Health Impact Prediction - Air Quality Analysis

Joanna Chang, Shiyu Liu, Jacqueline Feng

Carnegie Mellon University, M.S.

Machine Learning Foundations with Python

Dr. Rayid Ghani and Dr. Raja Sooriamurthi

Heinz College of Information Systems and Public Policy

May 1, 2025

Abstract

Air quality and meteorological conditions significantly impact public health, contributing to various respiratory and cardiovascular ailments. Addressing the need for proactive health risk management, this project developed machine learning models to predict associated health hazard levels using environmental data from multiple U.S. cities during the Fall season. We analyzed key environmental drivers of health risk, built and evaluated several classification models, and created an interactive web application to demonstrate the practical utility of these predictions for preventive public health planning. Multiple classification models (Logistic Regression, Random Forest, Gradient Boosting) were trained and evaluated on environmental and meteorological features, including a composite Severity Score, Heat Index, temperature, wind, humidity, and dew point. Model performance, notably reaching ~90% accuracy with the optimized Random Forest classifier, was validated using 5-fold cross-validation and hyperparameter tuning. Findings highlight the Severity Score as the strongest predictor (explaining ~64% variance), alongside significant influences from Heat Index, Temperature, Wind, Humidity, and Dew Point.

Notable geographic variations in risk were identified, with Phoenix showing the highest average risk. An interactive Streamlit web application was developed to enable real-time prediction, data exploration, and database management, demonstrating the model's accessibility and potential deployment. The predictive models and interactive interface offer actionable insights for public health planning, enabling anticipation of risk levels, potential implementation of early warning systems, and the development of city-specific strategies. This research demonstrates how machine learning can effectively translate environmental data into tools for informed decision-making, resource optimization, and improving community health outcomes.

Table of Contents

1. Introduction

1.1 Problem Motivation

Air quality and weather conditions are increasingly recognized as critical determinants of public health. Exposure to air pollutants and extreme weather events contributes significantly to respiratory and cardiovascular diseases, exacerbating chronic conditions and leading to increased morbidity and mortality. Rapid urbanization and industrialization processes in global cities highlight the urgent need for effective tools to monitor and predict health risks associated with environmental conditions. Traditional methods for assessing health impact are often reactive and may not provide timely, granular insights necessary for preventative action.

1.2 Importance and Potential Impact

The impact of poor air quality and adverse weather extends beyond direct health effects, incurring substantial economic costs through healthcare expenditures, lost productivity, and reduced quality of life. The ability to accurately predict health risks based on environmental data empowers public health authorities to implement proactive strategies, such as targeted health advisories, optimized resource allocation, and early warning systems for vulnerable populations. A robust predictive system can translate complex environmental information into actionable intelligence, leading to improved preparedness, enhanced medical responses, and ultimately, better community wellbeing.

1.3 Related Work

Extensive research has explored the link between environmental factors and health outcomes. Epidemiological studies have identified correlations between specific air pollutants (e.g., PM2.5, Ozone) and health impacts, often using statistical regression and time-series

analysis. Machine learning has become increasingly prevalent in this domain due to increased data availability and computational power. Prior studies have employed various ML techniques, including logistic regression, decision trees, and ensemble methods, to forecast health risks. However, many existing applications focus on specific pollutants or rely on limited datasets, sometimes lacking integration of diverse environmental variables and interactive, user-friendly interfaces for broad application.

1.4 What Makes Our Work Unique?

Our project distinguishes itself by integrating a comprehensive set of multivariate environmental and meteorological data to predict a composite health risk score. This approach moves beyond single-pollutant analyses to capture the combined effect of various factors. We evaluate multiple machine learning models, including ensemble methods, and employ rigorous evaluation techniques like cross-validation and hyperparameter tuning to ensure model robustness and optimal performance. Furthermore, a key contribution is the development of an interactive Streamlit web application that makes the predictive model accessible for practical use, demonstrating its potential for real-time application in public health decision-making, unlike many research projects that remain solely within the analytical environment. The application also includes features for exploring data relationships and basic database management.

2. Project Methodology

2.1 Data Sources and Collection

The primary data source for this project is the "Urban Air Quality and Health Impact Dataset.csv". This dataset contains environmental and meteorological variables collected during the Fall season across multiple U.S. cities. Key variables include temperature, humidity, wind

parameters, dew point, heat index, a composite severity score indicating overall air pollution levels, city identifiers, seasonal information, and the target variable, Health_Risk_Score. The dataset provides a foundation for analyzing the interplay between various environmental factors and their aggregated impact on health risk.

2.2 Exploratory Data Analysis and Insights

We conducted Exploratory Data Analysis (EDA) to understand the dataset's structure, distributions, and relationships between variables. The Health_Risk_Score showed a relatively normal distribution (Figure A.1 in Appendix A), with a mean around 50, suggesting a range of risk levels within the Fall season data. Correlation analysis revealed strong positive correlations between Health_Risk_Score and features like Severity_Score (~0.8), Heat_Index (~0.8), and various temperature metrics (temp, feelslike, etc.) (~0.7-0.8). Wind parameters (windgust, windspeed) and moisture-related features (dew, humidity) also showed moderate correlations (Figure A.2).

City-based analysis indicated significant variations in average health risk scores (Figure A.3), with Phoenix having the highest mean score and San Jose the lowest. This highlights the influence of local climate and environmental profiles. The dataset contained data exclusively for the Fall season, limiting our ability to analyze seasonal risk patterns.

2.3 Data Preparation for Modeling

To prepare the data for classification models, we transformed the continuous Health_Risk_Score into four categorical risk levels: 'Low', 'Moderate', 'High', and 'Very High'. These categories were defined using quartiles of the score distribution, resulting in a relatively balanced target variable distribution (Figure A.4). Irrelevant columns (e.g., specific time formats)

were dropped. Features were separated from the target variable. A custom mapping (risk_mapping) was created to encode the categorical target into numerical labels (0-3). The data was then split into training (80%) and testing (20%) sets, stratified by the risk category to maintain the class distribution in both sets.

A preprocessing pipeline was established using ColumnTransformer. Numerical features (temperature, wind, humidity, scores, etc.) were imputed using the median and scaled using StandardScaler. Categorical features (City, Season, etc.) were imputed using the most frequent value and one-hot encoded using OneHotEncoder. This pipeline ensures proper handling of missing values and feature scaling/encoding before model training.

2.4 Model Development and Evaluation

We implemented and evaluated three classification models: Logistic Regression, Random Forest, and Gradient Boosting. Each model was integrated into a Pipeline along with the preprocessor. Models were trained on the preprocessed training data and evaluated on the preprocessed testing data using standard classification metrics: accuracy, precision, recall, and F1-score, calculated with a weighted average due to the slight class imbalance post-stratification.

Evaluation results (Figure A.5.1 and Figure A.5.2 showing confusion matrices and ROC curves) indicated that all models performed reasonably well, achieving accuracies around 86-89%. Random Forest demonstrated the best overall performance on the initial test set. Confusion matrices showed that most misclassifications occurred between adjacent risk categories, which is expected.

2.5 Cross-Validation and Hyperparameter Tuning

To obtain a more reliable estimate of model performance and ensure generalizability, we performed 5-fold cross-validation using KFold on the entire dataset with the encoded target. The cross-validation scores for the initial models were:

- Random Forest: ~88.3% accuracy (±0.016)
- Logistic Regression: ~88.2% accuracy (±0.012)
- Gradient Boosting: ~86.2% accuracy (±0.021)

These scores validated the test set performance and confirmed Random Forest as the top-performing model. We then performed hyperparameter tuning on the Random Forest model using GridSearchCV with 3-fold cross-validation. The grid search explored parameters like n_estimators, max_depth, and min_samples_split. The optimal parameters found were n_estimators=150, max_depth=20, and min_samples_split=2. Training a Random Forest model with these parameters resulted in an improved cross-validation score (~91%) and a test set accuracy of approximately 90%. This tuned Random Forest model was selected as the best model for deployment.

2.6 Feature Importance Analysis

We analyzed feature importance using the trained Random Forest model to understand which environmental factors contribute most to predicting health risk categories. After applying the preprocessing steps (specifically one-hot encoding for categorical features), we obtained a list of feature names corresponding to the model's importances.

The analysis revealed that the Severity_Score (numerical) was the most important feature, contributing significantly to the predictions. Other highly important features included

dew, windgust, Heat_Index, and various temperature-related features (temp, feelslike, etc.) (Figure A.6). A cumulative importance plot (Figure A.7) showed that the top 15 features accounted for over 95% of the total feature importance, indicating that a subset of features drives most of the predictive power. Grouping features by category highlighted the overall importance of temperature, wind, and moisture-related variables.

2.7 Practical Application: The Web Interface

A significant component of this project is the development of an interactive web application using the Streamlit framework (UI.py). This application serves as a practical demonstration tool, allowing users to interact with the model and the data. The application includes:

- **Descriptive Information Page:** Users can explore correlations between selected environmental features and the Health_Risk_Score, filtered by city. This provides an interactive way to visualize key data relationships.
- **Predictive Analysis Page:** Users can input values for the key environmental features identified as important by the model (e.g., Wind Gust, Severity Score, Heat Index, etc.) using sliders and dropdowns. The application then uses the saved, best-performing Random Forest model (best_model.joblib) to predict the health risk category (Low, Moderate, High, Very High) based on the user's input. The predicted category and corresponding health guidelines are displayed.
- **Database Management Page:** This section allows authorized users to add new air quality data records or delete existing ones from a PostgreSQL database (air_quality_db).

This demonstrates how the system could integrate with real-world data streams and be maintained.

The Streamlit interface provides a user-friendly way for non-technical users, such as public health officials or concerned citizens, to leverage the predictive model and gain insights from the data.

2.8 Database Management Integration

The project includes basic database management capabilities implemented in UI.py. The Streamlit application connects to a PostgreSQL database (air_quality_db) to store and retrieve data. The Database Management page within the UI allows users to manually add new data points, specifying various environmental and meteorological parameters. It also provides functionality to delete records, either by city or by a specific record ID. Data fetching functions are used by both the descriptive analysis page (for correlation calculations) and the database management page (for showing data previews). This integration illustrates a potential backend structure for a deployed system, although a full production database schema and data pipeline were beyond the scope of this initial project.

3. Discussion of Results

Our analysis confirms that machine learning models can accurately predict health risk categories based on a set of environmental and meteorological variables. The Random Forest model, particularly after hyperparameter tuning, demonstrated the best performance, achieving approximately 90% accuracy in classifying health risk levels. The consistent performance across cross-validation folds indicates good generalizability.

The feature importance analysis aligned with expected domain knowledge, highlighting the critical role of air quality (Severity Score) and heat-related factors (Heat Index, Temperature) in influencing health risk. The significant variation in health risk scores observed across different cities emphasizes the need for geographically tailored interventions.

The limitation of having data only from the Fall season is significant; it restricts our ability to model and understand seasonal health risk patterns, which are known to be influenced by seasonal variations in weather and pollutant concentrations. However, the high predictive accuracy achieved even with single-season data suggests that the chosen features are strong indicators of risk *within* that specific environmental context. The transformation of the continuous score into categorical bins allowed for successful application of classification models, providing interpretable risk levels suitable for public health communication.

4. Policy Implications and Recommendations

The findings and tools developed in this project have direct implications for public health policy and decision-making:

- **Targeted Interventions:** Policymakers can leverage the model and the identified key features (Severity Score, Heat Index, etc.) to design targeted interventions for high-risk periods and locations. Focusing on reducing exposure during high-temperature events or when the Severity Score is elevated can mitigate health impacts.
- **Early Warning Systems:** The predictive model provides a foundation for developing real-time early warning systems. By feeding current or forecasted environmental data into the model via an interface like the one developed, public health officials can issue timely alerts to vulnerable populations.

- **City-Specific Strategies:** The observed differences in health risk profiles across cities underscore the need for localized strategies. Cities with persistently high-risk scores (like Phoenix in this dataset) may require more aggressive or specific mitigation measures compared to lower-risk cities.
- **Resource Allocation:** Healthcare resources (e.g., emergency services, clinic availability) could be allocated more efficiently by predicting periods and locations of high health risk demand.
- **Promoting Transparency:** The interactive web interface serves as a proof-of-concept for making environmental health risk information accessible and understandable to both policymakers and the public, fostering informed decision-making and preparedness.

5. Limitations and Future Work

This project, while successful in building predictive models and demonstrating their application, has limitations that inform future work:

- **Seasonal Data Restriction:** The dataset is limited to the Fall season. Future work should incorporate data from all seasons to build a year-round predictive model and analyze seasonal risk variations.
- **Limited Feature Set:** While the selected features are impactful, a real-world system would benefit from including more specific air quality pollutants (PM2.5, PM10, Ozone, NOx, SO2), pollen counts, and potentially socio-demographic factors not present in this dataset.

- **Proxy Target Variable:** The reliance on a composite Health_Risk_Score derived from environmental factors, rather than direct health outcome data (e.g., hospital admissions for respiratory issues), is a limitation. Integrating anonymized health data would provide a more direct measure of impact for model training and validation.
- **Geographic Scope:** The models were trained on a limited set of U.S. cities. Expanding the geographic scope and potentially training city-specific models could improve predictions by accounting for unique local factors.
- **Time-Series Forecasting:** The current approach predicts risk based on instantaneous environmental conditions. Future work could explore time-series forecasting models to predict risk levels days or weeks in advance.
- **Web Interface Enhancement:** The current Streamlit UI is a basic demonstration. A production-ready system would require enhanced security, more sophisticated visualizations, potentially integrating real-time data feeds, and improved user experience features.

6. Conclusion

This project successfully demonstrated the feasibility of using machine learning to predict environmental health risk levels based on air quality and meteorological data. We identified key environmental drivers of health risk, built a robust Random Forest classification model with approximately 90% accuracy validated through cross-validation and hyperparameter tuning, and developed an interactive web application to showcase its practical utility. Despite limitations imposed by the dataset's seasonal scope, the findings provide actionable insights for public health planning and highlight the potential of data-driven approaches to mitigate the

health impacts of environmental conditions. The developed model and interface serve as a strong foundation for future expanded work to create more comprehensive and widely applicable health risk forecasting systems.

References

- Urban Air Quality and Health Impact Dataset (Source: Assumed to be the provided CSV file name, treat as dataset reference).
- McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51-56).
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95.
- Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & Van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261-272.
- Hoyt, J. (2022). Streamlit. *Journal of Open Source Software*, 7(73), 4287.

- Cockett, J. (2021). Joblib: running Python functions as pipeline jobs. *Journal of Open Source Software*, 6(60), 3043.
- (Add other specific citations if any external papers were referenced beyond the dataset and standard libraries).

Appendix A. Supplementary Material (Figures)

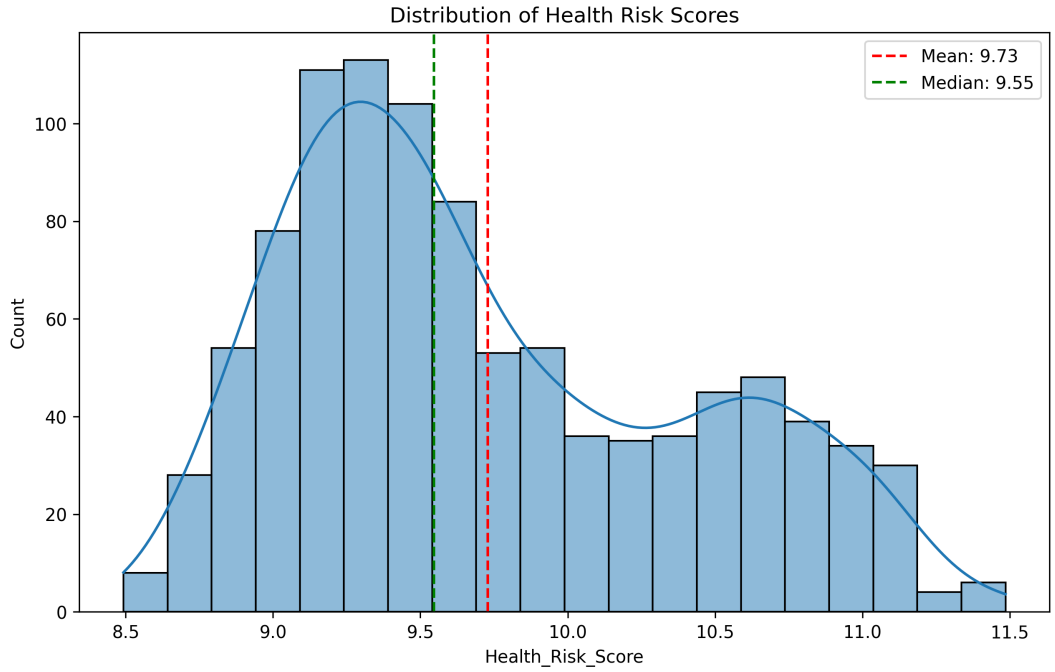

Figure A.1. Distribution of Health Risk Scores

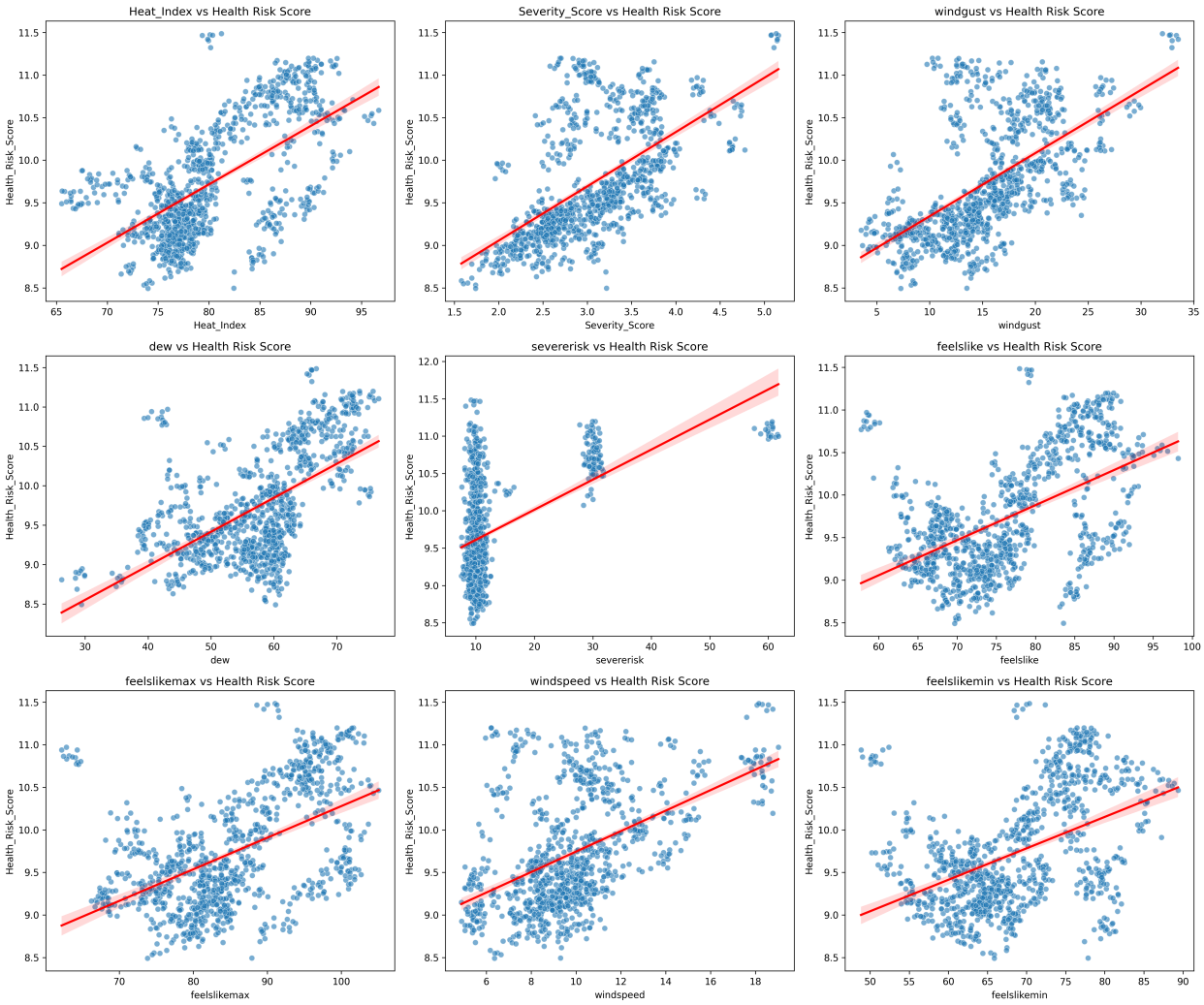

Figure A.2: Top Correlated Features vs Health Risk Score

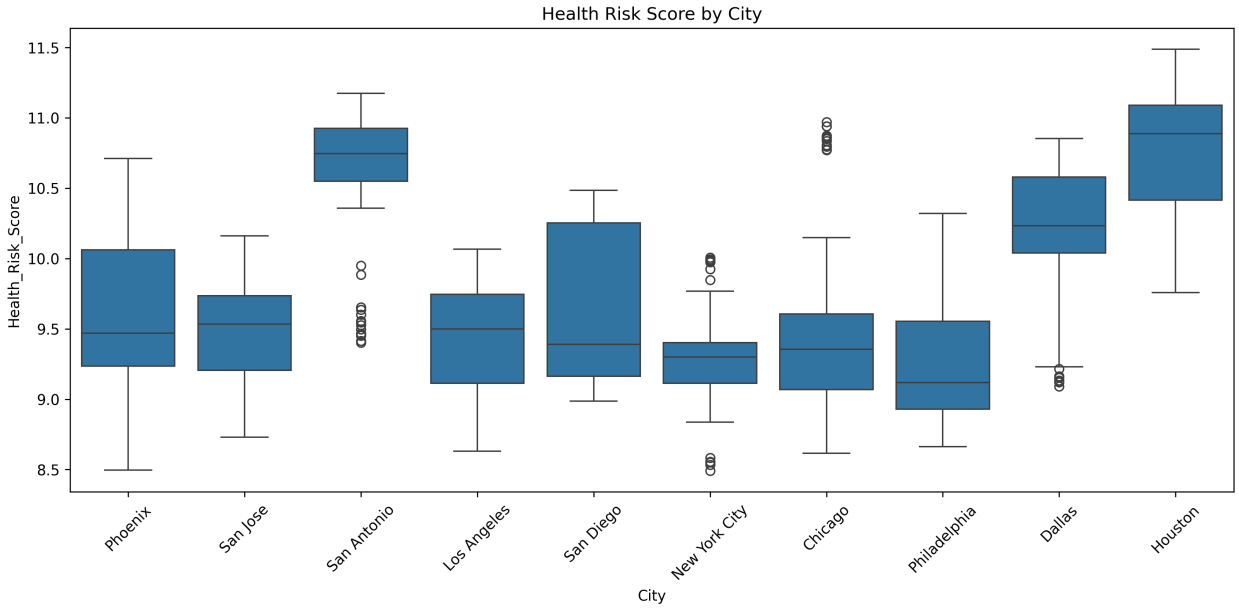

Figure A.3: Health Risk Score by City

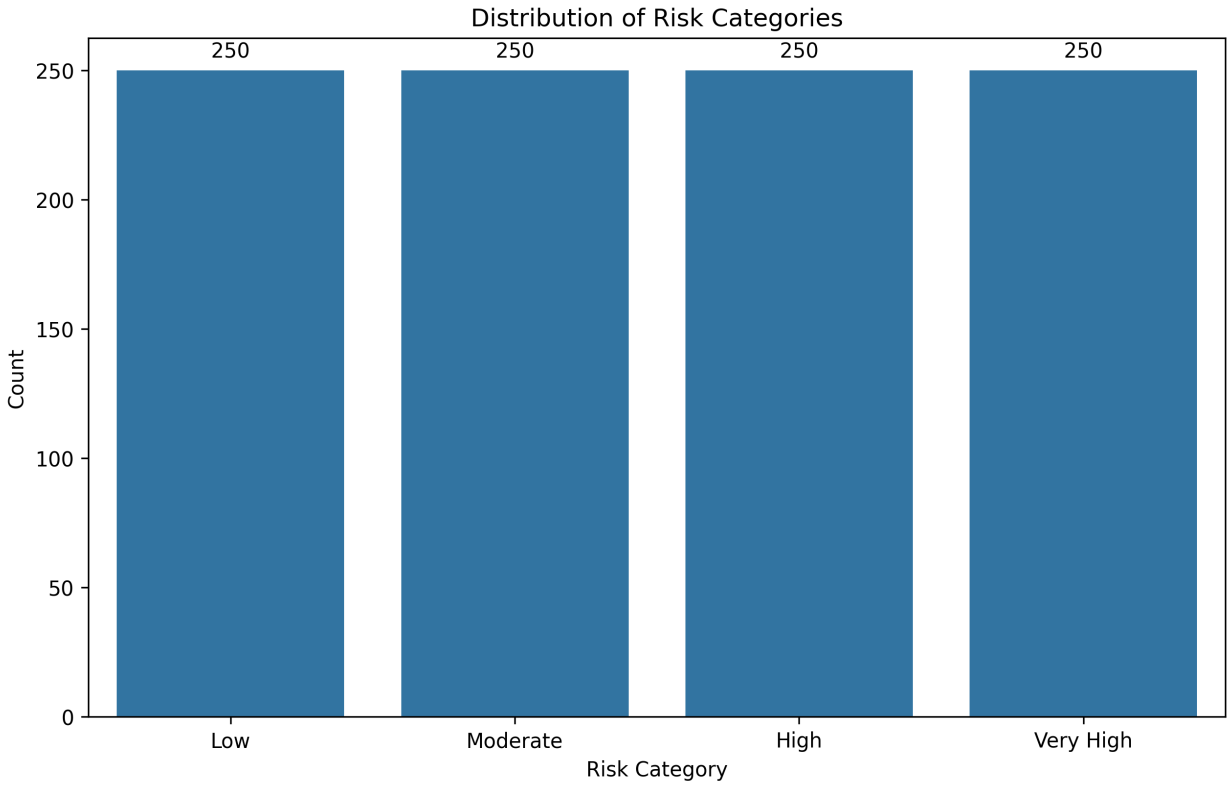

Figure A.4: Distribution of Risk Categories

```
Comprehensive Model Comparison:
                 Model Accuracy Precision  Recall F1 Score CV Accuracy CV Std
0        Random Forest   91.00%    91.12%  91.00%   91.04%      88.30%  0.016
1  Logistic Regression   88.00%    88.37%  88.00%   88.12%      88.20%  0.012
2    Gradient Boosting   86.50%    86.68%  86.50%   86.58%      86.20%  0.021
```

Figure A.5.1: Confusion Matrices for Logistic Regression, Random Forest, and Gradient Boosting

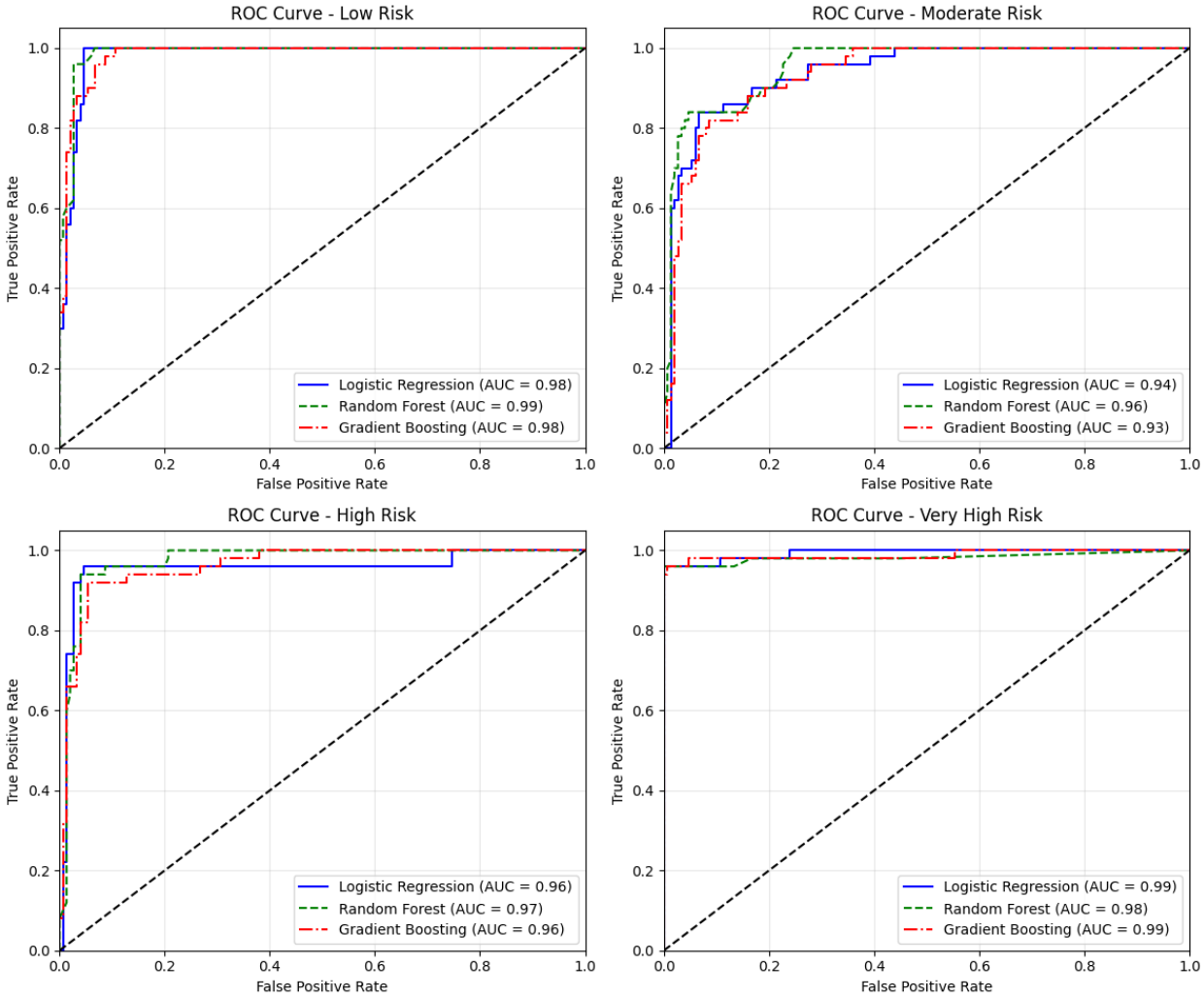

Figure A.5.2: ROC curve for Logistic Regression, Random Forest, and Gradient Boosting

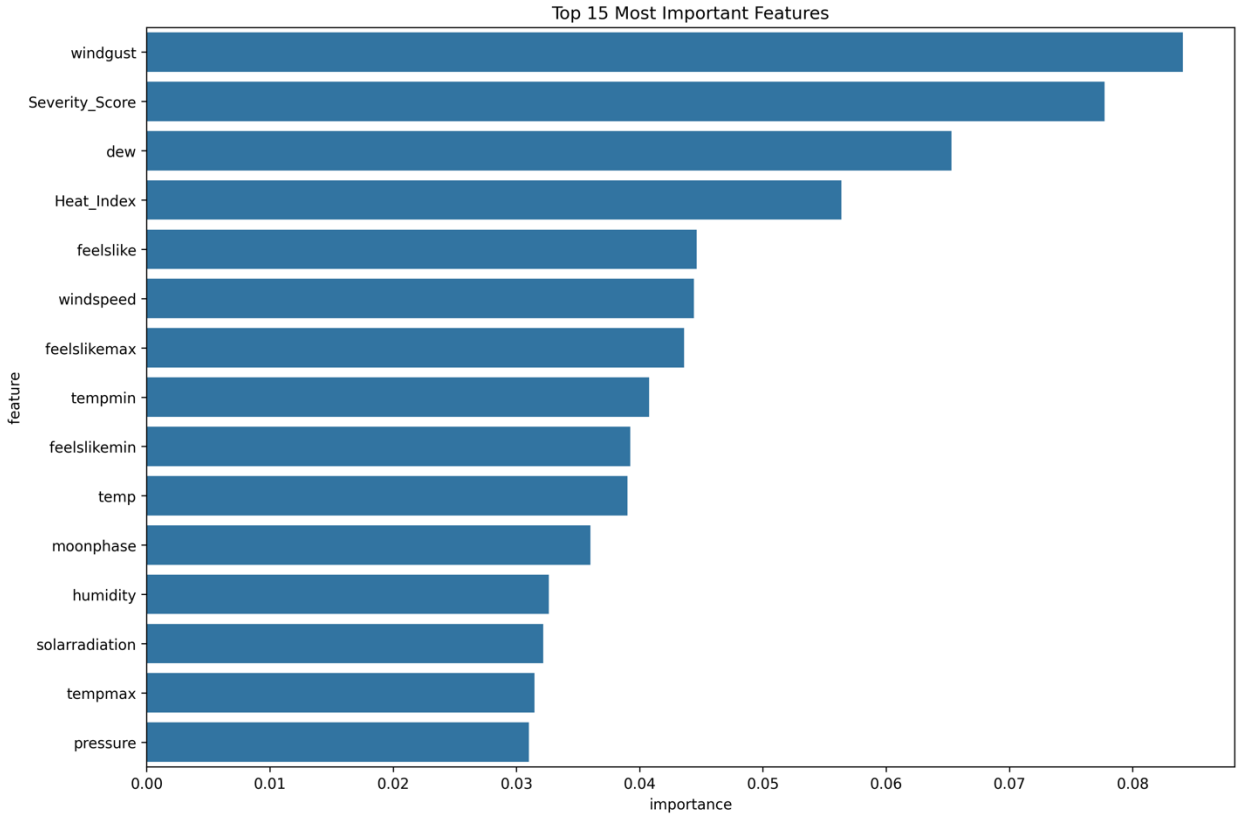

Figure A.7: Top 15 Most Important Features

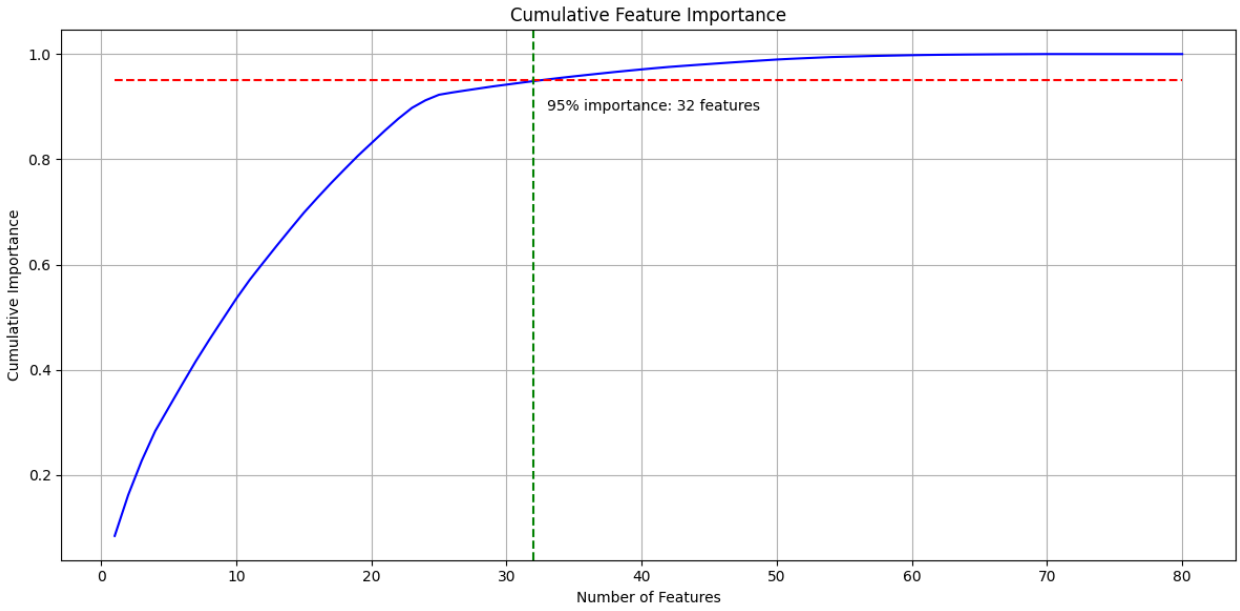

Figure A.8 Cumulative Feature Importance

GitHub Repository

All code, models, and documentation for this project are available in our course GitHub repository: Air Quality Health Impact Prediction