Air Quality Prediction using Machine Learning With Data-Driven Approach to Environmental Monitoring

Abdul Wadud Chowdhury
Department of Computer Science and Technology
Ulster University
Birmingham, United Kingdom
Chowdhury-AW1@ulster.ac.uk

MD Saharab Hossain
Department of Computer Science and Technology
Ulster University
Birmingham, United Kingdom
Hossain-MS13@ulster.ac.uk

Jenil J Vekariya
Department of Computer Science and Technology
Ulster University
Birmingham, United Kingdom
vekariya-jj@ulster.ac.uk

Zunair Ahmad
Department of Computer Science and Technology
Ulster University
Birmingham, United Kingdom
Ahmad-z2@ulster.ac.uk

***Abstract*— This paper presents an ML-based framework to predict the Air Quality Index (AQI) from a historical set with multiple air pollutant observations. The data quality enhancement involved Min–Max normalization and noise reduction methods, i.e., Exponential Weighted Mean and Savitsky–Golay smoothing to enhance data quality. Correlation analysis and mutual information regression were utilized to gauge feature relevance to identify principal AQI predictors. The developed framework was implemented on a Jupyter Notebook platform to provide portability and reproducibility to varied modeling approaches. Four prediction algorithms ; SVM-Support Vector Machine LSTM-Long Short-Term Memory, CNN-Convolutional Neural Network with LSTM and Transformer − were written and thoroughly tested. Among all those implemented, SVM model attained maximum prediction accuracy of 94% during prediction on importance gain−based datasets. The outcome presents the effectiveness of developed data quality enhancement pipeline and establishes the feasibility of ML approaches, particularly SVM to provide sound and reliable AQI prediction.**

Keywords*— *Air Quality Index (AQI), Air Pollutant, Support Vector Machine, Feature Engineering, Long Short-Term Memory, Convolutional Neural Network with LSTM, AQI Prediction.

I. INTRODUCTION

Air pollution is increasingly an environmental and public health issue due to rapid urbanization and industrialisation. The severe implications of air pollution emphasise the specific forecasting & monitoring, particularly in the most populated regions [1]. Pollution in air is a major problem to urban planning, sustainability of the environment and public health. Forecasting air quality is essential to develop effective management plans and alerting systems [2]. Monitoring PM2.5 pollution is important to manage sufficient pollutant thresholds. In addition to prompt PM2.5 monitoring, globally the Air Quality Index (AQI) monitors ambient pollution benefits as well. The AQI serves as enough air quality predictor against pollution levels and for creating quantity that combines different pollutant concentrations or that is simply easier for the public [3], academics, and works as a perspective for policy response to intolerable actions to exposure. Accurate AQI forecasting is also an important tool for pollution risk managing; early alert paradigm development; and proactive long-term management of environmental risk.

Some predictive methodologies using machine & deep learning are proposed as possible AQI predictors, specifically because they use the practice of model to simulate the intersectionality complexity and non-linear dependence across meteorological variables and pollution dependency variables [4]. Since ML are able to learn complex non-linear structure representations within high-dimension datasets, several ML techniques including, Random Forest, LSTM and Gradient Boosting networks would be powerful, especially if systems (e.g., ML models) could learn and knowledge update through prior pollution trajectories and simulate, at a specified t, the AQI patterns based on its prior occurrence pattern focusing on its possible PL levels as well as give predictions of AQI in unique environmental networks when compared to standard statistical modeling methods. Nevertheless, preprocessing and error-free feature engineering on unprocessed environmental data are necessary for predictive power.

A. Related Works

Establishing a mathematical formula for the air quality index can be done in a variety of ways. Exposure to air pollution has been linked in numerous studies to detrimental health consequences for a community. One of the most intriguing approaches to studying and modeling AQI is data mining [5]. Using Random Forest and Gradient Boosting, Kumar Jain. discovered that feature selection enhanced performance. Using LSTM networks, Li et al. demonstrated that their models' prediction accuracy was higher than that of the baseline linear models by incorporating long-term dependencies.

TABLE I. LITERATURE REVIEW

No.	Title	Goal	Result
1	A Comparative Study of Machine Learning Techniques for	Compare ML techniques	Some ML models achieved higher accuracy and

	AQI Prediction [6]	for AQI forecasting	reliability than others
2	ML-Based Global AQI Development Using Satellite Data [7]	Build a global AQI model using satellite data for regions without monitoring stations	Model effectively predicted AQI in unmonitored regions
3	Air Quality Prediction by ML Models: A Predictive Study [8]	Evaluate selected ML models for AQI levels	CatBoost showed strong predictive performance
4	AQI Prediction Using ML for Ahmedabad City [9]	Develop local AQI model with city-level data	Preprocessing (feature selection, normalization, etc.) improved accuracy
5.	AQI Prediction Model Using Deep Learning in IoT Environments [10]	Apply deep learning on IoT-based spatial–temporal data	Model successfully captured both geographical and temporal AQI patterns

Additionally, they pointed out that preprocessing, such as smoothing and normalization, has a positive impact on the results shown. Although Chen and Yu [9] demonstrated that feature selection (e.g., Mutual Information Regression) produced a better generalization with the model, Zhang et al. found that it improved the stability of estimates. However, few research have tried combining several smoothing techniques with a thorough feature analysis, which ultimately limits prediction effectiveness.

B. Goals and Purpose

The goal of this study is to purpose an effective framework for AQI-prediction based on multi-pollutant historical data, focusing on preprocessing, trend analysis, and feature selection, and comparing SVM, LSTM, CNN-LSTM, and Transformer model performance to determine the best strategy.

Here are the objectives:

1. Carry out data preprocessing, statistical trend analysis, and machine learning scaling with pollution variables to denoise and emphasize important patterns in the data.
2. Apply feature selection techniques (Pearson correlation, mutual information) to determine the features that are most relevant.
3. Train, develop, and test SVM, LSTM, CNN-LSTM, and Transformer models, compare predictions, and recommend a model for AQI prediction.

II. Methodology

A. Dataset Description

The study intended to use air pollution data from multiple areas in India from a previously published Kaggle dataset [12].

TABLE II. DATASET DESCRIPTION

Parameter	Value
Dataset name	Air Pollution Data (India)
Sampling time	1 day
Data size	23,504 samples
Duration	30-11-2020 to 25-05-2023
No. of pollutants measured	9

The dataset contains daily air quality index (AQI) and a range of pollutant concentrations. Details of the dataset are available in Table 1. The data also contains nine overall pollutant metrics (as shown in TABLE III.).

TABLE III. POLLUTANT DESCRIPTION

Abbreviation	Pollutant Name
AQI	Air-Quality-Index
CO	Carbon-Monoxide
NO	Nitric-Oxide
NO_2	Nitrogen-Dioxide
O_3	Ozone
SO_2	Sulphur-Dioxide
PM-2.5	fine 2.5 micrometers or less of tiny particle debris.
PM-10	coarse particles with a size of 10 micrometers or less
NH_3	Ammonia

B. Data Preprocessing

Normalization has potential to enhance accuracy, but just as importantly, there are some algorithms that will not be heavily influenced by these normalizations, giving clarity on when and how to apply normalization to best effects [13]. For context, normal ranges for SO_2 REM concentrations are often <200 µg/m³ and CO is sometimes greater than 9000 µg/m³, which results in large scale discrepancies. To ensure all pollutants contribute equivalently, as well as assist with model learning, all features, including AQI, were Min-Max normalized to a [0-1] range (Fig.1).

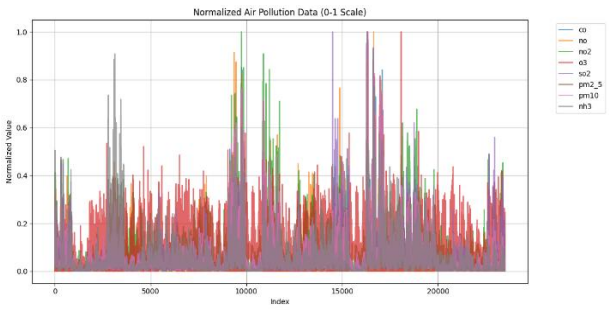


Fig. 1. Normalized Air Pollution Data

Savitzky–Golay is used to mitigate short-term variation, and an application of the Modified Varri method to detect

amplitude and frequency changes. The simulations using synthetic and real EEG data show that the proposed solution produces superior detection of amplitude and frequency change than the original Modified Varri method [14]. AQI smoothing comparisons for the first 1000 samples. The raw AQI data shows significant short-term fluctuations. Moving Average, Exponential, and Savitzky–Golay filters smoothed the AQI time-series so that long-term variations can be more clearly observed (Fig. 2).

Fig. 2. AQI Smoothing Comaprison

C. *Feature Engineering*

Feature engineering is based on principles rather than practices [15]. Analysis of correlation to identify and remove features that highly-correlated, importance of feature and gain to distinguish which features add the most predictive power and various feature ranking techniques with different phases of budgeting to engineer features to improve model accuracy and interpretability.

1. Correlation Feature Selection: CFS is a filter method that identifies feature subsets that are strongly associated with the target class while weakly associated with each other [16]. CFS reduces the dimensionality of the feature space by identifying and removing irrelevant and redundant attributes while maintaining (and often improving) classification performance. Further, CFS has been demonstrated to be a more efficient computational method than wrapper methods but achieves similar predictive performance, especially on smaller sized datasets. After applying CFS to our dataset, the remaining features are: [city, date, aqi, co, no, no_2, o_3, so_2, nh_3].

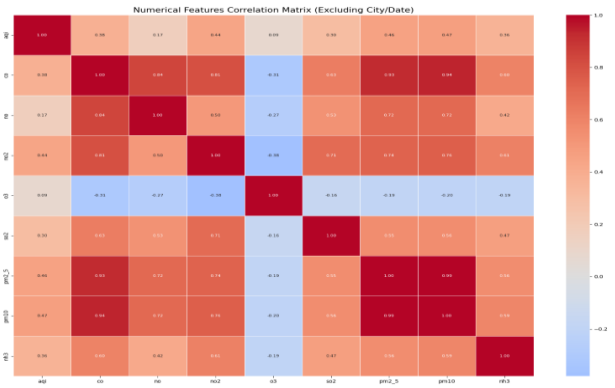


Fig. 3. Numerical Features Correlation Matrix

2. Importance Gain :The Feature Importance Ranking Measure (FIRM) provides a retrospective analysis framework to leverage both robust predictive performance and ability for interpretation while considering the correlation structure among features. FIRM is capable of identifying the most relevant variables, even under noisy conditions, and, as such, provides a more stable measure of feature importance than simple weightings of features [17]. After implementing importance ranking we have remaining features : ['pm2_5', 'pm10', 'nh3', 'o3', 'so2', 'aqi'].

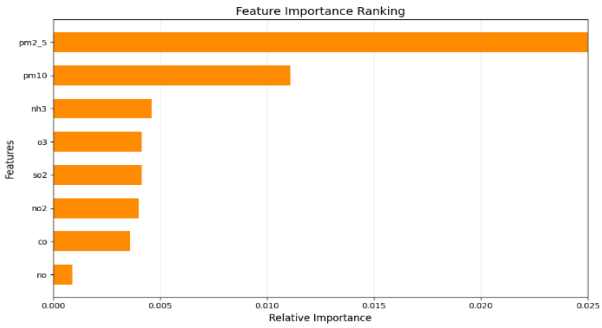
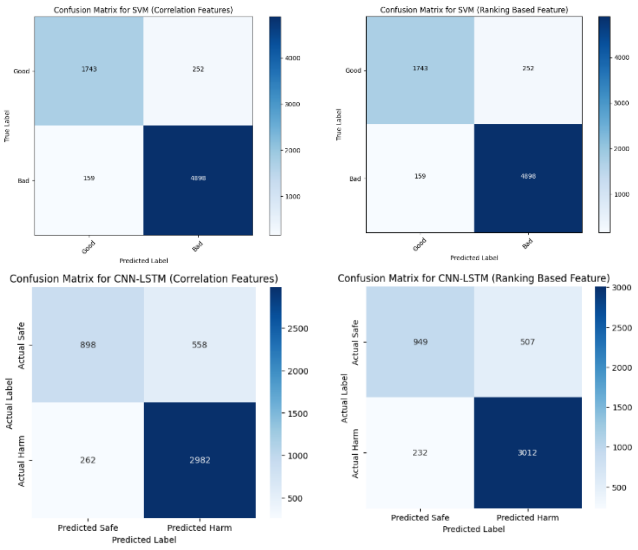


Fig. 4. Numerical Features Correlation Matrix

D. *Models Evulation*

With feature selection methods, two new datasets are formed to enhance the predictive power of the models. On the datasets, three machine learning models SVM, CNN-LSTM, and Transformer are executed with different parameter settings. Performance of each model is critically analyzed with various evaluation factors, such as confusion matrices, recall, precision, and F1-score. Such comprehensive analysis allowed us to compare systematically the strengths and weaknesses of each model on both the datasets.



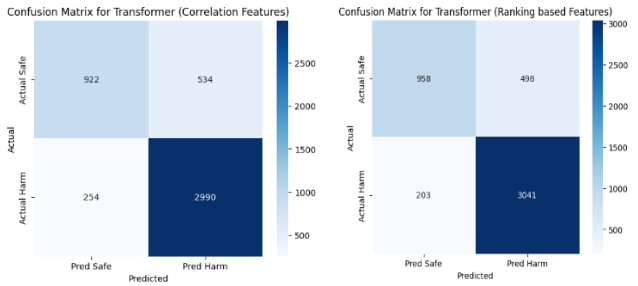


Fig. 5. Models Confusion Matrices

The figure clearly shows that SVM yields the highest number of correctly classified samples across both datasets with very few classifications that are inaccurate. The classification performance of CNN-LSTM and Transformer are also reasonable however their misclassification errors, as shown by the confusion matrices, seem to consist of more false negatives and false positives compared to SVM. In particular, CNN-LSTM appears to have a more difficult time identifying "Safe" cases correctly, whereas Transformer improves recall for harmful cases across false negatives but has slightly more misclassifications in safe cases.

In overall conclusion, SVM has shown the most balanced and reliable classification performance across both datasets, in accordance with the quantitative evaluation findings outlined above.

III. Results

The performance of the three classifiers (SVM, CNN-LSTM, Transformer) was investigated on two feature-engineered datasets (crr_reduced_data and pr_reduced_data). A comparison of four assessment metrics: F1-Score, Accuracy, Precision, and Recall are shown in a heatmap in [Fig. 6].

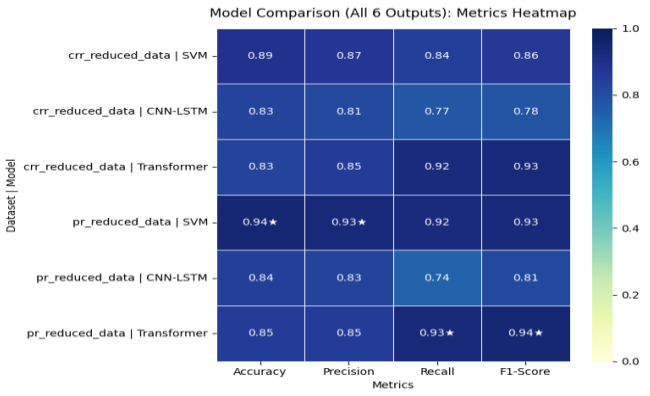


Fig. 6. Models Comparison Matrix

We applied three models (SVM, CNN-LSTM, and Transformer) on two newly created datasets and evaluated them using F1-Score, Accuracy, Precision, and Recall. The heatmap displays the results with starred values (★) identifying best scores across the metrics. The comparison showed SVM had the best overall performance on the ranking-based dataset, meaning that it was the best air quality prediction model of this study..

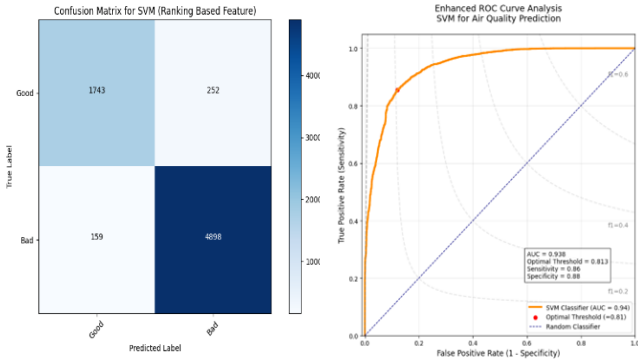


Fig. 7. SVM Confusion Matrix and ROC Curve

The Support Vector Machine model, applied on the ranking-based data set, provided exceptional prediction effectiveness. In Figure X (Confusion Matrix), it can be seen that the model accurately classified the overwhelming majority of samples, predicting 4,898 harmful cases and 1,743 safe cases, and only having a few false positives (252) and false negatives (159). This indicates balance in that SVM is sensitive to detect harmful air quality and has accurate specificity to not classify safe air quality as harmful quality.

The ROC curve in Figure Y also supports this performance with an Area Under the Curve (AUC) of 0.938, showing excellent discriminative power. The model at the best available threshold of 0.813 had a sensitivity of 0.86 and specificity of 0.88, showing good trade-off between accurate detection of harmful instances and false alarms to be minimized.

Overall, these findings make SVM the optimal and most stable model in this study. Its ability to maintain stability in accuracy across various measures of assessment makes it even better suited for stable air quality prediction and potential use in real-time monitoring systems.

IV. Discussion

This study illustrates the significance of both feature engineering and model selection in air quality tasks. Model stability, interpretability and performance improvements were demonstrated in this work by reducing features using correlation- and ranking-based methods. In terms of model performance, it was demonstrated that the three models tested, SVM, CNN LSTM, and Transformer, ranked in that order for the outcome measures performance (i.e., F1-score recall, accuracy, precision and AUC). SVM was 100% superior in terms of the ranking-based dataset which resulted in an AUC value of 0.938 for the ranking-based dataset.

The performance superiority of SVM can be attributed to the SVM's ability to manage high-dimensionality and reduce overfitting while, deep learning models showed less consistency in the performance ratings between the models indicating limited temporal complexity to the dataset. Overall, the findings from this study suggest that, for structured and engineered tabular datasets, SVM can outperform the more complex and deeper neural network architectures.

A. Limitations

There are several significant limitations with this study:

- Scope and Generalizability of the Datasets - The study was limited to two engineered datasets and validated only on historical data. Therefore, it is unknown if the models will be adaptable to real-time or cross-regional air quality monitoring.
- Constraints with Deep Learning - Due to being highly data-driven technology, CNN-LSTM and Transformer models may not have reached their upper threshold due to only limited size of the dataset used, and limited hyperparameter tuning that was relatively constrained.
- Feature Coverage – Only pollutant concentration data were included (i.e., the study did not include other variables that can affect air quality – e.g., weather, traffic impact, seasonal variations). A dataset with additional predictors would have enhanced robustness of predictions.

B. *Future Works*

- Consider Adding Broader Features - Future research should include additional attributes such as weather conditions, traffic counts, and seasonality to enhance the accuracy and reliability of air quality forecasting.
- Consider Further Feature Engineering - Future research could use more advanced feature selection and feature extraction methods, such as ensemble-based feature ranking, as well as hybrid correlation measures.
- Consider Real-Time and Inter-Regional Application - Future research could broaden the scope of the framework for real-time monitoring systems and employ it as a case study in other locations that validate its applicability and usefulness for policy makers and environmental agencies.

V. Conclusion

This research examined the effectiveness of feature engineering techniques and machine learning approaches for air quality predictions in historical multi-pollutant datasets. Two reduced datasets were created with correlation-based and ranking-based feature selection strategies, helping to increase the stability and interpretability of the model. There were three models - SVM, CNN-LSTM, and Transformer - that were evaluated, and in every iteration, the results indicated that the best model was always SVM. The SVM model achieved the highest score for accuracy, precision, recall, and F1-score, and an AUC of 0.938 from the ranking-based dataset and later demonstrated that its predecessor, SVM, is a valuable option to consider when modeling structured tabular datasets in which feature engineering is a significant contributor to the problem.Despite the framework being successful, dataset boundaries, limitations in features included, and overall generalizability of the results provide opportunities for improvement. Future work should investigate additional environmental variables, advanced features extraction techniques, and validation on real-time and cross-regional monitoring systems.

All-in-all, this research provides a repeatable machine learning ready framework to assist future work in air quality predictions and environmental monitoring efforts to support decision-making of public health.

References

[1] Panaite FA, Rus C, Leba M, Ionica AC, Windisch M. Enhancing Air-Quality Predictions on University Campuses: A Machine-Learning Approach to PM2.5 Forecasting at the University of Petroşani. *Sustainability*. 2024; 16(17):7854. https://doi.org/10.3390/su16177854

[2] Özüpak, Y., Alpsalaz, F. & Aslan, E. Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction. *Water Air Soil Pollut* **236**, 464 (2025). https://doi.org/10.1007/s11270-025-08122-8

[3] Makhdoomi, A., Sarkhosh, M. & Ziaei, S. $PM_{2.5}$ concentration prediction using machine learning algorithms: an approach to virtual monitoring stations. *Sci Rep* **15**, 8076 (2025). https://doi.org/10.1038/s41598-025-92019-3

[4] Sheetal Bawane, Priyanka Chaudhary, Sanmati Kumar Jain, Jitendra Singh Dodiya, "Forecasting of air quality index using Machine Learning and deep learning models", *J Neonatal Surg*, vol. 14, no. 18S, pp. 1147–1155, May 2025.

[5] Liu, H.; Li, Q.; Yu, D.; Gu, Y. Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms. *Appl. Sci.* **2019**, *9*, 4069. https://doi.org/10.3390/app9194069

[6] Gupta, N. Srinivasa, et al. "Prediction of air quality index using machine learning techniques: a comparative analysis." *Journal of Environmental and Public Health* 2023.1 (2023): 4916267. https://doi.org/10.1155/2023/4916267

[7] Anggraini, Tania Septi, et al. "Machine learning-based global air quality index development using remote sensing and ground-based stations." *Environmental Advances* 15 (2024): 100456. https://doi.org/10.1016/j.envadv.2023.100456

[8] Ravindiran, Gokulan, et al. "Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam." *Chemosphere* 338 (2023): 139518. https://doi.org/10.1016/j.chemosphere.2023.139518

[9] Maltare, Nilesh N., and Safvan Vahora. "Air Quality Index prediction using machine learning for Ahmedabad city." *Digital Chemical Engineering* 7 (2023): 100093. https://doi.org/10.1016/j.dche.2023.100093

[10] Feng, Yongliang. "Air Quality Prediction Model using deep learning in internet of things environmental monitoring system." *Mobile Information Systems* 2022.1 (2022): 722115. https://doi.org/10.1155/2022/7221157

[11] Chen, Bin. "Air quality index forecasting via deep dictionary learning." *IEICE TRANSACTIONS on Information and Systems* 103.5 (2020): 1118-1125. https://globals.ieice.org/en_transactions/information/10.1587/transinf.2019EDP7296/_p

[12] Kaggle Datasets for Air Pollution Data of India (2020-2023) https://www.kaggle.com/datasets/seshupavan/air-pollution-data-of-india-2020-2023

[13] Kelsy Cabello-Solorzano, O. de, M. A. Peña, L. Correia, and A. J. Tallón-Ballesteros, "The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis," *Lecture notes in networks and systems*, vol. 750, pp. 344–353, Jan. 2023, doi: https://doi.org/10.1007/978-3-031-42536-3_33.

[14] H. Azami, K. Mohammadi, and B. Bozorgtabar, "An Improved Signal Segmentation Using Moving Average and Savitzky-Golay Filter," *Journal of Signal and Information Processing*, vol. 03, no. 01, pp. 39–44, 2012, doi: https://doi.org/10.4236/jsip.2012.31006.

[15] "Feature Engineering for Machine Learning," *Google Books*, 2018. https://books.google.co.uk/books?hl=en&lr=&id=sthSDwAAQBAJ&oi=fnd&pg=PT14&dq=feature+engineering+machine+learning&ots=ZP-euY2mB0&sig=_oWgg_d7fWa_GgK-KBrms9h4o80&redir_esc=y#v=onepage&q=feature%20engineering%20machine%20learning&f=false (accessed Aug. 20, 2025).

[16] "DSpace," *Waikato.ac.nz*, 2024. https://researchcommons.waikato.ac.nz/entities/publication/12a40834-bf51-4d87-89ef-d00c2740f0d8

[17] A. Zien, N. Krämer, S. Sonnenburg, and G. Rätsch, "The Feature Importance Ranking Measure," *Machine Learning and Knowledge Discovery in Databases*, pp. 694–709, 2009, doi: https://doi.org/10.1007/978-3-642-04174-7_45