Smoking Status Prediction Based on Kaggle Data

Presenters: Luka namoradze, Luka kakriashvil, Giorgi apciauri, Nika iniashvili

Email: nika.iniashvili767@ens.tsu.edu.ge
Department of Computer Science, Faculty of Exact and Natural Sciences, Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia

This project aims to predict an individual's smoking status based on the analysis of a dataset from Kaggle. This task is important for public health, as timely and accurate identification supports planning of preventive measures.

The project is based on modern data science methodologies and includes data preprocessing, feature selection, application of classification algorithms, and analysis of results. Algorithms used include Logistic Regression, Random Forest, Gradient Boosting (XGBoost, LightGBM), implemented using Python libraries: Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn.

Model evaluation was performed using metrics such as Accuracy, ROC-AUC, Precision, Recall, and F1 Score. The best result was achieved by the Gradient Boosting algorithm (ROC-AUC ≈ 92%). Key factors associated with smoking were identified, including age, gender, and lifestyle.

The project represents a practical application of data science in the medical field and contributes to data-driven decision-making in healthcare.

References

[1] Kaggle – Smoking Status Prediction Dataset. https://www.kaggle.com