ISyE 6740 – Computational Data Analysis - Fall 2025
Final Report

Team Member Names: Olan Malcolm

Project Title: Predicting the Cost of Medical Insurance

Problem Statement

Accurately estimating healthcare costs is a crucial challenge in the field of health economics and predictive analytics. It is equally important for consumers to understand the potential costs associated with their healthcare coverage. Healthcare expenses are influenced by multiple factors – such as an individual's age, lifestyle, and medical history – making it difficult for insurance companies to determine fair and accurate premium rates.

This project aims to develop a machine learning model capable of predicting an individual's medical insurance cost based on the features provided in the dataset and from other outside data. Such a model can assist insurance companies in improving risk assessment and pricing strategies, while also offering insights into how personal characteristics influence healthcare costs. From the consumer's perspective, this model could enable individuals to estimate their own insurance costs and compare them to quotes provided by insurance companies, promoting transparency and informed decision-making.

A large portion of Americans rely on the Affordable Care Act (ACA) marketplace because they otherwise could not afford health insurance. As of early 2024, more than 45 million people were enrolled in ACA-related plans (*The Affordable Care Act and the Data: Who Is Insured and Who Isn't*, 2025). Many marketplace enrollees struggle with affordability – a 2023 survey found that 57% of people in marketplace plans reported difficulty paying for care (Collins et al., 2023), and a 2024 survey found that 37% said they delayed or skipped needed care due to costs (Center on Budget and Policy Priorities, 2024). Without these ACA subsidies, many of these individuals would likely go uninsured – roughly 44% of ACA enrollees reported they would forgo coverage if the marketplaces didn't exist (LaPick, 2022).

By modeling individual-level risk and cost, the proposed machine learning system not only helps insurers set fair premiums, but also sheds light on the financial struggles of ACA enrollees who depend on these subsidies to access care.

Data Source

There were two datasets used for this project. The bulk of the data comes from the "Medical Insurance Cost Dataset" available on Kaggle (Mosap Abdel-Ghany, 2025). The dataset contains information on 1,338 individuals, each described by six independent variables and one dependent variable. The independent variables include age, sex, BMI (Body Mass Index), number of children, smoker status, and region. The dependent variable is charges, which represents the individual's yearly medical insurance cost.

The second dataset used comes from the U.S. Bureau of Economic Analysis (BEA), which provides the 2024 Per Capita Personal Income for each U.S. state (BEA, n.d.). Because the Kaggle dataset categorizes individuals into four regions – Northeast, Southeast, Northwest, and Southwest – and these categories do not correspond to any official U.S. Census definitions, a custom regional mapping was used to align the BEA state-level data with the regions used in the Kaggle dataset. The custom regions and the states in each are:

- Northeast (NE):
  Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, Delaware, Maryland, New Jersey, New York, Pennsylvania, District of Columbia
- Southeast (SE):
  Virginia, West Virginia, North Carolina, South Carolina, Georgia, Florida, Alabama, Mississippi, Tennessee, Kentucky, Arkansas, Louisiana
- Southwest (SW):
  Texas, Oklahoma, New Mexico, Arizona
- Northwest (NW):
  Washington, Oregon, Idaho, Montana, Wyoming, Alaska

This custom classification ensured that the BEA state-level income data could be consistently merged with the regional categories present in the Kaggle dataset, enabling a unified analysis across both sources.

An example of a row in the final dataset looked like the following:

Column	Explanation	Example Value
Age	Age of insured	34
Sex	Gender of insured	Male
BMI	Measure of body fat based on height and weight	27.78
Children	Number of children of insured	2
Smoker	If insured smokes or not	No
Region	Region the insured lives in	Southeast
Avg_region_wage	The average wage of people who live in the region	75063.583333
Charges	Cost of medical insurance to the insured	21984.47061

Table 1. Example of Data

Methodology

Feature Engineering and Preprocessing

To prepare the dataset for modeling:

- Categorical Encoding: The categorical variables sex and smoker were encoded into binary (0,1) – (male=0, female=1) and (no=0, yes=1) respectively. The region column was converted, using one-hot encoding, into four binary indicators (region_northeast, region_southeast, region_southwest, region_northwest) using get_dummies to allow machine learning algorithms to process them effectively.

- Numerical Features: Continuous features (age, bmi, children, avg_region_wage) were standardized using StandardScaler to ensure they were on the same scale, which is important for models sensitive to feature magnitude, such as ridge regression and neural networks.
- Missing Data: The dataset was checked for missing values – none were found.
- Feature Selection: All Kaggle-provided features plus avg_region_wage were retained for modeling, as they are directly relevant to predicting insurance charges.

Dimensionality Reduction

A Principal Component Analysis (PCA) was performed on the numeric features (age, bmi, children, avg_region_wage) to examine potential multicollinearity and reduce dimensionality:

- PCA was applied to retain 95% of the variance.
- Standardization of numeric features ensured that differences in scale did not bias the PCA.
- PCA reduced 0 variables, indicating that all numeric features were needed to capture the desired variance. Therefore, PCA was not applied in the final modeling, as it did not simplify the dataset.

Model Development

Multiple models were trained and compared to predict individual medical insurance costs (charges). Each model was chosen to capture different types of relationships in the data:

- Linear Regression: Estimates the relationship between independent variables and the target by fitting a linear equation. It is simple, interpretable, and provides a baseline model for comparison (James et al., 2021).
- Ridge Regression: A linear model with L2 regularization, which adds a penalty proportional to the square of the coefficients. This reduces overfitting and handles multicollinearity by shrinking coefficients of less important features (Hoerl & Kennard, 2000).
- LASSO (Least Absolute Shrinkage and Selection Operator) Regression: Similar to Ridge but uses L1 regularization, which can shrink some coefficients to exactly zero. This performs automatic feature selection, identifying the most important predictors of insurance charges (Tibshirani, 1996).
- Elastic Net Regression: Combines L1 (LASSO) and L2 (Ridge) penalties to balance variable selection and regularization. Useful when features are highly correlated or when both sparsity and stability are desired (Zou & Hastie, 2005).
- Regression Tree: A decision tree that splits the dataset into subsets based on feature values to predict the target. Captures non-linear relationships and interactions between features but can overfit if not controlled (Awad & Khanna, 2015).

- Random Forest: An ensemble of multiple regression trees built on random subsets of the data and features. Averaging predictions from many trees reduces overfitting and improves predictive accuracy (Breiman, 2001).
- Gradient Boosting: Sequentially builds regression trees where each new tree focuses on correcting the errors of previous trees. Captures complex patterns and interactions, often achieving higher accuracy than single trees or random forests (Friedman, 2001).
- Neural Network (MLP): A multi-layer feed-forward neural network capable of modeling highly non-linear relationships. Learns complex interactions between features but requires careful tuning of architecture, learning rate, and regularization to avoid overfitting (Goodfellow et al., 2016).

Model Evaluation

Each model was built and evaluated by:

- Cross-Validation: All models were evaluated using 5-fold cross-validation (KFold, n_splits=5, shuffle=True, random_state=42) to ensure robust performance estimates.
- Performance Metric: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2, Average CV MSE, and Average CV RMSE were used to quantify predictive accuracy.
- Hyperparameter Tuning: Regularized models (Ridge, Lasso, Elastic Net), tree-based models (max depth, number of estimators), and neural networks (layer sizes, activation functions, learning rate) were tuned to optimize performance using GridSearchCV.

This methodology ensures that the models capture both linear and non-linear effects in healthcare costs while providing a robust framework for comparing predictive performance across different approaches.

Exploratory Data Analysis (EDA)

To understand the distributions and relationships of features in the original Kaggle dataset, an exploratory data analysis was done. The regional wages were not included in the EDA because it wasn’t part of the original dataset. Adding it could change the patterns and relationships between the original features, making it harder to understand the dataset as it was. By leaving it out, the analysis reflects the real distributions and relationships in the original data.

To examine the distribution, skewness, and presence of outliers in the numeric features (age, bmi, children, charges), histograms were plotted. Bar plots were used for the categorical features (sex, smoker, region) to visualize the total count of each category and identify potential imbalances.

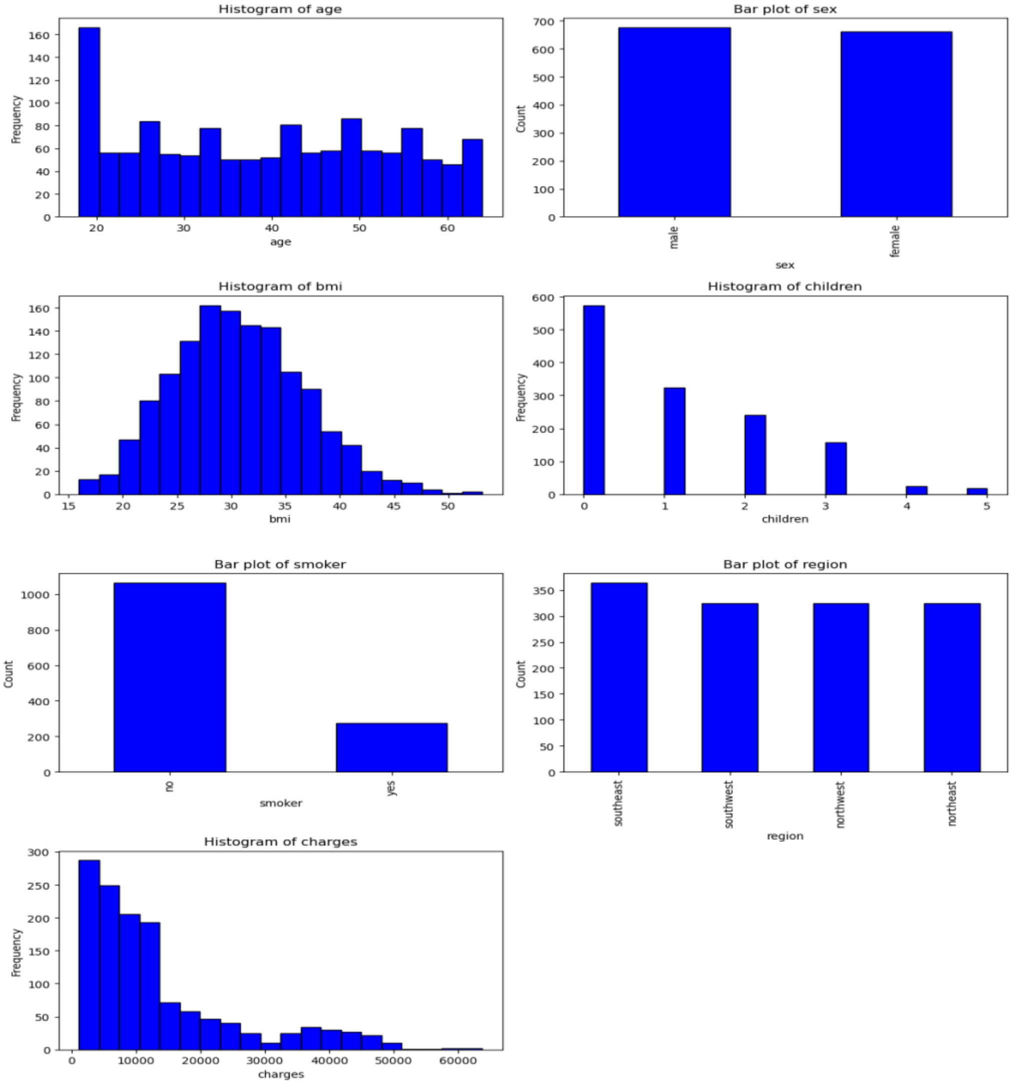

Figure 1. Histograms and Bar Plots of Features

The visualizations reveal several key patterns. There is a relatively large number of insured individuals in the 18–20 year-old range. The dataset appears to have an approximately even balance of males and females. The distribution of BMI values is roughly normal, with noticeable outliers occurring above 50. Most insured individuals have between 0 and 3 children, while those with 4 or 5 children are relatively rare. There are approximately 800 more non-smokers than smokers in the dataset. The insured population is pretty evenly distributed across regions. Finally, most of the healthcare charges fall below $15,000.

After these were looked at, boxplots were used to look at the relationship between charges and region, segmented by sex and smoker. This was meant to see the differences in charges across the demographic groups and identify any outliers.

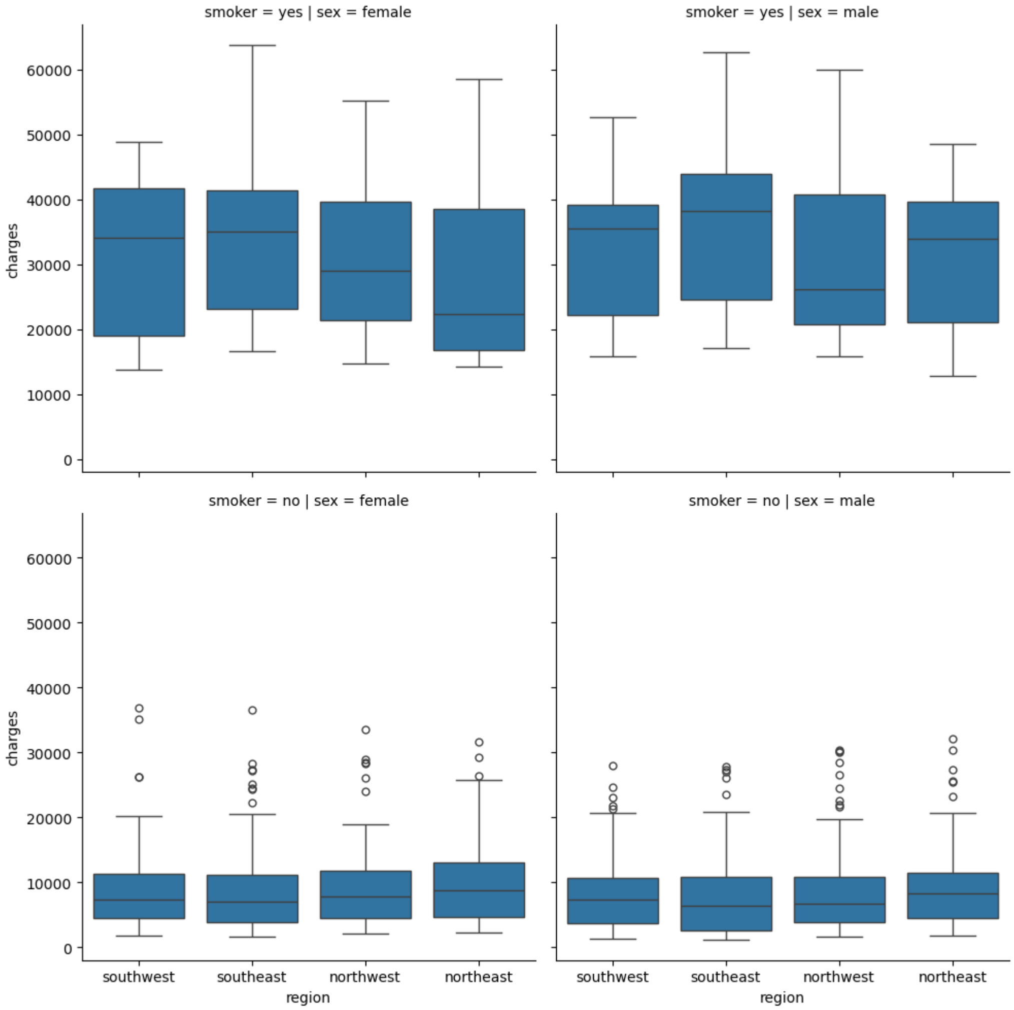

Figure 2. Boxplots of Charges by Region Segmented by Smoker and Sex

The boxplots reveal patterns in healthcare charges between smokers and non-smokers. For smokers, no outliers are observed, and their charges are consistently higher than those of non-smokers. In contrast, the boxplots for non-smokers show outliers across all regions, regardless of the insured's sex, indicating that a small subset of non-smokers incurs unusually high charges. The outliers were not removed, for they represent legitimate high-cost cases which are important for modeling.

The final item that was looked at was a correlation matrix of the features. The matrix was made using both numeric and encoded categorical features.

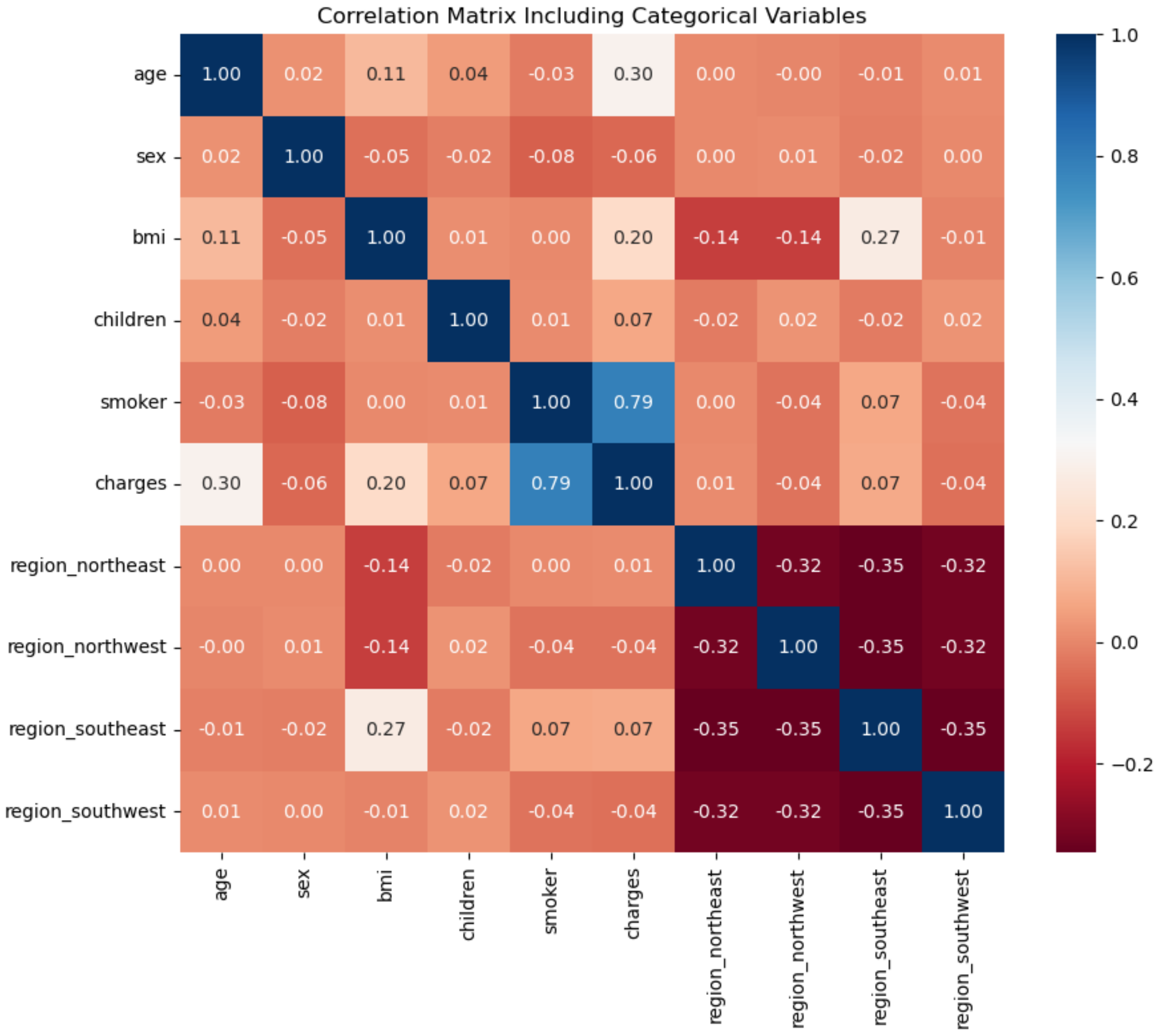

Figure 3. Correlation Matrix

Based on the correlation matrix, smoker shows the strongest positive correlation with healthcare charges, indicating it is the most influential predictor. Age shows the second highest positive correlation, followed by BMI. No features display a substantial negative correlation with charges, suggesting that all predictors either have a positive relationship or little impact on the target variable.

To ensure the dataset accurately reflects real-world insurance costs, selected charges were compared to benchmark premiums from HealthMarkets (2025) for individuals of specific ages, sexes, smoker statuses, and family sizes in representative cities from each region. This comparison helps verify the reasonableness of the dataset and highlights any significant deviations from typical marketplace costs.

Region	Example Profile	ACA Benchmark ($)	Dataset Charges ($)
Northeast	25 y/o male nonsmoker 0 children (Philadelphia, PA)	3,149.76 minimum	2,724.36

Northeast	23 y/o female smoker with 2 children (Philadelphia, PA)	6,577.44 – 19,728	38,511.62
Northeast	52 y/o female nonsmoker 1 child (Philadelphia, PA)	8,627.88 – 24,298.9	10,106.13
Southeast	25 y/o male nonsmoker 0 children (Atlanta, GA)	3,649.92 minimum	2,137.65
Southeast	23 y/o female smoker with 2 children (Atlanta, GA)	12,972.72 – 31,258.2	36,021.01
Southeast	53 y/o female nonsmoker 1 child (Atlanta, GA)	11,216.20 minimum	10,579.71
Southwest	26 y/o male nonsmoker 1 child (Phoenix, AZ)	4,824 minimum	2,904.96
Southwest	36 y/o female smoker with 2 children (Phoenix, AZ)	8,268 – 23,460	18,608.26
Southwest	51 y/o female nonsmoker 1 child (Phoenix, AZ)	7,716 – 21,864	9,871.26
Northwest	25 y/o male nonsmoker 0 children (Seattle, WA)	2,760 – 9,972	2,528.78
Northwest	32 y/o female smoker with 2 children (Seattle, WA)	9,624 – 32,796	32,734.19
Northwest	53 y/o female nonsmoker 1 child (Seattle, WA)	8,172 – 27,840	10,950.91

Table 2. Dataset Charges vs. ACA Benchmarks

The table shows that most dataset charges fall within or near the ranges observed in the ACA marketplace, suggesting that the data is generally realistic. Some high-cost cases, particularly for smokers with multiple children, exceed ACA minimum benchmarks, reflecting the presence of high-cost outliers that are important for modeling. These deviations are consistent with the real-world variability of healthcare costs and highlight the dataset's ability to capture extreme but legitimate insurance charges. This verification provides confidence that the Kaggle dataset is appropriate for predictive modeling of medical insurance costs.

Evaluation and Final Results

Linear Regression

Linear regression is a method that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation. The model assumes a constant linear relationship between features and the target. While highly interpretable, it can struggle to capture non-linear relationships and interactions between features, which may limit predictive accuracy for complex datasets (James et al., 2021). After running the linear regression, the equation produced was:

$$\begin{aligned} y = {} & -15064.2808 + 256.9757 age + 18.5917 sex + 337.0926 bmi + 424.2788 children + 23651.1289 smoker \\ & + 0.0374 avg\_region\_wage + 51.2244 region\_northeast + 116.9645 region\_southeast \\ & - 104.4884 region\_southwest - 63.7005 region\_northwest \end{aligned}$$

The regression results indicate that several factors are associated with the predicted outcome. The intercept here means that an individual's charge is 0 if they have no features (somebody cannot have a negative insurance charge). Age, BMI, and the number of children each have positive effects, with older individuals, those with higher BMI, and those with more children showing higher predicted values. Sex has a small positive effect, with females slightly increasing the predicted value compared to males. Smoking is the most influential predictor, raising the predicted outcome by over $23,000 for smokers. The average regional wage has a very minor positive effect. The model also includes indicators for region, which adjust the predicted value depending on location: living in the Northeast or Southeast increases the predicted value slightly, while living in the Southwest or Northwest reduces it. Overall, lifestyle factors like smoking, along with demographic characteristics such as age, BMI, and family size, drive most of the variation in the outcome.

Ridge Regression

Ridge regression is an extension of linear regression that introduces an L2 regularization term, penalizing the sum of the squared coefficients. This reduces the impact of multicollinearity and overfitting, particularly when predictors are highly correlated, while retaining all features in the model (Hoerl & Kennard, 2000). After running the ridge regression, the equation produced was:

$$\begin{aligned} \mathrm{y} = {} & 13346.0897 + 3611.3527 age + 8.6223 sex + 2034.4746 bmi + 516.7630 children + 9549.2731 smoker \\ & + 126.3444 avg\_region\_wage + 123.6425 region\_northeast - 27.5490 region\_southeast \\ & - 105.8791 region\_southwest + 10.0572 region\_northwest \end{aligned}$$

The Ridge regression model, with the hyperparameter alpha = 1 selected via GridSearchCV, was trained on standardized numeric features. This standardization ensures that all continuous predictors contribute proportionally to the penalty term, allowing the model to fairly shrink coefficients. Age and BMI, after scaling, show substantial positive effects, indicating that older individuals and those with higher BMI tend to have higher predicted charges. The number of children also contributes moderately. Females have slightly higher predicted outcomes than males, as indicated by the positive coefficient for sex. Smoking remains a strong predictor, increasing the predicted outcome by roughly $9,550. Average regional wage has a minor positive effect, and regional indicators show small variations: living in the Northeast slightly increases predicted charges, while residing in the Southeast or Southwest slightly decreases them, and living in the Northwest has minimal effect. Overall, Ridge regression shrinks coefficients slightly compared to ordinary linear regression but confirms that lifestyle factors, particularly smoking, along with age, BMI, and family size, are the main drivers of the predicted outcome.

LASSO

LASSO adds an L1 regularization penalty to linear regression, which can shrink some coefficients exactly to zero. This effectively performs feature selection, identifying the most influential predictors while reducing model complexity and preventing overfitting (Tibshirani, 1996). After running LASSO regression, the equation produced was:

$$\begin{aligned} \mathrm{y} = {} & 13346.0897 + 3570.4083 age + 0.0 sex + 1980.0103 bmi + 470.3008 children + 9506.8749 smoker \\ & + 153.7974 avg\_region\_wage \ + 65.6262 region\_northeast \ + 0.0 region\_southeast \\ & - 56.1539 region\_southwest + 0.0 region\_northwest \end{aligned}$$

The LASSO regression model, tuned using GridSearchCV with a best alpha value of 50, was trained on standardized numeric features. Age and BMI remain strong positive contributors after scaling, indicating that older individuals and those with higher BMI tend to have higher predicted outcomes. The number of children also adds moderately to the prediction. Smoking continues to be a dominant factor, with smokers showing a substantial increase of roughly $9,500 compared to non-smokers. Average regional wage has a minor positive effect. Through its regularization process, LASSO set the coefficients for sex, Southeast, and Northwest regions to zero, suggesting that these variables do not meaningfully improve the model once the stronger predictors are accounted for. The remaining regional indicators contribute small adjustments: living in the Northeast slightly increases the predicted value, while living in the Southwest slightly decreases it. Overall, the LASSO model emphasizes the key role of smoking, age, BMI, and family size, while effectively removing weaker predictors to reduce model complexity and enhance interpretability.

Elastic Net

Elastic Net combines the penalties of LASSO and Ridge regression to balance variable selection and coefficient shrinkage. This model is particularly useful when features are highly correlated, as it can maintain groups of correlated predictors while still performing regularization (Zou & Hastie, 2005). The equation Elastic Net produced was:

$$\begin{aligned} \mathrm{y} = {} & 13346.0897 + 3606.0796 age + 0.0 sex + 2025.0299 bmi + 507.4268 children + 9547.6735 smoker \\ & + 177.4419 avg\_region\_wage \ + 86.6166 region\_northeast \ + 0.0 region\_southeast \\ & - 78.0852 region\_southwest + 0.0 region\_northwest \end{aligned}$$

The Elastic Net model, tuned using GridSearchCV with a best alpha of 10 and an L1 ratio of 1.0, was trained on standardized numeric features to ensure fair regularization across predictors. The results closely resemble LASSO due to the full weighting on the L1 penalty. Several predictors remain the main drivers of the outcome. Age and BMI both show strong positive effects after scaling, indicating that older individuals and those with higher BMI tend to have higher predicted values. The number of children also contributes moderately to increasing the prediction. Smoking continues to be a major factor, with smokers exhibiting a substantial increase of roughly $9,500 compared to non-smokers. Average regional wage has a small positive influence on the predicted outcome. As in LASSO, the coefficients for sex, Southeast, and Northwest were shrunk to zero, suggesting they do not provide additional predictive value once the stronger features are accounted for. Among the remaining regions, living in the Northeast slightly increases predicted values, while living in the Southwest slightly decreases them. Overall, the Elastic Net model reinforces the importance of smoking, age, BMI, and family size, while confirming that several weaker predictors add little benefit and can be effectively excluded for a more interpretable model.

Estimated Coefficients by Regression Models

Below are the coefficients of each of the above models:

Model	Linear Regression	Ridge Regression	LASSO	Elastic Net
Intercept	-15064.2808	13346.0897	13346.0897	13346.0897
Age	256.9757	3611.3527	3570.4083	3606.0796
Sex	18.5917	8.6223	0.0	0.0
BMI	337.0926	2034.4746	1980.0103	2025.0299
Children	425.2788	516.7630	470.3008	507.4268
Smoker	23651.1289	9549.2731	9506.8749	9547.6735
Avg_Region_Wage	0.0374	126.3444	153.7974	177.4419
Region_Northeast	51.2244	123.6425	65.6262	86.6166
Region_Southeast	116.9645	-27.5490	0.0	0.0
Region_Southwest	-104.4884	-105.8791	-56.1539	-78.0852
Region_Northwest	-63.7005	10.0572	0.0	0.0

Table 3. Regression Coefficients

For the Ridge, LASSO, and Elastic Net models, all numeric features were standardized prior to modeling. This means the reported coefficients correspond to changes in the scaled units of each feature rather than the original units. As a result, coefficients should be interpreted in terms of relative effect sizes rather than absolute dollar changes per unit increase. Features with larger standardized coefficients indicate stronger influence on predicted charges compared to features with smaller coefficients.

The coefficient table reveals several consistent patterns across all four models. Smoking, age, and BMI stand out as the strongest and most reliable predictors, maintaining large positive coefficients even after regularization. The number of children also holds a stable positive effect across models, though to a lesser extent. In contrast, several predictors – including sex, Southeast region, and Northwest region – are shown to be weak contributors for both LASSO and Elastic Net shrink their coefficients to zero, indicating limited predictive value once stronger variables are accounted for. Regional effects overall are modest, with only small positive or negative adjustments depending on the model, suggesting that where one lives plays a relatively minor role. Standardization also substantially reduces the magnitude of the smoking coefficient and increases the importance of age and BMI compared to the unregularized model, highlighting how these methods redistribute influence among correlated predictors. Overall, the table shows the levels of predictor importance, with smoking, age, BMI, and family size being the primary drivers of the outcome.

Regression Tree

A regression tree splits the dataset into subsets based on feature values to predict a continuous target. Each split is chosen to minimize a loss function, typically mean squared error. Regression trees can capture non-linear relationships and feature interactions but are prone to overfitting unless depth or leaf size is controlled (Awad & Khanna, 2015). Unlike the above models, a regression tree does not produce a regression equation. Instead, it gives feature importances, which were:

Feature	Importance
Smoker	0.6931
BMI	0.1793
Age	0.1156
Children	0.0063
Region_Northwest	0.0034
Avg_Region_Wage	0.0022
Sex	0.0
Region_Northeast	0.0
Region_Southeast	0.0
Region_Southwest	0.0

Table 4. Regression Tree Feature Importances

The regression tree model, tuned with optimal hyperparameters (max_depth = 5, min_samples_split = 2, min_samples_leaf = 4), shows that feature importance is strongly focused in a few areas. Smoking dominates the predictions, accounting for nearly 70% of the model's decision-making power. BMI and age are secondary contributors, though much less important than smoking. Children, regional indicators, average regional wage, and sex play minimal roles. Several features receiving zero importance, indicating that the tree found no meaningful splits based on them. The selected hyperparameters ensure that the tree is deep enough to capture the main patterns but constrained enough to prevent overfitting, with each leaf containing at least 4 samples for stability. Overall, the regression tree confirms that smoking, BMI, and age are the primary drivers of variation in the outcome, while all other predictors have negligible influence.

<u>Random Forest</u>

A regression tree is grown by repeatedly splitting the dataset into subsets based on feature values. A single tree is usually inefficient due to its high variance and sensitivity to noise. Random Forest grows multiple trees and averages their predictions, improving accuracy and reducing variance through the law of large numbers (Breiman, 2001). The feature importance's of the Random Forest model were:

Feature	Importance
Smoker	0.6624
BMI	0.1890
Age	0.1265
Children	0.0110
Avg_Region_Wage	0.0041
Sex	0.0026

Region_Northwest	0.0014
Region_Northeast	0.0013
Region_Southeast	0.0010
Region_Southwest	0.0007

Table 5. Random Forest Feature Importances

The Random Forest model, tuned with optimal hyperparameters (n_estimators = 300, max_depth = 10, min_samples_split = 10, min_samples_leaf = 4), identifies smoking as the dominant predictor, accounting for approximately 66% of the model's overall importance. Age and BMI are also meaningful contributors, with smaller but notable importance, while children and average regional wage play minor roles. Sex and all regional indicators have minimal importance, reflecting their limited impact on predictions. The selected hyperparameters ensure the forest is deep enough to capture complex patterns and interactions while limiting overfitting and using 300 trees stabilizes predictions through aggregation. Overall, the Random Forest confirms the strong influence of smoking, age, and BMI on the outcome, while other predictors contribute minimally.

Gradient Boosting

Gradient Boosting sequentially builds regression trees where each new tree corrects the errors of the previous trees. By optimizing a loss function and iteratively adding trees, it captures complex patterns and feature interactions. While powerful, it requires careful tuning of learning rate, tree depth, and number of trees to prevent overfitting (Friedman, 2001). The Gradient Boosting feature importances were:

Feature	Importance
Smoker	0.8188
BMI	0.0979
Age	0.0432
Children	0.0132
Avg_Region_Wage	0.0089
Region_Southwest	0.0071
Sex	0.0059
Region_Northwest	0.0051
Region_Northeast	0.0
Region_Southeast	0.0

Table 6. Gradient Boosting Feature Importances

The Gradient Boosting model, tuned with optimal hyperparameters (n_estimators = 100, learning_rate = 0.05, max_depth = 3, subsample = 0.8, colsample_bytree = 1.0), highlights smoking as the dominant predictor, accounting for over 81% of the model's importance. BMI and age are secondary contributors, while children, average regional wage, sex, and several regional indicators play minor roles. The Northeast and Southeast regions have negligible importance, indicating little contribution to the model's predictions. The hyperparameters create a moderately shallow ensemble of trees with controlled learning, balancing the ability to capture nonlinear relationships and interactions with the need to avoid overfitting. Overall, the Gradient Boosting model reinforces the primary influence of smoking, along with smaller contributions from BMI and age, while other predictors have minimal impact.

Estimated Importances by Tree Models

Below is a summary of the tree models feature importances:

Feature	Regression Tree Importance	Random Forest Importance	Gradient Boosting Importance
Smoker	0.6931	0.6624	0.8188
BMI	0.1793	0.1890	0.0979
Age	0.1156	0.1265	0.0432
Children	0.0063	0.0110	0.0132
Region_Northwest	0.0034	0.0014	0.0051
Avg_Region_Wage	0.0022	0.0041	0.0089
Sex	0.0	0.0026	0.0059
Region_Northeast	0.0	0.0013	0.0
Region_Southeast	0.0	0.0010	0.0
Region_Southwest	0.0	0.0007	0.0071

Table 7. Summary of Tree Models Feature Importances

The feature importance comparison across the three tree-based models reveals a consistent level of importance of predictors. Smoking emerges as the most important in all models, accounting for 66–82% of importance, followed by BMI and age as the next most important features. Children, regional indicators, average regional wage, and sex contribute minimally across models, with some features set to effectively zero in certain models, highlighting their limited predictive value. Gradient Boosting places even greater emphasis on smoking compared to the Regression Tree and Random Forest, while BMI and age receive slightly less importance. Overall, all three tree-based models consistently identify lifestyle and demographic factors, particularly smoking, BMI, and age, as the key determinants of the outcome, whereas geographic and other minor features play only a secondary role.

Neural Network

A Multi-Layer Perceptron (MLP) is a feed-forward neural network consisting of an input layer, one or more hidden layers, and an output layer. Non-linear activation functions allow it to learn complex, non-linear relationships between features and the target. Neural networks are flexible and powerful but require careful tuning of architecture, learning rate, and regularization to avoid overfitting (Goodfellow et al., 2016).

The neural network model was implemented using an MLPRegressor with the tuned hyperparameters of one hidden layer of 50 neurons, ReLU activation, a learning rate of 0.01, and L2 regularization (alpha = 0.01). This allows the network to capture nonlinear relationships between the predictors and the outcome while controlling for overfitting. Unlike linear and tree-based models, the neural network does not produce coefficients or feature importances, so the relative contribution of individual predictors is not directly interpretable. However, based on model behavior and comparisons with other approaches, smoking, BMI, and age remain the primary drivers of the outcome. The single hidden layer provides sufficient capacity to model interactions among predictors, while the regularization helps maintain stable and generalizable predictions. Overall, the neural network confirms the importance of lifestyle and demographic factors while offering a flexible, nonlinear modeling approach compared to traditional regression and tree-based methods.

Model Performance Results

Here is a comparison of the predictive performance of all models evaluated in this analysis. The table below summarizes key metrics for each model, allowing for a clear assessment of how well each approach captures the relationships between the predictors and the outcome. These results highlight differences in accuracy and the ability of various modeling techniques – linear, regularized, tree-based, and neural networks – to generalize to unseen data.

Model	MSE	RMSE	R^2	CV Avg MSE	CV Avg RMSE
Linear Regression	33596915.8514	5796.2847	0.7836	37737178.5615	6123.3538
Ridge Regression	33604444.6417	5796.9341	0.7835	37736990.5916	6143.0441
LASSO	33720780.6302	5806.9597	0.7828	37704806.1866	6140.4239
Elastic Net	33623124.4151	5798.5450	0.7834	37728486.9520	6142.3519
Regression Tree	20442765.1582	4521.3676	0.8683	23118377.3289	4808.1574
Random Forest	18976221.6345	4356.1705	0.8778	21864327.8023	4675.9307
Gradient Boosting	18300366.8781	4277.8928	0.8821	20665273.2421	4545.9073
Neural Network	20511600.6610	4528.9730	0.8680	24368068.2304	4936.4024

Table 8. Summary of Model Predictive Accuracy

The table above summarizes the predictive performance of all models using several evaluation metrics. Mean Squared Error (MSE) measures the average squared difference between the predicted and actual values, with lower values indicating better predictive accuracy. Root Mean Squared Error (RMSE) is the square root of MSE, providing an error metric on the same scale as the outcome, making it easier to interpret. R^2 indicates the proportion of variance in the outcome explained by the model, with values closer to 1 reflecting stronger predictive performance. The cross-validated averages (CV Avg MSE and CV Avg RMSE) show the model's performance across multiple training/testing splits, providing a measure of generalizability and robustness.

Linear and regularized regression models (Ridge, LASSO, Elastic Net) perform similarly, explaining a substantial portion of the variance but with higher prediction errors. Tree-based models – Regression Tree, Random Forest, and Gradient Boosting – achieve lower errors and higher R^2 values. In particular, Gradient Boosting has a Cross Validated RMSE of approximately 4,546, meaning that, on average, its predictions deviate from the actual outcome by about $4,546. This indicates that the model can provide accurate and practically useful estimates for individual predictions. The Neural Network performs comparably to the Regression Tree but is less interpretable. Based on these metrics, Gradient Boosting is the recommended model, as it delivers the most accurate predictions while capturing complex, nonlinear relationships in the data.

Conclusion

Among the models tested, Gradient Boosting was the most accurate for predicting individual medical insurance costs, achieving the lowest RMSE and highest R^2. Tree-based methods, in general, outperformed linear and regularized regression models by capturing complex, non-linear relationships and interactions among features. For someone looking to estimate their own insurance costs, Gradient Boosting would provide the most reliable predictions, though the linear models could be used for more interpretable results while still maintaining reasonable accuracy. Across all models, smoking status, age, BMI, and the number of children consistently stood out as the strongest predictors of healthcare charges.

Overall, this project demonstrates that machine learning can effectively model healthcare costs using demographic, lifestyle, and regional features. The analysis confirms that certain personal characteristics have a large impact on insurance costs, while other factors, such as sex or regional economic indicators, play a smaller role. By combining accurate predictive modeling with realistic data verification against ACA benchmarks, the study highlights the potential for data-driven tools to support insurers in setting fair premiums and help individuals make informed decisions about their healthcare coverage.

Sources

Awad, M., & Khanna, R. (2015). *Machine learning.* In *Efficient learning machines* (pp. 1–18). https://doi.org/10.1007/978-1-4302-5990-9_1

BEA. (n.d.). *BEA interactive data application.* https://apps.bea.gov/itable/?ReqID=70&step=1&_gl=1

Breiman, L. (2001). *Random forests.* https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Center on Budget and Policy Priorities. (2024, April 10). *Building on the Affordable Care Act: Strategies to address marketplace enrollees' cost challenges.* https://www.cbpp.org/research/health/building-on-the-affordable-care-act-strategies-to-address-marketplace-enrollees

Collins, S., Roy, S., & Masitha, R. (2023, October 26). *Paying for it: How Health Care Costs and Medical Debt Are Making Americans Sicker and Poorer*. The Commonwealth Fund. https://www.commonwealthfund.org/publications/surveys/2023/oct/paying-for-it-costs-debt-americans-sicker-poorer-2023-affordability-survey

Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press. https://www.deeplearningbook.org/

HealthMarkets. (2025). *Health insurance quotes by age, region, and coverage*. https://shop.healthmarkets.com/en/

Hoerl, A. E., & Kennard, R. W. (2000). *Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 42*(1), 80. https://doi.org/10.2307/1271436

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *Linear regression.* In *Springer texts in statistics* (pp. 59–128). https://doi.org/10.1007/978-1-0716-1418-1_3

LaPick, M. (2022, January 3). *4 in 10 enrollees would go uninsured without Obamacare: Survey. HealthCareInsider.com.* https://healthcareinsider.com/affordable-care-act-snapshot-survey-2022

Mosap Abdel-Ghany. (2025). *Medical insurance cost dataset.* Kaggle. https://www.kaggle.com/datasets/mosapabdelghany/medical-insurance-cost-dataset/data

Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

USAFacts. (2025, January 17). *The Affordable Care Act and the data: Who is insured and who isn’t.* https://usafacts.org/articles/affordable-care-act-and-data-who-insured-and-who-isnt/

Zou, H., & Hastie, T. (2005). *Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x