**URS project on Kaggle medical insurance dataset:**

1. Kaggle Medical insurance dataset

Source: <https://www.kaggle.com/datasets/mosapabdelghany/medical-insurance-cost-dataset>

Variables:

* age: Age of primary beneficiary (int)
* sex: Gender of beneficiary (male, female)
* bmi: Body Mass Index, a measure of body fat based on height and weight (float)
* children: Number of children covered by health insurance (int)
* smoker: Smoking status of the beneficiary (yes, no)
* region: Residential region in the US (northeast, northwest, southeast, southwest)
* charges: Medical insurance cost billed to the beneficiary (float)
  1. **Generate scatter plot between charges and age/sex/bmi/children/smoker/region**
     1. **Scatter plot between charges and age**

Since the variables sex, smoker, and region are categorical rather than numerical, their data distributions require different visualization techniques. To explore these distributions, I initially used a swarm plot, which spreads individual data points horizontally to avoid overlap. This type of plot is particularly effective for displaying the distribution of numerical values across categorical groups. Interestingly, the overall shape of the swarm plot resembles that of a violin plot, which led me to further investigate the dataset using violin plots. Violin plots not only show the distribution density but also provide a more compact and informative summary of the data across categories.

A graph of a number of blue dots

AI-generated content may be incorrect.

Three distinct linear patterns between age and three charges range.

Figure 1 Scatter Plot between age and charges

*A graph of a graph of a number of points

AI-generated content may be incorrect.*

Figure 2 Segmented linear regression line between age and chages

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Segment | N | Linear regression Equation |  | Pearson r |
| Low charge  (0-17,000) | 1007 | Charges = -1803.15 + 235.71 \* Age | 0.665 | 0.816 |
| Medium charge  (15,000-32,000) | 203 | Charges = 15894.96 + 147.19 \* Age | 0.283 | 0.532 |
| High Charge (31,000-60,000) | 153 | Charges = 31783.55 + 227.14 \* Age | 0.394 | 0.628 |

Table 1 Summarized correlation tabel for linear regression between age and charge

To improve the best fit line accuracy, separting each segment into individual graph is essential to have a better focus.

1. **Low charge (0 – 17,000)**

Linear regression is not enough to fit the regerssion trend between age and low charge. Therefore, a polynomial regression should be applied.

Without making any procedures of cleaning data, the polynomial regression between age and charge is:

A graph with a red line

AI-generated content may be incorrect.

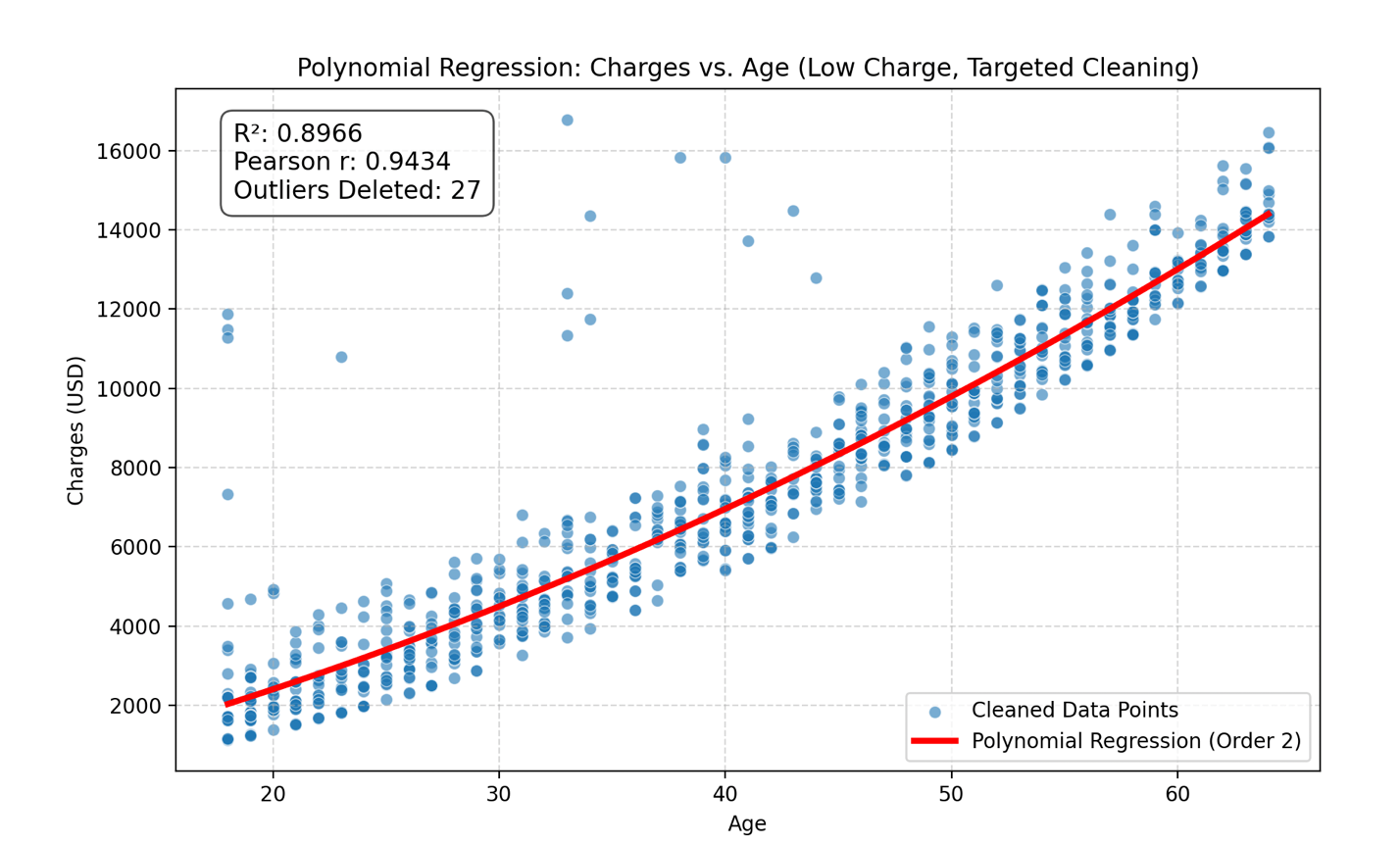
Figure 3 Polynomial regression between age and charge for low charge segment

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Segment | N | Polynomial regression Equation |  | Pearson r |
| Low charge  (0-17,000) | 1007 |  | 0.6817 | 0.8155 |

Table 2 Correlationa result for polynomial regression between age and low charge

Intuitively, the starting point of polynomial regression line should be located at the point around 2,100 to 2,600. The reason why current regression linear starts from 3,700 might be hugely influenced by noise data points on the left corner side. (age < 30 && charge >= 12,000). Therefore, a procedure of cleaning data is essential to be considered.

By using standard 1.5 \* IQR on both age and charges, no extreme outliers distorted the boundaries. Secondly, correspond to previous observation, adding a highly specific filter based on the visual scatter plot is essential: Age < 30 and Charges > 12,000.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Segment | N | Polynomial regression Equation |  | Pearson r |
| Low charge  (0-17,000) | 980 |  | 0.8966 | 0.9434 |

Figure 4 Correlationa result for polynomial regression between age and low charge after stricter data cleaning

1. **Meidum charge (15,000 – 32,000)**

According to the result in table 1, both the coefficient of determination and Pearson r values are low, which indicate low correlation between two variables. Therefore, to further improve the accuracy and relationship between these two variables, polynomial regression analysis is necessary to be included. Since, the data scattering pattern is not as clear as low charge one, incorporating various regression methods are necessary. In the following discussion, quadratic, cubic, and power regression.

A graph with colored lines and numbers

AI-generated content may be incorrect.

Figure 5 Non-Liner regression for medium charge segmented data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | N  (after cleaning) | Equation |  | Pearson r |
| Cubic polynomial | 203 |  | 0.3368 | 0.5322 |
| Quadratic polynomial | 203 |  | 0.2832 |
| Power law | 203 |  | 0.2542 |

Table 3 Correlationa result for polynomial regression between age and medium charge

The Pearson value of 0.5322 indicates a moderate positive linear relationship between Age and Charges in the raw data. Older people in this segment tend to have slightly higher costs.

However, the low value of (even for the best non-linear model, Cubic Polynomial) means that our chosen predictor, Age, accounts for only about 33.7% of the total variance in charges. The remaining of the cost variation must be explained by factors not included in the model, which will be further discussed.

1. High charge ( 31,000 – 60,000)

A graph with a line and a line

AI-generated content may be incorrect.

Figure 6 Non-Liner regression for high charge segmented data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | N  (after cleaning) | Equation |  | Pearson r |
| Cubic polynomial | 203 |  | 0.4502 | 0.6687 |
| Quadratic polynomial | 203 |  | 0.4474 |
| Power law | 203 |  | 0.4376 |

Table 4 Correlationa result for polynomial regression between age and high charge

The Pearson r value of 0.6687 indicates a moderately strong, positive linear correlation between Age and Charges in the raw data. However, the best-fitting still shows that more than half (55%) of the cost variation in this high-charge group is due to factors other than age.

* + 1. **Scatter plot between charges and bmi**

A diagram of blue dots

AI-generated content may be incorrect.

Below BMI 30: increasing variance

Below BMI 30: increasing variance

Below BMI 30: clustered

BMI Index

**Overweight**

**Normal weight**

**Extremely obese**

**Underweight**

**Obese**

Figure 7 Scatter plot between bmi and charges

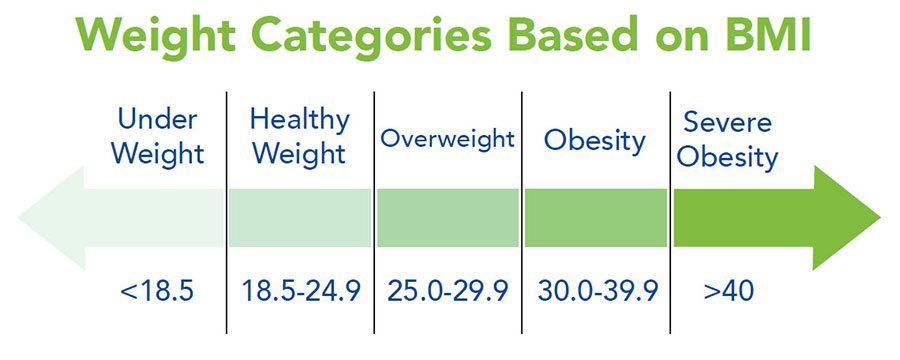


Figure 8 Weight categories based on BMI

1. **The BMI Threshold at 30**

The most striking feature is the dramatic change in the data's pattern around (the clinical definition of obesity).

* Below : Charges are relatively low, tightly clustered, and primarily follow a single, gentle upward trend.
* Above (Obesity): The costs suddenly bifurcate (split). A large segment of the population remains on the lower charge trend, but a new, distinct population appears, showing a steep increase in maximum and average charges.

1. **Increased Variance**

Beyond , the spread or vertical distance between the dots drastically increases.

* This means that for someone with a high BMI, we are much less certain of their exact charge. They could be anywhere from the low-cost baseline to the highest-cost bracket.

1. **Further investigation**

* Since BMI is a comprehensive indicator of body weight that also reflects factors such as age, sex, and other physiological characteristics, its relationship with medical charges is complex. In this project, I focus on the variables sex, age, and smoking status to better understand this relationship. The association between BMI and charges cannot be adequately captured by a simple linear or polynomial regression model. Instead, a more nuanced approach is required, potentially involving multiple regression, piecewise regression, or other advanced statistical techniques, to account for the interplay of multiple influencing factors.
  + 1. **Scatter plot between charges and children**

A graph of blue dots

AI-generated content may be incorrect.

Figure 9 Scatter plot between children and charges

* 1. A counter-intuitive observation: Increasing number of children covered by medical health insurance doesn’t correlate with higher charges. Plus, there is a huge drop from 0-3 children to 4-5 children. Thus, only focusing on two variables at a time is not enough. By combining factors of age, smoking status, and age together, we could conclude this table:

|  |  |  |  |
| --- | --- | --- | --- |
| Children Group | Typical Age Profile | Highest Risk Individuals | Maximum Charge |
| 0 Children | Bimodal: Includes very young adults (low risk) AND very old adults (highest risk). | Oldest smokers (50-64 years). | Very High |
| 1-3 Children | Primarily middle-aged adults (30s-50s) who are still accumulating medical issues. | Middle-aged smokers (30s-50s). | Very High |
| 4-5 Children | Tends to be dominated by younger to middle-aged adults (20s-40s) due to the physiological constraints of having many children. | Fewer/No older adults (50s-64s). | Lower (Capped) |

Table 5 Multiple factors are incorporated into the analysis of children and charges

* + 1. **Scatter plot between charges and sex**

The distribution for both male and female charges is heavily skewed, meaning most people have low charges, but there’s a long tail of very high charges. The density of points is highest in the to range for both sexes. The bulk of the points for male and female charges overlap almost entirely. It’s hard to visually identify a clear separation or group where one sex dominates the other.

A graph of a person and person

AI-generated content may be incorrect.

Figure 10 Swarm plot of sex and charges

The "violins" for both male and female charges are extremely wide at the bottom (low charges) and narrow quickly toward the top (high charges). This is definitive evidence of the high degree of positive skewness in insurance charges: most people incur low costs.

The most critical finding is that the shape and spread of the male and female violins are virtually identical. This confirms what was statistically observed: while the mean charges are statistically different by , the underlying risk distribution is not substantially different based on sex alone.

A diagram of two people

AI-generated content may be incorrect.

Figure 11 Violin plot of the distribution between sex and charges

The visual evidence from the violin plot strongly supports the idea that sex is a weak predictor of charges.

* + 1. **Swarm plot between smoking condition and charges**

Once an individual is a smoker, their insurance charge immediately jumps to a minimum of 15,000(with very few exceptions), regardless of other minor factors. Charges for smokers exhibits a scatter and massive range ( from 15,000 to 63,000), which confirms that the oberal charge for smokers is significanty largers than for non-smokers.

A graph of a diagram

AI-generated content may be incorrect.

Non-smoker:

Data clustering at the range of below 10,000 – 15,000

Figure 12 Swarm plot of smoker and charges

The distribution of charges for non-smokers exhibits extreme positive skewness. The violin plot is characterized by its maximal width and height at the lowest charge bracket (approaching zero), confirming that the vast majority of non-smokers incur minimal annual medical expenses. The rapid narrowing of the violin as charges increase demonstrates a high concentration of probability mass at the low end.

This shape is indicative of a low baseline risk profile: the typical non-smoker incurs only routine preventative and minor medical costs. The long, narrow tail extending toward higher charges represents a small minority of non-smokers who required high-cost intervention (e.g., due to accidents or non-smoking related chronic conditions). Overall, this distribution is highly predictable, centering around a low median value.

The violin for the "Smoker: Yes" group is substantially wider overall and is laterally shifted, indicating a much higher mean and median charge. Crucially, the density remains significant and relatively consistent across the to range, demonstrating that high costs are the norm, not the exception, for this group.

This broad distribution confirms that once the Smoker variable has dictated a high baseline cost, the remaining variance in charges is explained by a multitude of other factors (such as age, BMI, and children) whose relative contributions are spread out across a wide range.

A diagram of a distribution of charge

AI-generated content may be incorrect.

Sudden decrease in the width of smoker when the charge is 30,000

Figure 13 Violin plot of the distribution between smoking condition and charge

Noticeably, there is a sudden decrease in the width of the smoker when the charge reaches to 30,000. Here’s the hypothesis that I made to try to analyze this threshold:

* 1. **Policy Thresholds or Product Differentiation:** The sharp drop supports the hypothesis that 30,000 may act as a major cost boundary tied to insurance product design. The high density below this level likely captures the costs associated with common high-risk interventions for smokers (e.g., standard heart disease treatment, common respiratory disease management). The individuals whose costs exceed this boundary may represent those who have entered a smaller, more specialized risk pool.

After a sudden decrease, the width increases again around 45,000 to 50,000. This likely represents a cost cluster linked to a very specific, extremely expensive set of medical interventions that primarily affect smokers. Also, The individuals in this secondary bump are likely those who combine smoker status with other severe, high-risk factors (e.g., extremely high BMI, multiple children, or old age) that accelerate or complicate their conditions.

* + 1. **Swamr plot between regionn and charge**

According to the swarm plot below, all regiosn show a high concentration of individuals in the low charge bracket ( below 10,000). While shapes look similar, the southeast region appears to have a slightly higher density of points in the 35,000 to 50,000 range.

A graph of different colored lines

AI-generated content may be incorrect.

Figure 14 Swarm plots of region and charges

By generating a violin plot to systematically visualize data distribution and probability, the southeast region consistently demonstrate the highest average costs. The wider density in the moderate-to-high charge range suggests that costs for common medical procedures, or the average health profile of the region's population, systematically results in higher claims. This difference is statistically valuable, even if it accounts for a small percentage of the total variance, as it reflects genuine regional variations in healthcare pricing and utilization.

A diagram of different colored shapes

AI-generated content may be incorrect.

Figure 15 Violin plot of the distribution between region and charges

Based on previous analysis, analyzing relationships soely between two variables is not enough for the prediction of the charge. Therefore, a systematic and comprehensive preditive model should be applied. Before constructing a model, a correlation heatmap is useful in visualize linear correlation between various factors. Noticeably, in the raw data, answers given by smoking condition are descriptive but not numerical. A setp of converting descriptive variables, like “yes” or “no”, into numerical one is necessary because a mathematical model only understand numbers.

Here, I will use basic mutiple linear regression (MLR) to predict a continuous outcome (Charges) based on several predictor variables simultaneously (Age, BMI, Smoker, Region, etc.):

Where:

is the intercept;

are the coefficients (weight) for the variables;

are the coefficients for the regional dummy variables;

is the error term.

In regression analysis, dummy variables are often applied to convert descriptive variable into numerical data. A dummy variable can only takes two number, either 1 or 0. Take smoking condition as an example.

|  |  |  |
| --- | --- | --- |
| Original data | New Column: Smoker\_yes | Meaning to the method |
| Smoker: Yes | 1 | The feature is present/ON |
| Smoker: No | 0 | The feature is absent/OFF |

Table 6 Dummy variable example for smoking condition

Therefore:

if the person is a smoker (1), the model adds the full weight () to the charge.

If the person is a non-smoker (0), the model adds zero (0 ) to the charge.

For the ‘Region’ variable (Northeast, Northwest, Southeast, Southwest), creating N-1 columns and establish a matrix format:

|  |  |  |  |
| --- | --- | --- | --- |
| Original Region | Northeast Dummy | Northwest Dummy | Southeast Dummy |
| Northeast | 1 | 0 | 0 |
| Northwest | 0 | 1 | 0 |
| Southeast | 0 | 0 | 1 |
| Southwest | 0 | 0 | 0 |

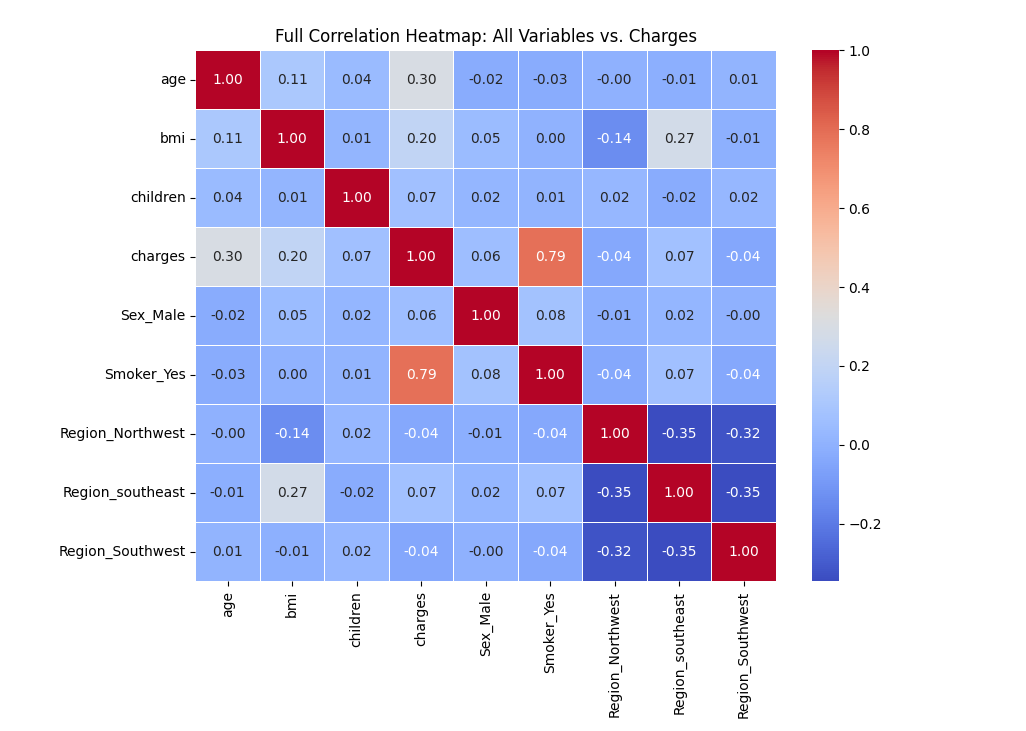


Figure 16 Correlation heatmap for all variables vs. charges

MLR model:

Charges = -11931.22 + 256.98 \* age + 337.09 \* bmi + 425.28 \* children - 18.59 \* Sex\_Male + 23651.13 \* Smoker\_Yes - 370.68 \* Region\_Northwest - 657.86 \* Region\_Southeast - 809.80 \* Region\_Southwest

R-squared (R²) on Test Set: 0.7836

Trying to test my MLR model accuracy:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Profile | Age | Sex | BMI | Children | Smoker | Region | Actual Charge | Predicted Charge | Error ($) | Abs. Error (%) |
| 1 | 19 | female | 27.9 | 0 | yes | southwest | $16,884.92 | $25,197.54 | 8,312.62 | 49.20% |
| 2 | 18 | male | 33.77 | 1 | no | southeast | $1,725.55 | $3,826.78 | 2,101.23 | 121.80% |
| 3 | 28 | male | 33 | 3 | no | southeast | $4,449.46 | $6,987.58 | 2,538.12 | 57.00% |
| 4 | 33 | male | 22.705 | 0 | no | northwest | $21,984.47 | $3,813.48 | -18,170.99 | 82.70% |
| 5 | 32 | male | 28.88 | 0 | no | northwest | $3,866.86 | $5,638.03 | 1,771.17 | 45.80% |
| 6 | 31 | female | 25.74 | 0 | no | southeast | $3,756.62 | $4,054.00 | 297.38 | 7.90% |
| 7 | 46 | female | 33.44 | 1 | no | southeast | $8,240.59 | $10,929.57 | 2,688.98 | 32.60% |
| 8 | 37 | female | 27.74 | 3 | no | northwest | $7,281.51 | $7,833.08 | 551.57 | 7.60% |
| 9 | 37 | male | 29.83 | 2 | no | northeast | $6,406.41 | $8,464.40 | 2,057.99 | 32.10% |
| 10 | 60 | female | 25.84 | 0 | no | northwest | $28,923.14 | $11,827.31 | -17,095.83 | 59.10% |

1. Building ANN model to improve accuracy of predicting insurance cost

ANN is computational model inspired by the structure of human brains. Similar to the struture of brains, ANN also composed of neurons that process information and learn patter form data. A basic ANN structure comprises of three types of layer: input, hidden, and output. Hidden layers are the most complex one that refered as blackbox of ANN, which has the ability to learn complex, non-linear relatioships between unput variables and output predictions without being programmed with specific rules. Instead, ANN adjusts its internal parameters through a process called training.

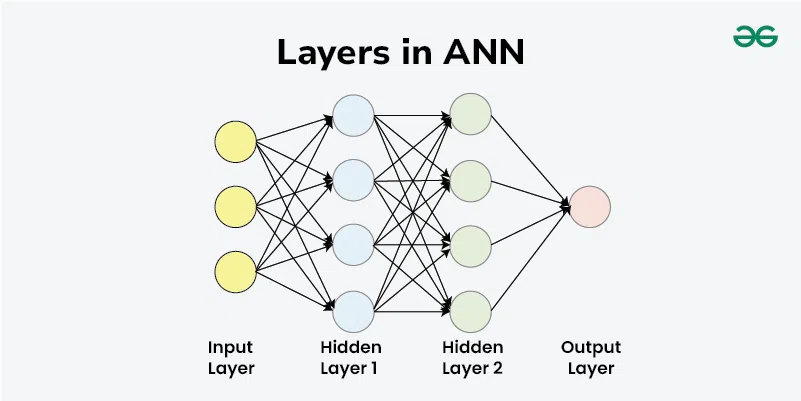


Figure 17 Basic structure of ANN

The operation of ANN goes through following steps:

1. Forward propagation:
   1. When data flows form unput layer to hidden layer, neurons in the first hidden layer receives weighted inputs.
   2. Hidden layer sums weighted inputs and a bias term.
   3. It applies an activation function to employ non-linearly alogrithm.
   4. The results from first hidden layer are passed to neurons in the next layer.
2. Activation function:
   1. Determining whether a neuron should be activated could be achieved by following methods:
      1. ReLu (Rectified Linear Unit): f(x) = max(0,x) ;By doing so, positive value is unchanged while setting negative one to zero, which makes ANN data propagation efficiently.
      2. Linear activation f(x) = x; It is used in the output layer for regression tasks, which allows the network to predict any continuous value.
3. Loss calculation:
   1. At each layer, when data reaches from input layer to the output one, biases are aded up. Mean Squared Error (MSE) is used to measure the average squared difference between predictions and actual costs.
4. Backpropagation:
   1. The network calculates how much each weight contributed to the error and adjusts the weights in the opposite direction to reduce the error. This process uses gradient descent.
5. Learning through iterations:
   1. Steps 1-4 are repeated for multiple epochs. With each epoch, the network's predictions become more accurate as the weights are optimized.

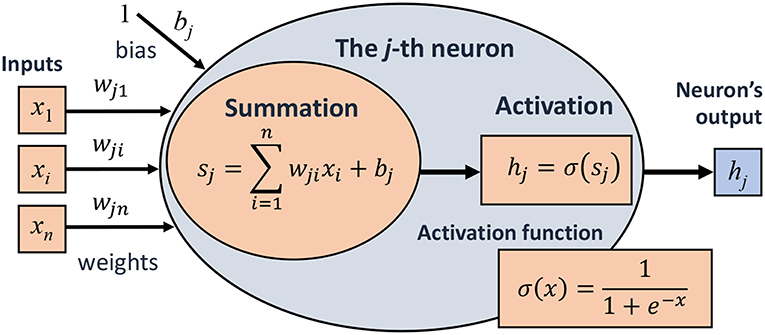


Figure 18 ANN mathematical representation

Figure 19 Data processing flowchart

Figure 20 Neural Network Archietcture

Figure 21 Model Training Process

A graph of a diagram

AI-generated content may be incorrect.

Figure 22 Actual and predicted insurance costs comparson with the line of perfect prediction

This diagram indicates that the model explains 83.49% of the variance in insurance costs, which means this model successfully captures the major patterns and relationships in the data. By calculating, the root mean squared error (RMSE) of $ 5063.29 means predictions deviate from actual costs by approximately this amount.

= $5063.29

The value or RMSW seems intimidating at first, but the interpretation of this RMSW should be interpretated in the context. The insurance charges range from about $1,100 to over $63,000. In other words, the potential variation in actual costs is large, so some prediction error is unavoidable. In this context, an RMSE of ~$5,000 means that the model’s predictions are typically within the normal spread of insurance costs for most policyholders.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 23 Distribution of absolute and percentage error

The left of the error distribution histogram shows how far each prediction is from the true value. Most predictions have an absolute error under $4,000, with a few extreme values reaching higher error levels. The red dashed line marks the mean absolute error ($3355.92), meaning the typical prediction is off by this much. The distribution is skewed right, indicating most errors are small, but a few large differences exist (mainly for outlier or rarely-seen cases).

The right of the same figure displays percentage errors. The majority of cases have errors below 50%, but there is a long tail with a few much higher percentages. This is expected in datasets with a wide range of target values: small actual values can have high percentage errors even if the absolute error is low. The mean percentage error (42.29%) reflects both many accurate predictions and a handful of more significant deviations.

A graph of training and training

AI-generated content may be incorrect.

model isn’t improving much more with each additional epoch

Both lines decrease rapidly in the first 20 epochs, indicating fast initial learning. For model loss, After epoch 40, the curves flatten, suggesting the model has largely converged. The training and validation losses track closely together, indicating the model is not overfitting: it generalizes well to unseen data. MAE decreases from approximately $13,500 to $3,300, representing a 74% improvement in prediction accuracy. The parallel movement of both curves confirms good generalization.

Q: If the MAE decreases quickly after 30th epoch, does this mean the training of model should stopped around 60 or earlier to save compuational efficiency?

Q: The convergence of training and validation loss represents? I tried to add extra layers but this results in a divergence of tranining loss and validation loss.

Q: MAE (~$3,356) is lower than RMSE (~$5,063) both are concepts relating to the discussion of errors but the huge difference between them suggest?

Q: last week, we’ve mentioned the interaction between variables might work. But how to find which two or even more variables should be paired up? Like smoker \* age \* BMI in this form?