdığını gösterir.

Bu değerlendirme metrikleri sayesinde her bir ensemble yöntemin başarı düzeyi objektif olarak karşılaştırılmıştır. Ayrıca stacking yöntemiyle, temel modellerin birlikte nasıl daha güçlü bir sınıflandırıcı oluşturduğu analiz edilmiştir.

**5.Deneysel Çalışmalar**

**TABLO I. Farklı Sınıflandırıcıların Performans Metrikleri ve Eğitim Süreleri**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **T**  **M** | **Accuracy** | **Precision** | **Recall** | **F1** | **Training Time** |
| **Random Forest** | **0.972222** | **0.972222** | **0.972222** | **0.972222** | **0.673412** |
| **AdaBoost** | **0.597222** | **0.487689** | **0.597222** | **0.508079** | **0.206697** |
| **Stacking** | **0.944444** | **0.943827** | **0.944444** | **0.943587** | **10.591519** |

Random Forest modeli, %97.2 doğruluk ve en yüksek F1-skoru ile en başarılı performansı sergilemiştir. Öte yandan, AdaBoost modeli en düşük başarı metriklerine sahipken, Stacking modeli yüksek doğruluk oranı ile dikkat çekmiş; ancak en uzun eğitim süresine ihtiyaç duymuştur.

**6. Sonuçlar**

### Bu çalışmanın amacı, dermatolojik hastalıkların sınıflandırılmasında farklı ensemble yöntemlerinin performansını karşılaştırmak ve en etkili modeli belirlemektir. Kullanılan veri seti üzerinde Random Forest, AdaBoost ve Stacking gibi yöntemler uygulanarak doğruluk (accuracy), hassasiyet (precision), duyarlılık (recall) ve F1 skoru gibi metriklerle değerlendirmeler yapıldı. Elde edilen sonuçlara göre, Random Forest modeli %97.2 doğruluk oranı ve en yüksek F1 skoru ile en başarılı performansı sergiledi. AdaBoost ise diğer yöntemlere kıyasla daha düşük başarı metrikleri gösterirken, Stacking yöntemi yüksek doğruluk sağlamasına rağmen uzun eğitim süresi gerektirdi.

### Bu çalışma, dermatolojik hastalıkların teşhisinde makine öğrenmesi yöntemlerinin potansiyelini ortaya koymuştur. Özellikle Random Forest gibi ensemble yöntemlerinin, karmaşık ve çok sınıflı veri setlerinde yüksek performans sağlayabileceği görülmüştür. Ancak, modelin genellenebilirliğini artırmak için daha büyük ve çeşitli veri setleri üzerinde çalışılması önerilir. Ayrıca, derin öğrenme modellerinin (örneğin CNN'ler) bu alanda nasıl bir performans sergileyeceğinin araştırılması ve farklı özellik çıkarımı tekniklerinin denenmesi gelecek çalışmalar için önemli adımlar olabilir. Sonuç olarak, bu tür modellerin klinik uygulamalarda destekleyici bir araç olarak kullanılması, dermatolojik teşhis süreçlerinin hızlandırılmasına ve doğruluğunun artırılmasına katkı sağlayabilir.

**Kaynakça**

[1] A Life Sağlık Grubu. (2017, 25 Nisan). *Dermatoloji Birimi*.<https://www.alifesaglikgrubu.com.tr/birimler/dermatoloji/103>

[2] Almustafa, K. M. (2025). Predictive modeling and optimization in dermatology: Machine learning for skin disease classification. *Computers in Biology and Medicine*, *189*, 109946, doi: 10.1016/j.compbiomed.2025.109946

[3] Girne Üniversitesi Hastanesi. (t.y.). *Deri ve Zührevi Hastalıklar*.<https://hospital.kyrenia.edu.tr/tibbi-birimler-ve-merkezler/deri-ve-zuhrevi-hastaliklar/>

[4] Tedavisi Bu. (t.y.). *Yüzdeki Cilt Hastalıkları: Nedenleri ve Tedavi Yöntemleri*.<https://tedavisibu.com/soru-cevap/yuzdeki-cilt-hastaliklari-nedenleri-ve-tedavi-yontemleri/>​

[5] S. Borade and D. Kalbande, "Survey paper based critical reviews for Cosmetic Skin Diseases," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India, 2021, pp. 580-585, doi: 10.1109/ICAIS50930.2021.9395803.

[6] VB Kumar, SS Kumar ve V. Saboo, "Görüntü işleme ve makine öğrenimi kullanılarak dermatolojik hastalık tespiti," *2016 Üçüncü Uluslararası Yapay Zeka ve Desen Tanıma Konferansı (AIPR)* , Lodz, Polonya, 2016, s. 1-6, doi: 10.1109/ICAIPR.2016.7585217.

[7]I. Gonçalves Ferreira, M. Blessmann Weber, and R. Rangel Bonamigo, "History of dermatology: the study of skin diseases over the centuries," *Anais Brasileiros de Dermatologia*, vol. 2021, no. 16, pp. 1-5, 2021, doi: 10.1016/j.abd.2020.09.006.

[8] BK Patel ve R. Patel, "Cilt Hastalıklarının Tanısı İçin Makine Öğrenmesi Algoritmaları Üzerine Bir İnceleme", *Biyomedikal Görüntülemede Makine Öğrenmesi Dergisi* , cilt 1, sayı 4, s. 345-360, 2023.

[9] Bolat, Olcay. "Dermatology Dataset (Multi-class classification)." *Kaggle*, 2023,<https://www.kaggle.com/datasets/olcaybolat1/dermatology-dataset-classification>.

[10] Tabachnick, B. G., & Fidell, L. S. (2013). Using Multivariate Statistics. Pearson Education.

[11] Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling tests. Journal of Statistical Modeling and Analytics, 2(1), 21-33.

[12] Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. International Journal of Endocrinology and Metabolism, 10(2), 486-489.

[13] Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. Journal of Statistical Computation and Simulation, 81(12), 2141-2155.

[14] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182.

[15] McDonald, J. H. (2014). Handbook of Biological Statistics (3rd ed.). Sparky House Publishing.

[16]Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. Springer.

[17] Breiman, L. (1996). Bagging Predictors. Machine Learning, 24(2), 123-140.

[18] Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences, 55(1), 119-139.

[19] Wolpert, D. H. (1992). Stacked Generalization. Neural Networks, 5, 241-259.

[20] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437.

[21] Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2(1), 37-63.

[22] Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233-240).

[23] Chinchor, N. (1992). MUC-4 Evaluation Metrics. In *Proceedings of the 4th Conference on Message Understanding* (pp. 22-29).

[24] Provost, F., & Fawcett, T. (2013). Data Science for Business. O'Reilly Media.