Assessing the Impact of Hospital Ratings and Conditions on Risk-Adjusted Mortality: Evidence from California Hospitals in 2023

Hilda Joseph [1] **Caroline Ngyuen** [1] **Eddie Zhang** [1]

1. Abstract

This project investigated how hospital quality ratings and patient medical conditions relate to risk-adjusted mortality outcomes across California hospitals in 2023. The goal was to determine whether widely reported quality indicators meaningfully predict mortality outcomes and to evaluate how different modeling approaches capture these relationships. Understanding these relationships is critical for improving patient care, informing policy decisions, and helping hospitals identify areas for quality improvement. After compiling statewide hospital performance data, we conducted extensive cleaning to address missing values, inconsistent numeric formats, low-volume instability, and suppressed entries. This cleaning process ensured that the dataset was as accurate and reliable as possible, minimizing potential biases that could arise from data inconsistencies. We standardized variables, converted text-based rates into numeric fields, flagged outliers, and encoded hospital ratings and medical conditions to create a model-ready dataset. Each step was carefully documented to maintain transparency and reproducibility. Three predictive models were implemented: linear regression for baseline interpretability, decision trees to capture non-linear interactions, and K-Nearest Neighbors (KNN) regression to estimate outcomes using local similarity among hospitals. These models were chosen to provide complementary perspectives on the data, allowing us to compare simple, interpretable methods with more flexible, pattern-based approaches. Models were trained on a train/test split and evaluated using R^2, RMSE, MAE, and cross-validation, with tuning applied through decision-tree pruning and varying k-values for KNN. Decision tree regression was used as a model aimed to capture the continuous nature of the variables. The tree was limited, though, with regards to the depth of the model. Across all methods, hospital rating and condition type emerged as meaningful predictors of mortality, though substantial unexplained variance remained. Linear regression revealed clear directional trends, decision trees highlighted interaction effects, and KNN captured localized patterns but was sensitive to noise. Overall, the results show that while rating and condition provide important information about mortality risk, they do not fully account for differences in outcomes across hospitals, indicating that additional unmeasured clinical or structural factors likely influence hospital performance. These findings highlight the need for more comprehensive data collection and modeling strategies to better understand the drivers of hospital quality and patient outcomes.

	RAMR
count	2135.000000
mean	4.196862
std	4.904603
min	0.000000
25%	1.200000
50%	3.400000
75%	5.900000
max	93.200000

Figure 1. Basic Statistics Table of the Data (RAMR = Risk Adjusted Mortality Rate

2. Data

2.1. Data

The dataset is from data.gov, and refers to a California set about hospital mortality rates for certain conditions and their quality ratings. The data was publicly accessible, and the idea of the paper was established due to the authors' interest in the healthcare field.

Dataset
https://catalog.data.gov/dataset/california-hospital-inpatient-mortality-rates-and-quality-ratings-8e21f

2.1.1. Exploratory Data Analysis

The summary statistics for the Risk-Adjusted Mortality Rate (RAMR) show that there is quite a bit of variation across the 2,135 hospital–condition entries in our dataset. The average RAMR is 4.20, and the median is slightly lower at 3.40, which suggests the data is somewhat right-skewed

because of a few very high mortality values. Most hospitals fall within a lower mortality range, since the middle 50% of values (1.20 to 5.90) are relatively close together. However, the maximum RAMR of 93.2 stands out as an extreme outlier, likely tied to rare or very high-risk conditions. Overall, these statistics show that while most hospitals have similar mortality rates, there are a few cases that significantly raise the upper end of the range.

The three graphs provide an initial overview of how hospital ratings, medical conditions, and risk-adjusted mortality rates (RAMR) relate across Californian hospitals in 2023. The first graph, a boxplot of RAMR by hospital rating, illustrates notable differences in mortality performance: hospitals rated worse exhibit higher median and upper-range RAMR values, while those rated better show lower and more compressed mortality ranges, suggesting more consistent outcomes (Figure 2). The second graph displays the frequency of each medical condition in the dataset, revealing that conditions such as pneumonia, heart failure, and acute stroke occur most frequently, while more specialized surgical procedures appear less common (Figure 3). The final scatterplot shows the distribution of RAMR across conditions, indicating considerable variability in mortality rates within and between conditions (Figure 4). These visualizations highlight meaningful patterns in hospital performance and suggest that both condition type and hospital rating may play important roles in explaining differences in risk-adjusted mortality outcomes.

2.2. Key Variables

Hospital, rating, condition, and risk-adjusted mortality rate

The hospitals refer to various hospitals in California. The ratings of the hospitals are either 'worse,' 'better,' or 'as expected' than normal. The condition is the disease the person in the case has. And the risk adjusted mortality rate is the hospital's observed death rate to its expected death rate after accounting for patient-specific risk factors like age. **Data Overview Feature and Target Variables:**

To analyze how hospital characteristics relate to patient outcomes, we structured the dataset into clearly defined predictors and outcomes: **Feature Variables (Predictors)**: **Hospital Rating (categorical)**: Better, Average, Worse **Medical Condition (categorical):** Includes pneumonia, heart failure, acute stroke types, hip fracture, carotid endarterectomy, PCI, pancreatic conditions, and others Hospital ID (optional control) **Target Variable (Outcomes): Risk Adjusted Mortality Rate (RAMR):** A continuous measure of mortality per 100 cases, adjusted for clinical risk. These variables allow both categorical and numeric modeling, making the dataset suitable for linear models, trees, and distance-based methods.

	RAMR
count	2135.000000
mean	4.196862
std	4.904603
min	0.000000
25%	1.200000
50%	3.400000
75%	5.900000
max	93.200000

Figure 2. Sample Table of Key Variables

2.3. Research Question

For each Californian hospital in 2023, how are the hospital ratings and medical conditions associated with the risk adjusted mortality rate and the proportion of deaths out of total cases?

2.4. Scope

The data focuses on Californian hospitals in 2023 because it represents the most recent year available in the dataset. Using the latest year ensures that our analysis reflects the most up-to-date information on hospital performance, operations, and outcomes, making the findings more relevant to our research questions. The dataset includes medical conditions such as: AAA Repair Unruptured, AMI Acute Stroke, Acute Stroke Hemorrhagic, Acute Stroke Ischemic, Acute Stroke Subarachnoid, Carotid Endarterectomy, Esophageal Resection, GI Hemorrhage, Heart Failure, Hip Fracture, PCI Pancreatic Cancer, Pancreatic Other, Pancreatic Resection, Pneumonia.

2.5. Challenges in Reading, Cleaning, and Preparing the Data

Since we are working with real-world administrative or performance data, it comes with multiple challenges. Below are key issues and ideas for how to address them:

Small Sample Sizes and Statistical Stability

Some hospitals may have very low case volume for certain conditions or procedures in some years. It is possible that mortality rates derived from small denominators can be highly volatile. Steps need to be taken to normalize the smallest sample sizes. For different case volumes, we will weight each hospital's contribution by the number of cases to reduce the distortion from small-sample hospitals.

We will exclude conditions with too few cases that are not statistically meaningful.

Numeric Formatting

Some entries might include non-numeric characters, which must be cleaned or parsed carefully. The risk-adjusted mortality rates come in formats such as X.X per 100 cases, we need to clean or adjust the scale to make it consistent We will convert mortality rates, case counts, and volume numbers to numeric types and watch out for cells with text.

Outliers, Extreme Values, and Anomalies

Before modeling, we need to explore distributions and flag extreme outliers is crucial. We will flag suspiciously high or low mortality rates and investigate whether these are valid small-sample effects or reporting errors.

Missing, Suppressed Values - Lack of Cases

Many hospitals have no values for many conditions, meaning there will need to be a lot of adding of NAs to better sort the data. We will identify which fields are empty and exclude them from rate calculations.

2.6. Steps for Workflow

Read In with Care

We will use a library depending on the language that tolerates missing or malformed entries and reads numerical/string variables.

Initial Sanity Checks

We plan to count missingness per column, check ranges of numeric variables, look for impossible values (i.e. negatives, greater than 100 percent mortality), narrow down the dataset to the year 2023 only, and inspect sample sizes.

Standardized Column Names and Formats

It is important to unify naming (e.g. "AMI" vs "Acute Myocardial Infarction") and convert rates to consistent numeric scale, coerce types.

Model-Ready Dataset

Ensure to drop or flag residual missing entries, impute if justifiable, and divide into training/test splits if doing an explanatory modeling.

Documentation

We will keep track of cleaning rules for the dataset (e.g. how missing values are treated) to ensure transparency and reproducibility.

3. Methods

3.1. Overview

In the methods section, this paper will discuss the various approaches taken to implement the three models we used. The methods section will begin with a discussion of the models used: linear regression, decision trees, and K-Nearest Neighbors (KNN) regression, and then provide a detailed justification for each of their uses. The methods will continue with the training procedure for each of the models, followed by the model validation plan, and details about each of the models implementations. The methods section aims to be a complete summary on the models used to unpack, for each Californian hospital in 2023, how the hospital ratings and medical conditions are associated with the risk adjusted mortality rate and the proportion of deaths out of total cases.

3.2. Models Used and Justification

In this analysis, we will employ three distinct predictive modeling approaches (linear regression, decision trees, and KNN regression to explore how hospital ratings and medical conditions relate to risk-adjusted mortality rates and death proportions across Californian hospitals in 2023. Linear regression will serve as our baseline model due to its interpretability and capacity to estimate the linear association between predictors, such as rating and condition, and continuous outcomes (mortality rate). Decision trees will be used to capture potential nonlinear interactions and hierarchical decision patterns that may emerge between hospital characteristics and outcomes, allowing for an intuitive visualization. KNN regression will provide a non-parametric perspective, estimating outcomes based on local similarity in hospital profiles and patient conditions. Together, these models represent a balanced methodological framework, enabling us to identify both general trends and localized variations in hospital performance.

3.3. Model Training Procedure

The three models we are training each have unique procedures. Linear regression models optimally weigh explanatory variables in order to predict the outcome variable. To train the linear model, the end point is plotting the x or independent variables or feature against the single linear regression function's production of $\hat{y}$, which is

$$\hat{y} = b_0 + b_1 x.$$

Besides x,

$$b_0 = \bar{y} - b_1 \bar{x},$$

where $\bar{y}$ is the mean of y and $\bar{x}$ is the mean of x. Finally,

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}.$$

A decision tree is a set of decision nodes, with a set of edges, and a set of terminal nodes. These represent the decision points, choices, and outcomes. The goal is to build a tree from data that predicts outcomes for future cases. After importing from `sklearn.tree` the `DecisionTreeRegressor` and `plot_tree`, we will create a target variable y dataframe and an X with features/predictors that want to be used. Using the classifier and `plot_tree`, we will have a decision tree created.

KNN regression computes the distance from $\hat{x}$ to each observation x_i. Then it finds the "nearest neighbors" $x_1^*, x_2^*, \ldots, x_k^*$ to $\hat{x}$ in the data, with outcomes $y_1^*, y_2^*, \ldots, y_k^*$. Finally, you compute the average nearest neighbor outcome as the prediction for $\hat{x}$:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i^*.$$

To create a KNN regression function, follow the code. First, defining the variable:

```
def knn_reg(x_hat, gdf, K):
```

Then, computing distances between $\hat{x}$ and the data:

```
squared_differences =
(x_hat - gdf.loc[:,['x1','x2']])**2
distances =
np.sum(squared_differences,axis=1)
```

Afterwards, finding the k smallest values in dist:

```
neighbors = np.argsort(distances)
[:K].tolist()
```

Find y values for the nearest neighbors:

```
y_star = gdf['y'].iloc[neighbors].
tolist()
```

Average neighbor values to get prediction:

```
y_hat = np.mean(y_star)
```

Finally, return a dictionary of computed values of interest:

```
return {'y_hat': y_hat, 'y_star':
y_star,'neighbors': neighbors}
```

Importantly, the choice and implementation of these models are directly tied to the characteristics of our California hospital dataset and our research question. Our data contains both categorical variables, such as hospital ratings and procedure types, as well as continuous outcomes like risk-adjusted mortality rates. Linear regression allows us to quantify independent effects of hospital quality and procedure type on mortality, providing interpretable coefficients that answer our primary research question. Decision trees complement this by capturing potential nonlinear interactions between hospital characteristics and procedure types, identifying combinations of factors that may lead to high or low mortality rates. KNN regression employs the local similarity among hospitals and procedures estimating mortality outcomes based on observed patterns in similar cases. This is particularly valuable given the variability in case columns and procedure-specific risks across hospitals in our dataset.

3.4. Model Validation Plan

To ensure our models are accurate and generalize well, we will split the data into training and testing sets. This helps us see how the models perform on data they have not seen before. For the linear regression model, we will check how well it fits the data using R-squared, adjusted R-squared, and root mean squared error. For the decision tree and KNN regression, we will use cross-validation to prevent overfitting and to find the best parameters. The decision tree will be tuned and pruned to keep it from becoming too complex, while KNN will be tested with different values of k to find the best number of neighbors. We will compare all models using mean absolute error and root mean squared error to see which one predicts the outcomes most accurately.

3.5. Details about Model Implementation

All models will be built in Python using standard machine learning tools. Before modeling, we will clean the dataset by fixing missing or inconsistent data and convert text categories, like hospital rating and condition (such as heart attack or pneumonia), into numerical form using one-hot encoding. The linear regression model will be used to measure how ratings and conditions relate to the risk-adjusted mortality rate and deaths per case. The decision tree will help us find nonlinear relationships and show which features are most important in predicting outcomes. The KNN regression model will predict outcomes by comparing each hospital to others with similar ratings and conditions. After training the models, we will compare their results using the same validation metrics to decide which one performs best.

4. Results

4.1. Linear Regression Model Results

To examine the factors influencing risk-adjusted mortality rates, three linear regression models were estimated using Ordinary Least Squares (OLS) with HC3 robust standard errors to account for uneven variability in the data. The first model included only hospital ratings as a predictor and showed that hospital ratings as a predictor and showed that

hospital quality significantly affects mortality outcomes. In particular, hospitals rated as "Better" had a coefficient of -1.20, indicating lower mortality compared to the reference category, while hospitals rated as "Worse" had a coefficient of 7.06, suggesting substantially higher mortality. This model explained approximately 11.4% of the variability in mortality rates (R^2 = 0.114), suggesting that hospital rating alone is an important but partial predictor of patient outcomes.

The second model examined procedure or condition type as the sole predictor. The results showed that mortality varies widely depending on the procedure, with acute stroke associated with higher mortality (coefficient = -3.98) and procedures such as hip fracture (coefficient = -3.57) and carotid endarterectomy (coefficient = -3.98) associated with lower mortality. This model explained 18.8% of the variation in mortality rates (R^2 = 0.188), supporting that procedure type is a key factor affecting patient outcomes.

The third combined model incorporated both hospital ratings and procedure/condition, increasing explanatory power to 28.2% (R^2 = 0.282). In this model, hospital rating effects remained significant after controlling for procedure type, with "Better" hospitals associated with lower mortality (coefficient = -2.60) and "Worse" hospitals associated with higher mortality (coefficient = 6.01). The procedure-specific effects were largely consistent with the procedure-only model, with high-risk procedures such as hip fracture and percutaneous coronary intervention associated with lower mortality. All coefficients in this model were statistically significant (p ¡ 0.0001), highlighting the independent impact of hospital quality and procedure type on mortality rate.

To make these findings easier to interpret, predicted mortality rates were calculated from the combined model along with 95% confidence intervals. Figure 7 of predicted mortality by hospital rating showed that "Worse" hospitals consistently had higher predicted mortality, while "Better" hospitals had lower predicted rates, with relatively narrow confidence intervals. In Figure 8, predicted mortality for the top ten most frequent procedures highlighted substantial differences: high-risk procedures such as acute stroke and pneumonia had elevated predicted mortality.

4.2. Decision Tree Regression Model Results

The decision tree regression model had an R-squared of 0.173. The RMSE was 5.238. Some of the hyperparameters of the tree, including the tree depth was 3 to prevent over fitting.

The four groups created at the third level of the model had varying sample sizes. The first group consisted of less than or equal to 0.5 for Pneumonia, and it consisted of 1394 samples with a value of 3.319. The second group consisted of less than or equal to 0.5 for Providence Redwood Memorial Hospital, and it consisted of 218 samples with a value of 6.885. The third group consisted of less than or equal to 0.5 for Colusa Medical Center, and it consisted of 95 samples with a value of 10.499. The last group was only 1 sample with the value of 64.4. (Figure 1). These groups split into a final set of pairs without a comparison. The residuals plot highlights that most of the residuals stayed around zero, which is ideal (Figure 2).

4.3. KNN Regression Model Results

To investigate how hospital ratings and medical procedures relate to risk-adjusted mortality rates, two K-Nearest Neighbors (KNN) regression models were estimated. KNN is a non-parametric, instance-based learning method that predicts outcomes based on the average of the closest K observations in feature space. This approach allows for local similarity effects to influence predictions, rather than assuming a linear or hierarchical relationship.

The first KNN model (Figure 3) compared predicted mortality rates from the KNN model using hospital rating as the predictor. Most predictions cluster around the lower range of mortality rates, reflecting that most hospitals have mortality rates between 0 and 10. The dashed 45-degree line represents perfect predictions, and while the majority of points align closely with this line, there are a few outliers at higher mortality values that deviate substantially. This pattern suggests that the KNN model captures general trends associated with hospital ratings. "Better" hospitals tend to have lower predicted mortality, and "worse" hospitals are higher. However, in extreme cases, it may be harder to predict accurately due to local variability and limited neighbors in those ranges.

The next plot (Figure 4) shows residuals versus predicted values. Most residuals cluster around zero, showing the model is generally unbiased. A few outliers at high predicted mortality suggest some high-risk procedures are mispredicted, but overall, KNN captures typical procedure-level mortality well.

The last plot (Figure 5) about the RMSE vs. K (Conditional Model) shows RMSE for the KNN model predicting mortality by condition across different K values. RMSE is high at K = 1, then decreases and stabilizes around K = 12 - 15, indicating an optimal balance between bias and variance. Larger K values add little improvement and may oversmooth local patterns, emphasizing the need to tune K for best predictive accuracy.

4.4. Neural Network Analysis

While neural network models were not the primary focus of this analysis, we also explored a simple feedforward neural

network to examine predictive patterns in hospital mortality. The results of this model are included in the Appendix (Figure 3). The training accuracy increased steadily over epochs, while the test accuracy remained relatively low and fluctuated, indicating potential overfitting and sensitivity to the dataset's structure. This figure provides supplementary context on how more complex, non-linear models behave with these predictors, though the linear, tree-based, and KNN models remain the main focus of interpretation.

4.5. Benchmark Details

To establish a performance baseline, we first implemented a mean-prediction benchmark model, which predicts the average mortality rate across all hospitals. All regression models: linear regression, decision tree, and K-nearest neighbors (KNN) were trained using an 80/20 train-test split. Features were standardized using z-score normalization to ensure correct scaling for distance-based models.

Linear regression: Three linear regression models were estimated. Predictors included hospital rating, procedure/condition type, and a combination of both. No hyperparameter tuning was required for linear regression beyond ensuring standard errors.

Decision tree regression: The decision tree model was implemented using a maximum tree depth of 3 to prevent overfitting. Other hyperparameters, such as minimum samples per leaf and splitting criteria, were tuned via 5-fold cross-validation to minimize out-of-sample RMSE. The final tree balanced model complexity with predictive performance.

KNN Regression: KNN is a non-parametric, instance-based method that predicts outcomes based on the average of the K nearest neighbors in feature space. Hyperparameters were tuned using a grid search over k =1 to k= 40, evaluated with 5-fold cross-validation. The optimal K was selected by minimizing RMSE on the validation folds.

4.6. Specific Insights

Comparing the three models used, we saw that the linear regression model had the highest overall R-squared, with the decision tree regression in second and the KNN regression in last. (Figure 9) If the neural networks are added to the mix, even at their best performance, they still come behind the decision tree and linear regression models. Overall, while the three main models did not fit the predicted data that well, they demonstrated key insight into what kind of results could be found with the data that the project was based on. Across models, hospitals rated as "Better" consistently showed lower predicted mortality, while high-risk procedures such as acute stroke or pneumonia exhibited elevated mortality regardless of hospital. Overall, these findings emphasize that both hospital quality and procedure type independently influence patient outcomes and that model choice affects the ability to capture global trends versus local variation.

Model Comparisons

Linear Regression The linear regression models provided a clear, interpretable baseline. Hospital rating and condition were both statistically significant predictors, with "better" hospitals showing meaningfully lower RAMR and high-risk conditions showing higher RAMR. However, linear models captured only broad trends and left unexplained variance due to nonlinear clinical relationships.

Decision Tree The decision tree model captured nonlinear patterns and produced clear splits showing which conditions and ratings most strongly drive mortality. Trees modeled interactions naturally kike for example, identifying that certain conditions only lead to high RAMR in hospitals with low ratings. However, trees risk overfitting, especially with many rare conditions, requiring pruning to improve generalization.

KNN Regression KNN regression modeled RAMR using localized similarity among hospitals. The model performed best at moderate K values (around 12–15), balancing noise and bias. KNN captured clusters of similar procedures well but struggled with rare or high-mortality conditions where few close neighbors exist. Residual plots showed the model is generally unbiased, with some outliers for extreme cases.

5. Conclusion

This project set out to examine the relationship between hospital ratings, medical conditions, and risk-adjusted mortality rates in California hospitals in 2023. We analyzed hospitals' performance across a range of procedures and conditions, applying three modeling approaches: linear regression, decision tree regression, and KNN regression. The aim was to quantify how hospital quality and the type of medical condition contribute to patient outcomes.

5.1. Finished Model Considerations

The linear regression models provide clear evidence that hospital ratings and procedure types independently influence mortality outcomes. In particular, hospitals rated as "Better" were associated with lower risk-adjusted mortality, while "Worse" hospitals showed significantly higher mortality rates. The magnitude of these effects remained consistent even when controlling for procedure type, demonstrating that hospital quality contributes meaningfully to patient outcomes. The procedure-specific analyses revealed that certain high-risk conditions, such as acute stroke and pneumonia, were associated with elevated mortality rates, while other procedures, such as hip fracture and carotid endarterectomy, had relatively lower morality. The combined

linear regression model, which incorporated both hospital ratings and procedure types, explained nearly 28% of the variability in mortality rates, highlighting the importance of considering multiple dimensions of hospital performance.

The decision tree analysis produced more modest results. Although decision trees are well suited for uncovering non-linearities and interaction effects, the predictive accuracy of the model was relatively low despite pruning and depth constraints designed to prevent overfitting. This is likely due to several data limitations. The tree was limited to only 3 layers of depth, and the tree was limited in its regression capabilities. The model had limited success in the accuracy of the regression overall. More analysis on different hyper-parameters is necessary.

KNN regression provided a non-parametric perspective by modeling local similarity effects. The KNN method effectively captured trends associated with hospital ratings, showing that hospitals rated "Better" generally had lower predicted mortality rates and "Worse" hospitals had higher rates. The residual analysis indicated that the model was generally unbiased, though high-risk procedures remained challenging to predict precisely. The optimal performance emerged around K 12-15. This approach underscores the value of using flexible, data-driven models to understand outcomes without imposing strict parametric assumptions.

5.2. Limitations

Despite the strengths of these models, the project has several limitations. First, the dataset captures only risk-adjusted mortality and a limited set of explanatory variables. Important predictors such as hospital size, patient demographics, patient socioeconomic status, case mix complexity, and staffing ratios were not included. Their omission likely contributes to the unexplained variance in all the models. Second, some of the mortality data were highly skewed, with a small number of hospitals or procedures exhibiting very high mortality rates. These outliers can influence both linear and non-linear models, making it difficult to achieve perfect prediction accuracy. Third, the KNN model's performance is contingent on sufficient local observations; sparse data for rare procedures or atypical hospitals limit the model's reliability in those cases.

While our study is limited by these factors, the methodological approach was carefully designed to ensure robust and interpretable results. Linear regression provided clear, baseline estimates of the independent effects of hospital ratings and procedure type, while KNN regression allowed for flexible, data-driven insights into local patterns in mortality rates. Decision trees offered an additional perspective on non-linear interactions. We mitigated data limitations by weighting hospitals by case volume, excluding extremely small-sample procedures and performing thorough data cleaning and pre-processing. These choices reflect a careful balance between methodological rigor and practical constraints, and our results remain meaningful, highlighting the independent role of hospital quality in patient outcomes.

5.3. Future Research

Future research could strengthen and expand upon this work in several ways. For instance, incorporating additional predictors like patient demographics, hospital resources, case severity indices, and regional healthcare characteristics could improve predictive performance and provide a more comprehensive understanding of the determinants of mortality. Using more advanced machine learning models, such as random forests could potentially help capture complex interactions and non-linear relationships that both linear and KNN models may miss. Using a neural network is another avenue to look into. After some preliminary exploration, it was determined that even at best, a neural network model could only achieve 0.167 test accuracy. (Figure 3) Additionally, assessing temporal patterns by integrating multiple years of hospital data could provide insight into whether hospital performance is stable or fluctuates over time. Finally, extending the analysis to other quality metrics, such as readmission rates, surgical complications or patient satisfaction scores, would help contextualize mortality outcomes within a broader framework of healthcare quality.

Overall, this project demonstrates that both statistical and machine learning approaches offer valuable tools for understanding hospital performance. Linear regression provides clear, interpretable estimates that help identify broad trends and the independent effects of hospital ratings and procedure types. KNN regression, meanwhile, offers flexible modeling of localized relationships, capturing subtle patterns not easily identified through parametric models. Decision tree regression allows for a unique method that is both interpretable and captures less linear trends. Together, these methods underscore the importance of combining traditional statistical techniques with modern machine learning approaches to produce richer and more robust insights into healthcare outcomes.

6. References

Department of Health Care Access and Information. (November 23, 2025). Data.gov. https://catalog.data.gov/dataset/california-hospital-inpatient-mortality-rates-and-quality-ratings-8e21f

7. Appendix

7.1. Credit Assignments Annotation

7.1.1. Abstract Credit Assignment

The abstract was written by Hilda with couple edits from Eddie and Caroline.

7.1.2. Data Credit Assignment

Eddie did the data section, key variables, research question, and the missing, suppressed value challenge. Hilda and Caroline worked both on the rest of the challenges and the scope, as well as the EDA in the data section. Hilda worked on Read in with Care, Standardized Column Names and Formats, and Model-Ready Dataset. Caroline did documentation and initial sanity checks. Basic statistics table of the data figure 1 and the sample table of key variables figure 2 was done by Hilda. Data Overview was done by Hilda.

7.1.3. Methods Credit Assignment

Overview and procedure were done by Eddie. Models used and justification were by Caroline. Validation plan and implementation were by Hilda.

7.1.4. Results Credit assignment

All linear regression work was performed by Caroline. All decision trees were by Eddie. All KNN regression was by Hilda. Neural networks were created by Eddie and written about by Hilda. Benchmark details was done by Hilda. Specific Insights was done by Hilda and Eddie. Model Comparison was done by Hilda.

7.1.5. Conclusion Credit Assignment

All linear regression work was performed by Caroline. All decision trees were by Eddie. All KNN regression was by Hilda. Hilda wrote future research and limitations. Caroline wrote the defense. Eddie wrote about and created neural networks.

7.1.6. References Credit Assignment

Done by Eddie.

7.1.7. Appendix Credit Assignment

Figures 6, 7, 8, and 14 are by Eddie. 9, 10, and 11 are by Hilda. 3, 4, 5, 12, and 13 are by Caroline.

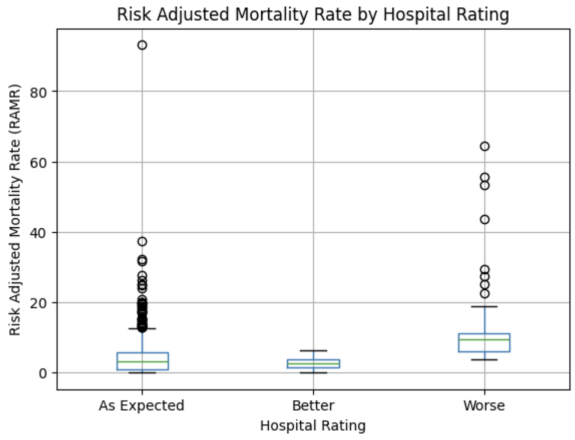

Figure 3. Risk Adjusted Mortality Rate by Hospital Rating

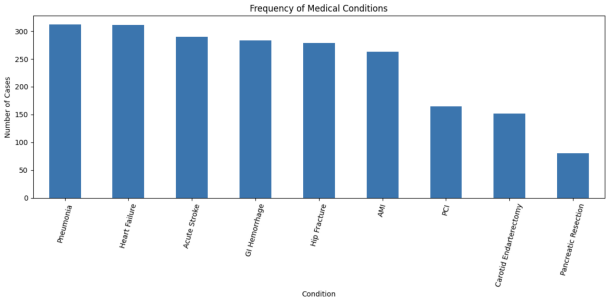

Figure 4. Frequency of Medical Conditions

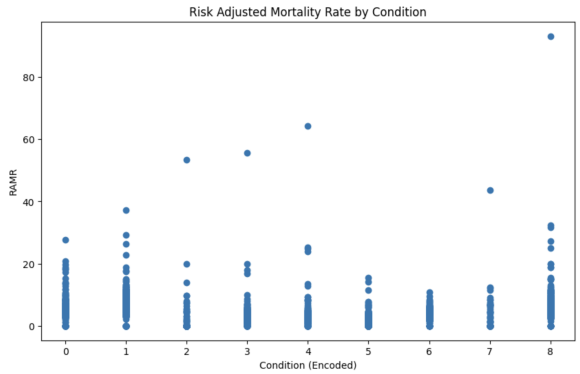

Figure 5. Risk Adjusted Mortality Rate by Condition

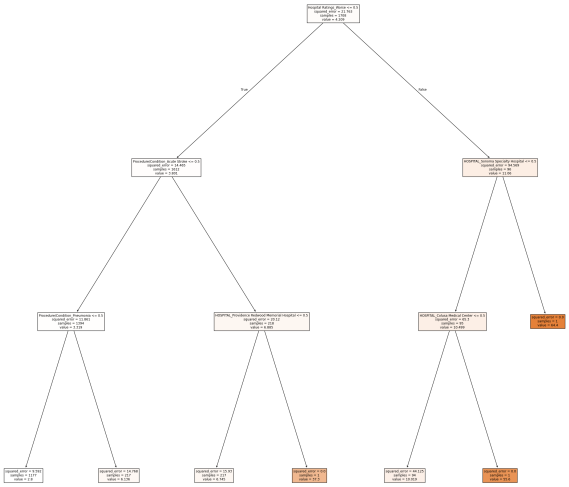

Figure 6. Decision Tree Regression

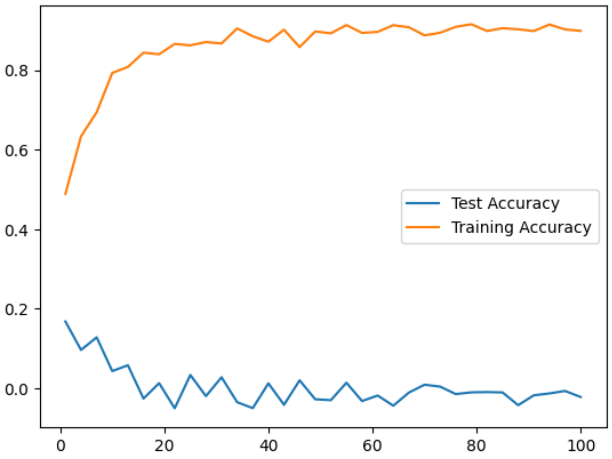

Figure 8. Neural Network Accuracy

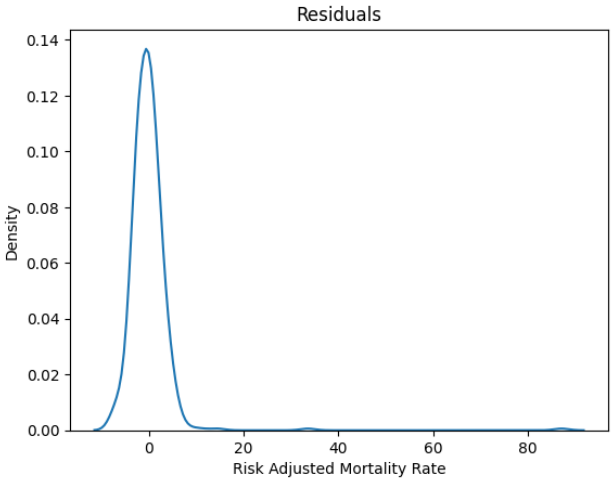

Figure 7. Residuals Plot for Decision Tree

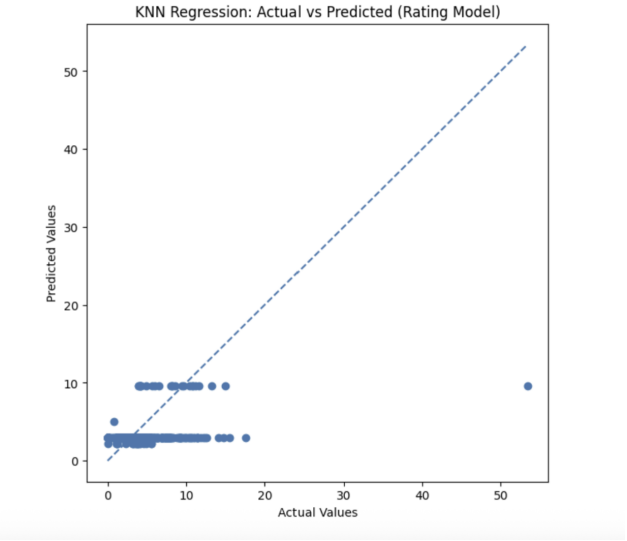

Figure 9. KNN Regression Actual vs Predicted Rating Model Plot

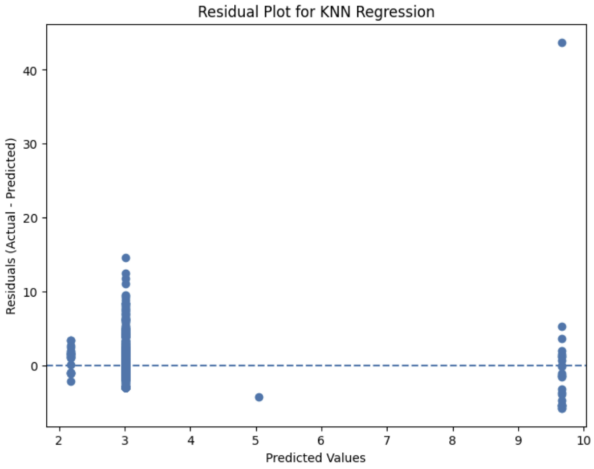

Figure 10. KNN Regression Residual Plot

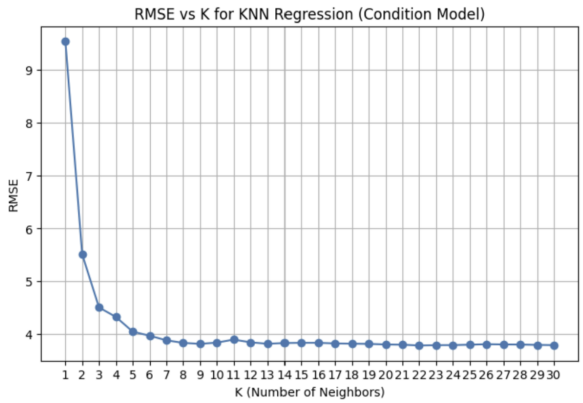

Figure 11. KNN Regression Condition Model

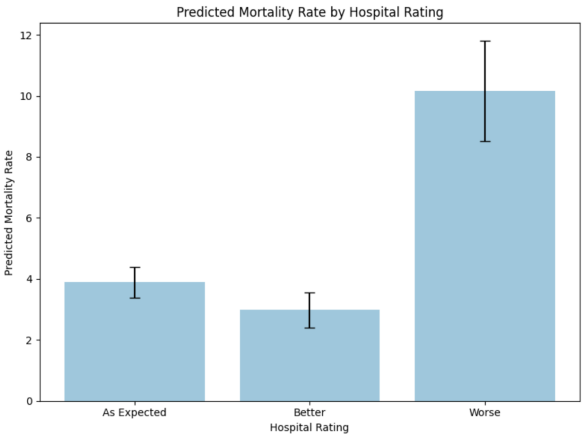

Figure 12. Predicted Morality Rate by Hospital Rating Bar Plot

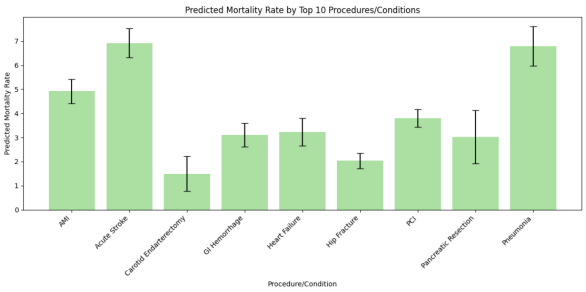

Figure 13. Predicted Morality Rate by Top 10 Procedures/Conditions Bar Plot

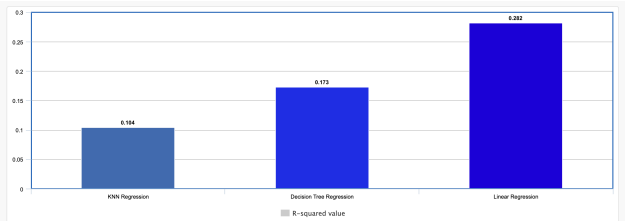

Figure 14. R-Squared Model Comparison