

Üniversite	:	T.C. İstanbul Kültür Üniversitesi
Enstitüsü	:	Lisansüstü Eğitim Enstitüsü
Anabilim Dalı	:	Bilgisayar Mühendisliği
Program	:	Bilgisayar Mühendisliği
Tez Danışmanı	:	Doç. Dr. Özgür Koray ŞAHİNGÖZ
Tez Türü ve Tarihi	:	Yüksek Lisans – Ağustos 2020

ÖZET

MAKİNE ÖĞRENME Sİ YÖNTEMLERİ İLE YAZILIM HATA TAHMİNİ

Günümüzde artan şekilde devam eden kritik işlemler ve bu işlemlerin detaylı süreçleri büyük çoğunlukla yazılımlar ile geliştirilmektedir. Bu yüzden ürünün yazılımının kalitesi ve yazılımının kusursuz olması; yazılımı geliştiren, yazılımı takip eden, ortaya çıkan ürünü test eden veya ürünü kullanan herkes için önemli bir uzmanlaşma alanı haline gelmiştir. Kalite modeli, bir ürünün kalitesini değerlendirmede sisteminin temel yapı taşıdır. Kalite modeli, bir yazılım ürününün özelliklerini değerlendirirken hangi kalite özelliklerinin dikkate alınacağını belirler. Bir sistemin kalitesi, sistemin çeşitli paydaşlarının belirtilen ve ima edilen ihtiyaçlarını karşılama ve bu sayede değer sağlama derecesidir. Bu paydaşların ihtiyaçları (işlevsellik, performans, güvenlik, sürdürülebilirlik vb.), ürün kalitesini özellikler ve alt özellikler olarak kategorize eden kalite modelinde tam olarak temsil edilen şeydir. ISO / IEC 25010'da tanımlanan ürün kalite modeli sekiz kalite özelliğini içermektedir. Bunlar; fonksiyonel uygunluk, performans verimliliği, uygunluk, kullanılabilirlik, güvenilirlik, güvenlik, idame ve taşınabilirliktir.

Yazılım Mühendisliği kavramında, yazılım hatalarının tahmini, yazılım geliştirme yaşam döngüsünün en kritik ve pahalı aşamalarından biri olan yazılım sistemlerinin kalitesini artırmada hayati bir rol oynamaktadır. Günlük yaşamımızda yazılım sistemlerinin kullanımı artarken dolayısıyla bağımlılıkları ve karmaşıklıkları da artmaktadır ve bu da hatalar için uygun bir ortam sağlamaktadır. Yazılım hatalarından dolayı, yazılım yanlış sonuçlar ve davranışlar üretmektedir. Hatalardan

daha kritik olan şey ise bu hataların meydana gelmeden bulunmasıdır. Bu nedenle, yazılım hatalarının tespiti ve ayrıca tahmini, yazılım yöneticilerinin bakım ve test aşamaları için kaynakların verimli bir şekilde tahsis edilmesini sağlar. Literatürde, yazılım hatalarının tahmini için farklı öneriler vardır. Bu çalışmada, Karar Ağacı, Naif Bayes, K-En Yakın Komşu, Destek Vektör Makinesi, Rastgele Orman, Ekstra Ağaçlar, Adaboost, Gradient Boosting olarak toplamda 10 makine öğrenmesi algoritması üzerinde çalışılarak ve bu algoritmaları birbirleri ile karşılaştırarak makine öğrenme tabanlı yazılım hatası tahmin sistemleri hakkında karşılaştırmalı bir analiz yapılmıştır.

Günümüzde yazılım hata tahmini araştırmalarında kullanılan veri kümeleri dengeli veri ve dengesiz veri olarak adlandırılmaktadır. Dengeli olmayan veri birbirinden farklı sınıflarda kayıt sayılarının eşit olmadığı bir yapıda sınıf dağılımı olacak şekilde olmasıdır. Dengeli veri ise sınıf dağılımı eşit olan veri kümelerine denilmektedir.

Anahtar Kelimeler: Yazılım Hata Tahmini, Makine Öğrenmesi, Sınıflandırma, Python ile Sınıflandırma.

University : T.C. İstanbul Kültür University
Institute : Institute of Graduate Studies
Department : Computer Engineering
Program : Computer Engineering
Thesis Advisor : Assoc. Prof. Özgür Koray ŞAHİNGÖZ
Degree Awarded And Date : MA – August 2020

ABSTRACT

SOFTWARE DEFECT PREDICTION WITH MACHINE LEARNING METHODS

Today, increasingly ongoing critical processes and detailed processes of these processes are mostly developed by software. Therefore, the quality of the software and the flawless software of the product; It has become an important area of specialization for anyone who develops the software, follows the software, tests the resulting product, or uses the product. The quality model is the basic building block of the system in evaluating the quality of a product. The quality model determines which quality features to consider when evaluating the features of a software product. The quality of a system is the degree to which the system meets the stated and implied needs of the various stakeholders and thus provides value. The needs of these stakeholders (functionality, performance, safety, sustainability, etc.) are exactly what is represented in the quality model that categorizes product quality into features and sub-features. The product quality model defined in ISO / IEC 25010 includes eight quality characteristics: functional suitability, performance efficiency, suitability, usability, reliability, security, maintenance, and portability.

In the concept of Software Engineering, the prediction of software errors plays a vital role in improving the quality of software systems, one of the most critical and expensive stages of the software development life cycle. As the use of software systems increases in our daily life, their dependencies and complexities are also increasing, and this provides a suitable environment for errors. Due to software errors, the software produces incorrect results and behavior. More critical than errors is that these errors are found before they occur. Therefore, the detection, as well as the prediction of software errors, enables software administrators to efficiently allocate resources for maintenance and testing phases. In the literature, there are different

suggestions for the prediction of software errors. In this study, by working on a total of 10 machine learning algorithms such as Decision Tree, Naive Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest, Extra Trees, Adaboost, Gradient Boosting and comparing these algorithms with each other, machine learning-based software error prediction systems a comparative analysis was made about.

Data sets used in software error prediction research today are called balanced data and unbalanced data. Unbalanced data is that there is a class distribution in a structure where the number of records in different classes is not equal. Balanced data are data sets with the equal class distribution.

Keywords: Software Defect Prediction, Machine Learning, Classification, Classification with Python.

