**ARTIFICIAL INTELLIGENCE OPTIMIZATION FOR COMPUTER SCIENTISTS — D682**

TASK 1: IMPLEMENTING AND TESTING AI SOLUTIONS

**B1: Identification of Suitable Algorithms**

For the urban air‑quality and health‑risk prediction problem, three candidate algorithms were selected and assessed for their fit, strengths, and limitations:

1. **Linear Regression**  
   Linear Regression fits a straight line (or hyperplane) to the data by minimizing the sum of squared errors. As a baseline, it offers complete transparency; each coefficient directly indicates the influence of a feature (e.g., PM₂.₅ or humidity) on the health risk score. However, its assumption of linearity makes it prone to underfitting when relationships between weather and pollution variables are complex or non‑linear.
2. **Random Forest Regressor**  
   Random Forest builds an ensemble of decision trees on bootstrapped samples, then averages their predictions to reduce variance. It naturally captures non‑linear interactions among multiple features (e.g., wind speed amplifying pollutant dispersion under certain humidity conditions) and is robust to outliers and missing values. The trade‑off is that it produces a large “forest” of trees that can be difficult to interpret and may require more memory at prediction time.
3. **Gradient Boosting Machine (GBM)**  
   GBM trains decision trees sequentially, each new tree correcting the errors of the combined prior ensemble by optimizing a specified loss function. This stage‑wise approach excels at modeling subtle, non‑linear dependencies in tabular data and includes built‑in regularization controls (learning rate, tree depth, subsampling) to mitigate overfitting. The downside is longer training times and sensitivity to hyperparameters, necessitating systematic tuning for optimal performance.

**B2: Description of Selected Algorithm – Gradient Boosting Machine**

**B2A. Relation to the Specified Optimization Problem**

Gradient Boosting Machine (GBM) is well‑suited to forecasting urban air quality and predicting health risks because it:

* **Models Complex Interactions:** GBM captures non‑linear effects (e.g., how a spike in PM₂.₅ combined with high humidity sharply increases health risk).
* **Iterative Error Correction:** Its sequential tree‑building aligns with the need for iterative refinement, improving predictions by focusing on previous errors.

**B2B. Strengths of GBM in This Context**

1. **High Predictive Accuracy:** By incrementally correcting mistakes, GBM often outperforms simpler models on noisy, multi‑feature datasets.
2. **Regularization Mechanisms:** Parameters like learning rate and max tree depth allow precise control over model complexity, reducing the risk of overfitting to synthetic or incomplete data.

**B2C. Potential Limitations of GBM**

1. **Computational Cost:** Sequential tree training can be time‑ and resource‑intensive, especially with large datasets or high tree counts.
2. **Hyperparameter Sensitivity:** Performance depends heavily on tuning several parameters (e.g., number of estimators, learning rate); poor choices can lead to under‑ or over‑fitting.

**C1. Implementation of the Selected Algorithm (GBM)**

|  |  |
| --- | --- |
| **Item** | **Details** |
| Programming language | Python 3.13 |
| Frameworks /Libraries | pandas, numpy, scikit‑learn 1.6.1, joblib |
| Process & Flow (commented in code) | 1. Load CSV (clean & sort by datetimeEpoch) 2. Chronological **train–test split** to avoid leakage  3. ColumnTransformer (numeric scaling + categorical one‑hot) 4. GradientBoostingRegressor wrapped in a Pipeline 5. GridSearchCV with **time‑series CV** (5 folds) for tuning  6. Hold‑out evaluation (RMSE, R²)  7. Persist best model to models/gbm.pkl |
| Key hyperparameters searched | n\_estimators ∈ {200, 400}; learning\_rate ∈ {0.05, 0.1}; max\_depth ∈ {3, 4}; subsample ∈ {0.8, 1.0} |
| Evaluation printout | Test RMSE: 0.196 Test R²: 0.931 |

**C2. Submission of Functional Code to GitLab**

|  |  |
| --- | --- |
| **Item** | **Details** |
| Repository URL | https://gitlab.com/wgu-gitlab-environment/student-repos/cetoeiro/d682-ai-optimization-for-computer-scientists.git |
| Branch used | working\_branch |
| Key commits | “C1/C2: functional GBM implementation and initial results” |
| Branch‑history file | branch\_history.txt |

**D1. Evaluation Metrics Selected**

|  |  |
| --- | --- |
| **Metric** | **Reason for Selection** |
| Root‑Mean‑Squared Error (RMSE) | Standard measure of average prediction error for continuous targets; penalizes large errors more heavily than MAE. |
| Coefficient of Determination (R²) | Indicates proportion of variance in healthRiskScore explained by the model; easy to interpret (0–1 scale). |

**D2. Implementation Results (from gbm\_model.py)**

|  |  |
| --- | --- |
| **Metric (Test Set)** | **Value** |
| RMSE | 0.196 |
| R² | 0.931 |

*Printed by the script after grid‑search tuning; also logged in console output.*

**D3. Analysis of Metric Results**

1. **Low Error Magnitude:** RMSE 0.196 means, on average, predictions deviate from actual health‑risk scores by < 0.2 units on the target’s scale, indicating high numerical accuracy.
2. **High Explained Variance:** R² 0.931 shows 93 % of variance in the health‑risk score is captured, leaving only 7 % unexplained noise; excellent for environmental data, which is often noisy.
3. **Consistency with Training Validation:** Cross‑validation RMSE values during grid‑search were within ±0.01 of the hold‑out RMSE, suggesting minimal overfitting.

**D4. Areas for Improvement**

|  |  |  |
| --- | --- | --- |
| **Improvement Area** | **Rationale** | **Planned Action** |
| |  | | --- | | **Add Pollution Features** |  |  | | --- | |  |  |  | | --- | |  | | Dataset currently lacks actual PM and NO₂ values; model relies heavily on weather proxies. | Integrate real sensor data or impute realistic pollution levels to boost explanatory power. |
| **Advanced Ensemble Techniques** | |  |  | | --- | --- | |  | GBM is strong but may plateau. | | Experiment with Extreme Gradient Boosting (XGBoost) and stacked ensembles to potentially reduce RMSE below 0.15. |
| **Model Interpretability** | |  |  | | --- | --- | |  | GBM predictions are accurate but opaque. | | Use SHAP values to identify which weather variables drive risk spikes, aiding public‑health decision‑making. |