UNIVERSITETI I PRISHTINËS “HASAN PRISHTINA”

FAKULTETI I SHKENCAVE MATEMATIKE-NATYRORE

DEPARTAMENTI I MATEMATIKËS

PROGRAMI SHKENCË KOMPJUTERIKE

Punim Diplome

Shkallëzimi automatik i mikroshërbimeve në kohë reale përmes të Mësuarit e Makinës

Mentori: **Kandidati:**

Prof. asoc. Dr. Ermir Rogova **Bsc . Erblin Halabaku**

Prishtinë

Qershor 2025

Abstrakt

Kompleksiteti në rritje i mikroshërbimeve dhe modelet dinamike të ngarkesës në mjediset e përshtatura për në kloud kërkojnë mekanizma më inteligjentë dhe në kohë reale për shkallëzimin automatik të burimeve. Strategjitë tradicionale të autoshkallëzimit shpesh bazohen në pragje statike ose rregulla reaktive, të cilat nuk mjaftojnë për të përballuar shpërthime të papritura të ngarkesës apo për të garantuar përdorim optimal të burimeve. Kjo temë propozon një qasje inovative që shfrytëzon modelet e të mësuarit të makinës për të zbatuar autoshkallëzimin në kohë reale të mikroshërbimeve, duke parashikuar tendencat e ngarkesës. Sistemi monitoron metrikat e përdorimit të resurseve nga shërbimet e brendshme dhe i furnizon ato në një model të trajnuar parashikues, duke e bërë të mundur rritjen ose uljen e numrit të replikave të kontejnerëve në varësi të parashikimeve. Zgjidhja e propozuar është testuar në një mjedis të kontenierizuar me Kubernetes, duke përdorur simulime të bazuara në gjurmë reale të ngarkesës. Rezultatet tregojnë përmirësime të dukshme në zvogëlimin e shkeljeve në lidhje me kohën e kthimit të përgjigjes dhe përdorim më efikas të burimeve, në krahasim me metodat tradicionale të autoshkallëzimit. Tema bazohet në një analizë të strukturuar dhe rigoroze që filtron dhe analizon kërkime të fundit nga bibliotekat digjitale më të njohura, duke siguruar që metodologjia të jetë në përputhje me njohuritë më të avancuara bashkëkohore. Kjo temë përfaqëson një kontribut teorik dhe praktik në fushën e kompjutimit në kloud, me aplikime të mundshme në optimizimin e performancës, reduktimin e kostove dhe rritjen e besueshmërisë së shërbimeve.

Fjalë kyçe: Të Mësuarit e Makinës, Mikroshërbime, Autoshkallëzim

Abstract

The increasing complexity of microservices and dynamic workload patterns in cloud-native environments demand more intelligent, real-time scaling mechanisms. Traditional autoscaling strategies often rely on static thresholds or reactive rules, which are insufficient to handle sudden workload bursts or ensure optimal resource usage. This thesis introduces an innovative approach that leverages machine learning models to predict workload trends and perform automatic, real-time autoscaling of microservices. By monitoring resource usage metrics from internal services and feeding them into a trained predictive model, the system anticipates demand and adjusts container replicas accordingly. The proposed solution is tested in a containerized Kubernetes environment using trace-driven simulations. It demonstrates improved performance in terms of reduced response time violations and more efficient resource allocation compared to rule-based autoscaling methods. The study is grounded in a rigorous mapping study that filters and analyzes recent research from top digital libraries, ensuring the methodology is aligned with state-of-the-art knowledge. This work provides both a theoretical and practical contribution to the field of cloud computing, with potential applications in performance optimization, cost reduction, and enhanced service reliability.

Keywords: Machine Learning, Microservices, Autoscaling