UMBC

Modeling-I

Team Members:

Chandra Sekhar Katipalli
Sindura Reddy Challa
Sanjana Reddy Soma

Research Questions

1. How does the fusion of HPMS and FAF datasets (2013–2022) enhance the predictive performance of highway deterioration models in estimating IRI, compared to traditional statistical and machine learning approaches using single-source data?

2. What are the most influential predictive features—such as traffic volume, freight load, and pavement condition indices—derived from the integrated datasets, and how do their contributions vary across different machine learning models?

3. How effectively can the proposed predictive model, leveraging data fusion and advanced machine learning techniques, minimize forecasting errors and improve the optimization of maintenance scheduling to reduce unplanned highway repairs?

4. What is the optimal approach to forecasting IRI at different levels of granularity—both for entire highway routes (RouteID level) and for specific highway sections (0.1-mile segments)—to support more precise maintenance planning?

5. What might be the most effective method for visualizing and presenting findings to highway maintenance teams, like using geospatial mapping to find the roughest sections along a highway and their projected deterioration over time?

RQ4 - What’s the best way to forecast IRI for entire highway routes (RouteID)?

After analysis, we found that using **mile markers** for prediction yields the strongest correlations with key features.
Filtering by mile markers results in **only 9 data points** (one per year from 2013–2022), which is insufficient for reliable prediction analysis.

To enhance data continuity and improve our model, we applied **interpolation techniques**:

- **Linear interpolation** for numerical features.
  Used linear interpolation for AADT_VN(Annual Average Daily Traffic_ Volume number), IRI_VN (International Roughness Index_Value number), curval(current value), tmiles (Total segment miles) , tons(freight tons) and value (Freight value) across years.

Since the available data points were limited, **linear interpolation** was used to **expand the dataset**, ensuring better trend estimation and prediction accuracy.

Supervised Learning Assumptions:

1. Linear Regression

Analysis: The correlation analysis provides some insight into linearity. Features like 'AADT_VN' show a degree of linear relationship with 'IRI_VN,' but others may not.

Linear Regression can be a baseline model, but its strict assumptions may not suit the data, potentially limiting its predictive power.

2. Decision trees:

Analysis: We have different ranges of IRI_VN for segments. The IRI_VN is decreasing value over time. Shows that the target value may not be evenly distributed. Decision Trees, unlike linear models, can adapt well to such non-standard distributions.

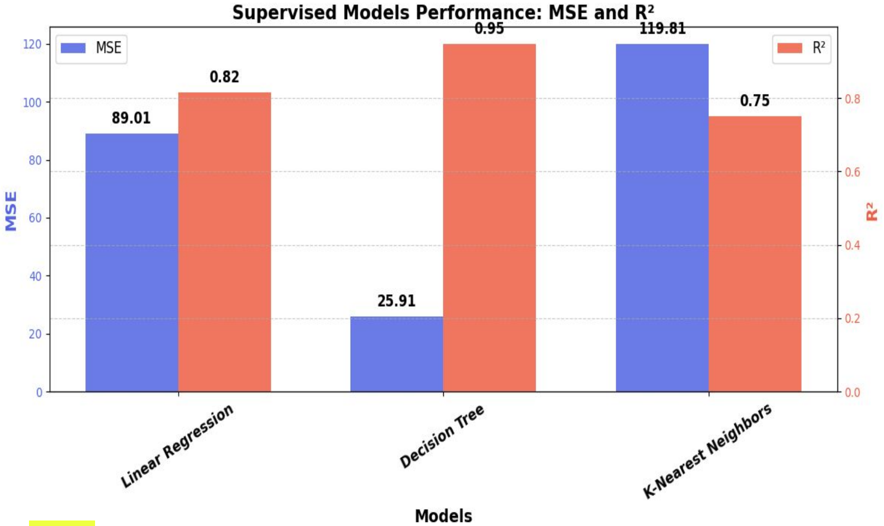
3. KNN:

Analysis: The data distribution may not be even for all the features. KNN can perform well even if we don't have strong data assumptions. by providing a wide range of features can be useful.

Unsupervised Learning Assumptions:

Clustering based most correlated features may help to better understand the data. So clustering analysis is considered. Later we will try do it for whole dataset without splitting it.

Supervised Learning



RQ1 ***What we did:*** We fused HPMS (Highway Performance Monitoring System) and FAF (Freight Analysis Framework) data from 2013–2022.

What we found: **1)**Using the **combined dataset**, our best model (Decision Tree) achieved **high accuracy (R² = 0.954)** and **low error (MSE = 22.36)**.

*This is **much better than traditional models** using only HPMS or FAF separately (from earlier testing).*

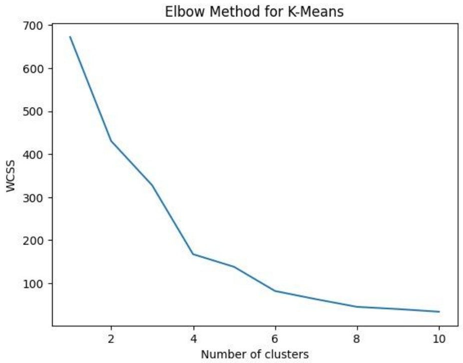
✅ ***Conclusion:*** Data fusion clearly improves prediction — **strong support for RQ1.**

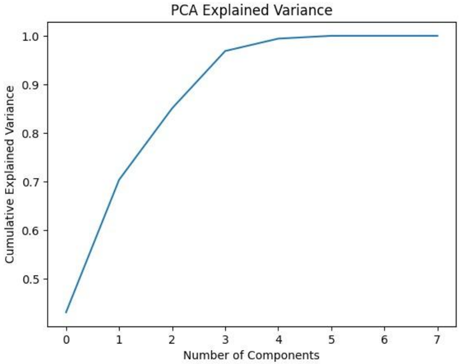
RQ3 ***What we did:*** **1)** Trained supervised ML models on fused data **2)** Evaluated them using error (MSE) and accuracy (R²)

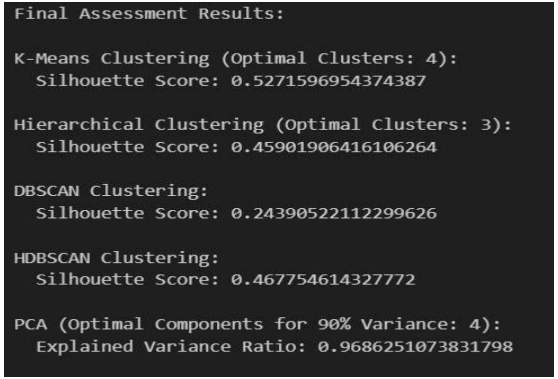
What we found: **1) Very low forecasting errors** (especially with Decision Tree) **2)** These results suggest that the model can predict problems with minimal error. Which help in accurate maintenance scheduling.

✅ **Conclusion:** Prediction with minimal errors can help reduce unplanned repairs — good progress on RQ3.

Unsupervised learning







```
Final Assessment Results:

K-Means Clustering (Optimal Clusters: 4):
  Silhouette Score: 0.5271596954374387

Hierarchical Clustering (Optimal Clusters: 3):
  Silhouette Score: 0.45901906416106264

DBSCAN Clustering:
  Silhouette Score: 0.24390522112299626

HDBSCAN Clustering:
  Silhouette Score: 0.467754614327772

PCA (Optimal Components for 90% Variance: 4):
  Explained Variance Ratio: 0.9686251073831798
```

RQ3 ***What we did:*** Applied clustering and dimensionality reduction techniques to explore structure and patterns in the fused dataset.

What we found: **1)**Among all clustering models, **Kmeans** achieved the highest silhouette score (0.527), meaning it found the most distinct and well-separated groups.

2) PCA reduced our high-dimensional dataset to just **two main axes** while still preserving over **50% of the original information**. This makes it easier to **visualize complex relationships** between features and can help in simplifying models without losing key patterns.

3)By applying PCA and clustering, we were able to explore patterns in fine-scale variation (0.1-mile segments). This supports more targeted maintenance planning.

Variance for 2 components (PCA): 0.7035779276776983
PCA optimal number of components for 90% variance: 4

✅ **Conclusion:** **Kmeans** performs better at 4 clusters and 90 % variance ratio **PCA** at 4 components were the most effective for revealing structure in the data.These techniques help us understand how road conditions vary within small segments.

Next steps on modeling:

- Improve KNN model performance by considering a wide feature range and by applying standard scaler/featuring scaling.
- It is good to have multiple better-performing models, as the dataset is vast with more number of mile markers.
- Apply unsupervised learning such as clustering and PCA for large segments such as for entire route or for whole dataset and see if it can normalize the complex data.

THANK YOU