#### MET CS 777

#### Name : Surya Gopalakrishnan

#### Project Submitted : 12/Oct/2025

#### Introduction

**Research Question: How well can we estimate state healthcare performance rates using minimal input variables?**

Accurate estimates of state healthcare performance rates can help policymakers identify underperforming areas, allocate resources efficiently, and monitor national healthcare progress. This project investigates how well we can estimate state healthcare performance using a minimal set of input variables.

#### Data Set

**My dataset source:** <https://healthdata.gov/CMS/2024-Child-and-Adult-Health-Care-Quality-Measures-/pahp-sncm/about_data>

**2024 Child and Adult Health Care Quality Measures Quality**

**Columns I’ve Considered:** State, Domain, Reporting Program, Measure Abbreviation, Measure Type, Population, Methodology, Core Set Year, State Rate.

**Train/Test Split:**

**There are about 11,000 rows of data in the dataset. After cleansing the Null values in “State Rate” column the final datasets were,**

* Train dataset rows: 8945 → 6981 after dropping NULLs
* Test dataset rows: 2170 → 1646 after dropping NULLs

#### Machine Learning Model

Lasso Regression (implemented as LinearRegression() in Spark with elasticNetParam=1.0) was selected because it performs **automatic feature selection** by shrinking less important coefficients to zero, which is **valuable for high-dimensional datasets** with multiple categorical variables. The regParam parameter controls the strength of regularization.

**Encoding used:**

* Categorical features → StringIndexer + OneHotEncoder
* Numeric features → cleansed and directly included (Core Set Year)
* Combined via VectorAssembler → final features vector

To prepare categorical variables for the regression model, each categorical feature (such as *State* or *Measure Type*) was first transformed using **StringIndexer**, which assigns each category a unique numeric index (e.g., “Texas” → 1, “California” → 2).

These indexed values were then passed to the **OneHotEncoder**, which converts them into binary indicator vectors (e.g., “Texas” → [1, 0, 0], “California” → [0, 1, 0]). This ensures that the model treats categories as distinct and non-ordinal, preventing misleading assumptions about numerical relationships between them.

**VectorAssembler** will assemble all values together, so that one row is now one numeric vector.

#### Evaluation

1. **R² is fairly stable** around ~0.53–0.535 for smaller regularization (regParam=0.01–0.2).
   * Indicates that ~53% of the variance in State Rate is explained by the features.
   * Performance slightly decreases at regularization 0.5 means too much regularization shrinks coefficients excessively.
2. **Root Mean Squared Error trends**:
   * Slight improvement at regularization 0.1 means minimal prediction error (~15.38).
   * Higher regularization (0.5) increases RMSE means model underfits.

Therefore, lower regularization gives better fit.

**Best parameter:** regParam = 0.1

* R² = 0.5352 → highest explained variance
* RMSE = 15.378 → lowest prediction error

R² value of 0.54 is solid for a real world dataset because healthcare performance is often influenced by other factors such as policy changes, local demographics. Capturing just over half of the variability with a small set of input features is a meaningful result in real-world applications.

**Limitations:**

* The model **assumes linear and additive relationships between features and outcomes**. As a result, it may not fully capture nonlinear patterns or interactions. Additionally, the dataset omits key factors such as patient demographics, state healthcare funding, and infrastructure quality, which may limit predictive performance.

#### How Lasso has highlighted necessary features



From this output, we can see the model has retained only a subset of features that positively or negatively impact the model. The **magnitude** shows relative importance (larger → stronger influence).

E.g.: The encoded State category has the strongest impact. That means certain states naturally have higher performance rates compared to others—just being in that state “adds” to the predicted rate.

Whereas, certain Measure Types decrease the predicted rate.

#### How do I interpret the results

About half of the variability in state healthcare rates can be explained using the features we have in a **simple, interpretable model**.

The other half is **unexplained**—either due to missing features or complex non-linear relationships.

#### Some Future Work:

**Future work could explore nonlinear models or include additional features such as patient demographics and state healthcare spending to improve predictive power.**

**These enhancements could lead to a more complete understanding of what drives healthcare performance variation across states.**