Thesis abstract

“Air pollution forecasting system in Sofia”

Air pollution is one of the most significant environmental problems in the modern world, posing serious risks to human health and the environment. This thesis investigates the process of air quality prediction by applying machine learning and data analysis methods. The main objective is to develop an integrated system that collects, processes, stores and analyzes data from sensor devices to provide accurate predictions of ambient air pollutant concentration levels.
The work starts with an analysis of existing approaches and technologies for air monitoring and pollutant concentration prediction, and a comparative analysis of different prediction models and algorithms including physical, statistical and machine learning models. It then moves on to an in-depth comparative analysis of different types of data mining, processing and storage technologies, as well as platforms for hosting infrastructure and deploying artificial intelligence models.
The next step is a detailed analysis of the system requirements, which includes both functional and non-functional requirements. The functional requirements cover the ability of the system to collect big data from a dynamic sensor network, process and store it in a secure environment, apply machine learning models, and provide the results through a user-friendly interface. Non-functional requirements include scalability, security, reliability and processing efficiency to ensure the robust operation of the system in real-world conditions.
After analyzing the requirements, we proceed to design a system architecture that includes modules for retrieval, storage, and processing of large amounts of data, as well as mechanisms for building machine learning models and inference. Based on the analysis, an LSTM (Long Short-Term Memory) model is selected. In addition to the core layers focused on processing data streams, auxiliary components are designed into the system to provide support and optimization of the core workflows. These include mechanisms for orchestration, monitoring, security, model retraining and other functionalities that contribute to the reliability of the overall system.
The system implementation uses the AWS cloud platform, where data is stored and processed through services such as Redshift, S3 and Glue, and the forecasting model is run in SageMaker. Testing includes both modular and system testing to ensure the correct operation of the individual components and their interaction in the overall workflow.
Finally, the system is able to produce predictions on the levels of the main pollutants, making it a useful tool for monitoring and managing air quality. Opportunities for future development include integrating additional data sources, refining the prediction model, and extending the system with functionalities to predict other environmental factors such as noise pollution and water quality.