HW 1: Project

Kevin Lin

1/16/2026

1

I chose the medical insurance dataset, as healthcare costs and insurance coverage in general is an important issue unique to the United States. While this rudimentary dataset only contains a few features and definitely won't make any groundbreaking predictions, it is a good starting point to explore insurance costs and how they vary across different demographics. In my undergraduate senior year, I worked with a dataset from MIMIC-III, which contained de-identified health data from critical care patients. I used that dataset to predict hospital readmission rates with an AUC of 0.8 matching a study done by Google. Hopefully I can find some interesting trends in this dataset as well, or uncover some biases in insurance coverage and costs.

2

There is no direct research article associated with this dataset. The dataset was curated specifically for educational purposes for Brett Lantz's book *Machine Learning with R* from PACKT Publishing. However, Lantz acknowledges that the dataset was originally sourced "using demographic statistics from the U.S. Census Bureau, and thus approximately reflect[s] real-world conditions." The variables in the dataset are as follows:

Data	Description	Variable Type	Units
age	age of primary benificiary, excluding anyone 64+	integer	years
sex	policy holder's gender (male or female)	binary string	N/A
bmi	policy holder's body mass index	float	kg/m
children	number of children / dependents covered by the plan	integer	# of children
smoker	whether the insured regularly smokes (yes or no)	binary string	N/A
region	beneficiary's place of residence in the US, divided into 4 regions (northeast, northwest, southeast, southwest)	categorical string	N/A
charges	insurance charge amount in dollars	float	$USD