Sveučilište u Zagrebu

FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE

Marija Jelena Lovrić Štefiček

RAZVOJ MODELA STROJNOGA UČENJA ZA PROCJENU RAZINA ONEČIŠĆUJUĆIH TVARI U UNUTARNJEM ZRAKU

DOKTORSKI RAD

Zagreb, 2026.

University of Zagreb

FACULTY OF CHEMICAL ENGINEERING AND TECHNOLOGY

Marija Jelena Lovrić Štefiček

DEVELOPMENT OF MACHINE LEARNING MODELS FOR ASSESSMENT OF POLLUTANT LEVELS IN INDOOR AIR

DOCTORAL THESIS

Zagreb, 2026

SAŽETAK

Praćenje i upravljanje kvalitetom unutarnjeg zraka postalo je jedno od ključnih javnozdravstvenih pitanja budući da ljudi u razvijenim zemljama većinu vremena provode u zatvorenim prostorima. Dugotrajna izloženost lebdećim česticama, policikličkim aromatskim ugljikovodicima (PAU) vezanim na lebdeće čestice i radonu povezana je s respiratornim i kardiovaskularnim bolestima, moždanim udarom, rakom pluća te pojavom glavobolja, umora i respiratornih infekcija. Unutarnji zrak je izložen kombiniranom utjecaju vanjskih i unutarnjih izvora, pri čemu kompleksna varijabilnost emisija i blizina zona disanja predstavljaju ključni izazov u analizi kvalitete unutarnjeg zraka.

PM_1 frakcija lebdećih čestica, tj. čestice s aerodinamičkim promjerom manjim od 1 μm, zadržavaju se dugo u zraku, prodiru duboko u respiratorni sustav i ulaze u krvotok, pri čemu prenose toksične organske spojeve i teške metale. PAU-i nastaju pri nepotpunom izgaranju i vezani su na lebdeće čestice, a rizik izloženosti pojačan je zimi i u sezoni grijanja zbog povećanih emisija i smanjene atmosferske disperzije. Radon je prirodno prisutan radioaktivni plin koji nastaje radioaktivnim raspadom radija, a nalazi se u tlu i stijenama. U unutarnji zrak dospijeva kroz pukotine i procijepe u temeljima.

Primjena modela strojnog učenja za procjenu kvalitete unutarnjeg zraka sve je učestalija, no postoje ograničenja vezana uz mali broj promatranih onečišćujućih tvari, specifične tipove objekata te nedostatak integriranog pristupa koji uključuje fizikalne, kemijske, okolišne i ljudske čimbenike. Primarni cilj ovog istraživanja je razvoj i evaluacija regresijskih modela strojnog učenja za procjenu razina PM_1, PAU-a i radona u unutarnjem zraku kućanstava, uz naglasak na usporedbu učinkovitosti, robusnost i interpretabilnost modela.

Istraživanje se temelji na podacima prikupljenima u sklopu EDIAQI projekta na području Zagreba i okolice, gdje su paralelno uzorkovane tjedne koncentracije PM_1 i PAU-a u unutarnjem i vanjskom zraku te koncentracije aktivnosti radona u kućanstvima. Primijenjen je strukturirani analitički pristup koji obuhvaća kemometrijsku analizu, analizu glavnih komponenti, inženjerstvo značajki i SHAP metodu za interpretaciju, s ciljem pouzdane procjene razina onečišćujućih tvari u realnim uvjetima i identifikacije dominantnih izvora onečišćenja.

Ispitane su razlike u masenim koncentracijama PM_1 frakcije lebdećih čestica s obzirom na mikro-lokaciju kućanstava te je analizirana sezonska raspodjela PM_1 i PAU-a na području Zagreba i okolice. Prvi korak razvoja modela strojnog učenja obuhvatio je detaljnu predobradu i definiranje skupova podataka, uključujući usporedbu različitih metoda predobrade i njihov utjecaj na rezultate modela. Za odabir najučinkovitijih pristupa napravljena je usporedba regresijskih modela temeljenih na stablima i udaljenostima. Novost istraživanja je strukturirani analitički okvir koji kombinira tradicionalne statističke pristupe, inženjerstvo značajki te razvoj i interpretaciju modela strojnog učenja.

Među analiziranim modelima, model pojačavanja gradijenta (GBR) je pokazao najbolje rezultate za sve promatrane onečišćujuće tvari. Međutim, uočene razlike u optimalnim hiperparametrima, odabiru relevantnih značajki i postupcima validacije modela naglašavaju važnost predobrade podataka i optimizacije modela. Time se potvrđuje potreba za prilagodbom pristupa svakoj pojedinoj onečišćujućoj tvari. Razvijeni modeli primjenjivi su prvenstveno kao interpolacijski alati za kućanstva sličnih karakteristika na području Zagreba i okolice.

Ključne riječi: kvaliteta unutarnjeg zraka, onečišćenje vanjskog zraka, lebdeće čestice, policiklički aromatski ugljikovodici, regresijski algoritmi, strojno učenje, *SHapley Additive exPlanations*, javno zdravstvo

ABSTRACT

Monitoring and management of indoor air quality has become one of the key public health issues since people in developed countries spend most of their time indoors. Long-term exposure to particulate matter, polycyclic aromatic hydrocarbons (PAHs) bound to particulate matter and radon is associated with respiratory and cardiovascular diseases, stroke, lung cancer and the occurrence of headaches, fatigue and respiratory infections. Indoor air is exposed to the combined influence of external and internal sources, where complex variability of emissions and proximity to breathing zones represent a key challenge in the analysis of indoor air quality.

PM_1 fraction of particulate matter, i.e. particles with an aerodynamic diameter of less than 1 μm, remain in the air for a long time, penetrate deep into the respiratory system and enter the bloodstream, carrying toxic organic compounds and heavy metals. PAHs are formed during incomplete combustion and are bound to particulate matter, with the exposure risk increasing in winter and during the heating season due to increased emissions and reduced atmospheric dispersion. Radon is a naturally occurring radioactive gas that is formed by the radioactive decay of radium and is found in soil and rocks. It enters indoor air through cracks and crevices in the foundation.

The application of machine learning models for indoor air quality assessment is becoming more frequent, but there are limitations related to the small number of observed pollutants, specific types of facilities, and the lack of an integrated approach that includes physical, chemical, environmental and human factors. The primary goal of this research is to develop and evaluate regression machine learning models for assessing PM_1, PAHs and radon levels in household indoor air, with an emphasis on comparing model performance, robustness and interpretability of the models.

The research is based on data collected as part of the EDIAQI project in the Zagreb area, where parallel weekly PM_1 and PAH concentrations in indoor and outdoor air, as well as radon activity concentrations in households were sampled in parallel. A structured analytical approach was applied, including chemometric analysis, principal component analysis, feature engineering, and the SHAP method for interpretation, with the aim of reliably assessing pollutant levels in real conditions and identifying dominant sources of pollution.

The differences in mass concentrations of PM_1 fraction of particulate matter were examined with respect to the micro-location of households, and the seasonal distribution of PM_1 and PAHs in the area of Zagreb and its surroundings was analysed. The first step in the development of machine learning models included detailed pre-processing and data set definition, including a comparison of different pre-processing methods and their impact on model results. A comparison of regression models based on trees and distances was made to select the most effective approaches. The novelty of the research is a structured analytical framework that combines traditional statistical approaches, feature engineering, and machine learning model development and interpretation.

Among the analysed models, the gradient boosting regression (GBR) model showed the best results for all observed pollutants. However, the observed differences in optimal hyperparameters, feature selection, and model validation procedures highlight the importance of data pre-processing and model optimization. This confirms the need for tailoring the approach to each individual pollutant. The developed models are primarily applicable as interpolation tools for households with similar characteristics in the Zagreb area.

Keywords: indoor air quality, outdoor pollution, particulate matter, polycyclic aromatic hydrocarbons, regression algorithms, machine learning, SHapley Additive exPlanations, public health