MEDICAL COST PREDICTION

REGRESSION

MACHINE LEARNING PORTFOLIO

DEREK RODRIGO SÁNCHEZ SEGUAME

1.Task Definition

2.EDA

3.Model and Results

01

Task Definition

PROBLEM STATEMENT | MEDICAL PREMIUM CHARGE

SaludVital Analytics is the advanced-analytics division of a **leading Mexican health insurer** that underwrites policies for individuals and families across middle- and upper-income segments. Leveraging an extensive distribution network and product portfolio, the insurer **has achieved strong premium growth** but now **faces claim costs that are rising faster than actuarial forecasts**.

To address this, we **will implement a machine-learning–powered predictive engine** using the Medical Charges dataset, which includes age, sex, region, BMI, smoking status, number of dependents and actual annual medical charges. By applying sophisticated regression models and robust feature selection, **the project will refine premium pricing to align rates with predicted claim costs**. It will also optimize reserve adequacy and regulatory capital planning through more precise loss forecasts. Additionally, the engine will identify high-cost segments for targeted wellness programs and personalized member engagement—**enhancing underwriting profitability and strengthening the insurer's competitive positioning in Mexico's health-insurance market.**

MEDICINE AND DATA | MEDICAL MANAGEMENT POWERED BY PREDICTIVE MODELS

Business Context

- The client, **SaludVital** is the analytics division of a Mexican insurer
- Their target market are families across **middle and upper income** segments

The Problem

- Due to the insurer's extensive distribution network and portfolio, they've **achived significant premium growth**
- The downside: they are facing **claims costs higher than expected**

Solution

- We will implement a **Machine Learning Model** to refine premium predictions to **ensure fair pricing and a sustainable business model**
- By applying an advanced **Regression Model**, premium costs will be aligned with claim costs

02

Exploratory Data Analysis

OVERVIEW | FEW, BUT POWERFUL PREDICTORS

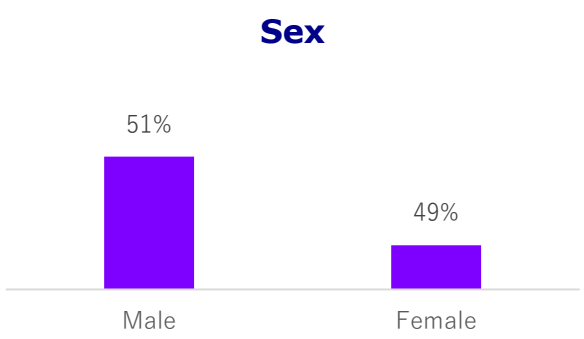
of Features | By data type

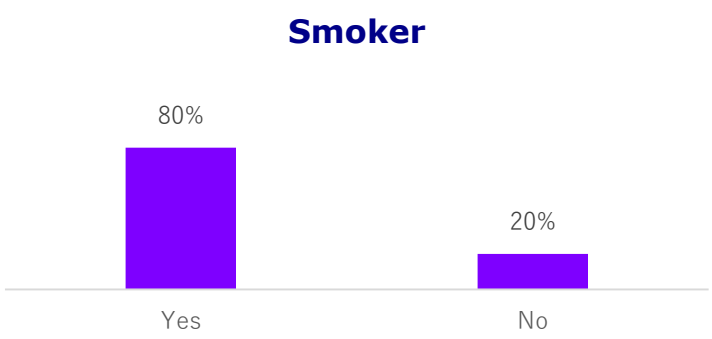
	#Features
Numerical	3
Categorical	3
Total	**6**

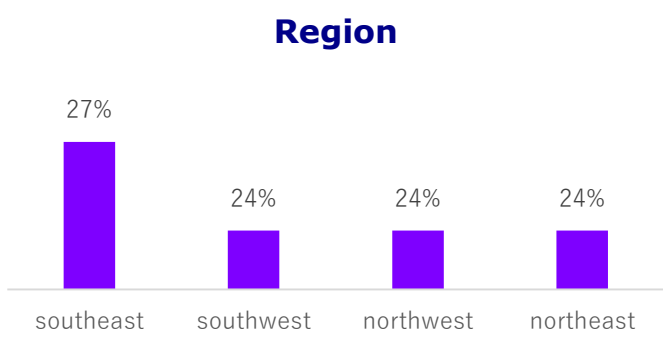
# of Categorical Features | By # of unique observations

	Two Unique Obs	Four Unique Obs	Five+ Unique Obs
#Features	2	1	0

Categorical Features Distributions



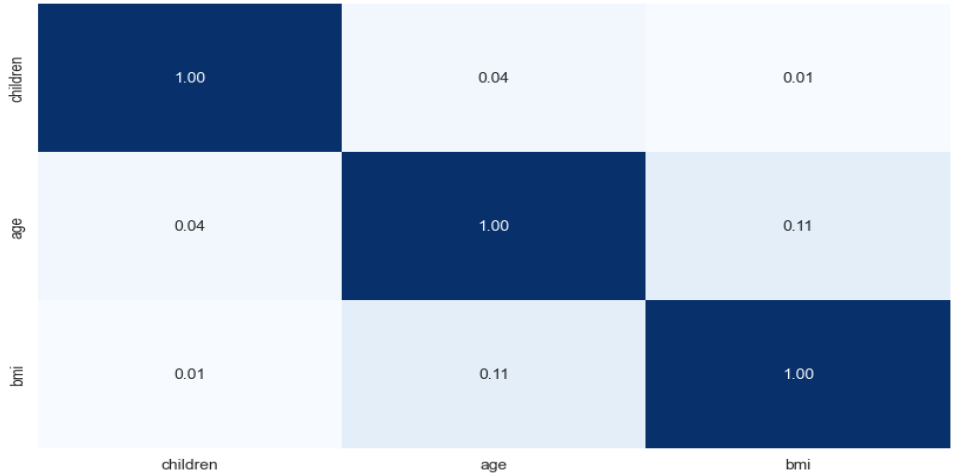




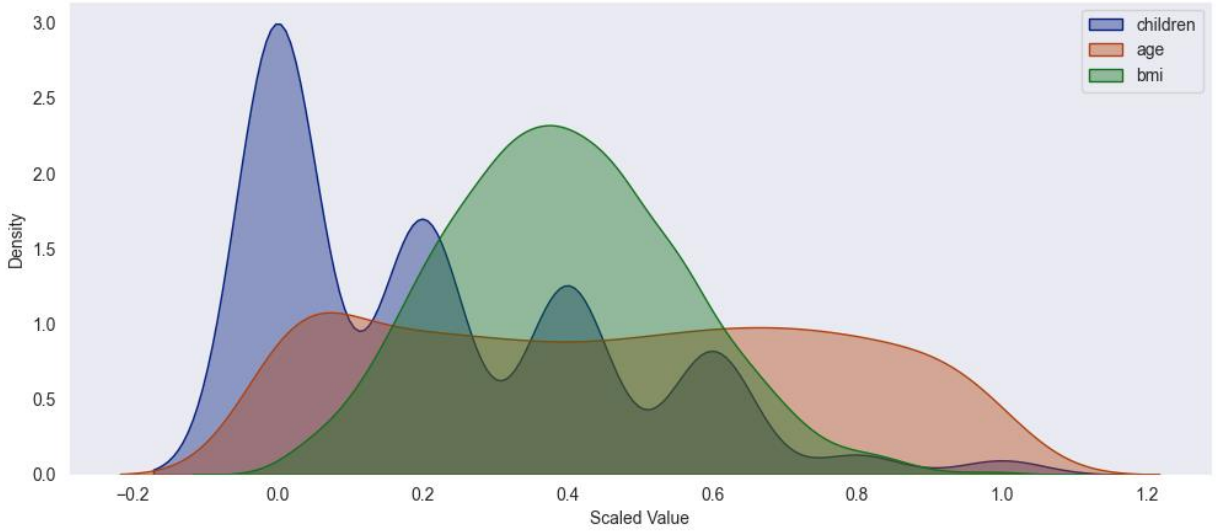
Note:

NUMERICAL VARIABLES | HEATMAP, DISTRIBUTIONS AND BOXPLOTS

Correlation Matrix Heatmap

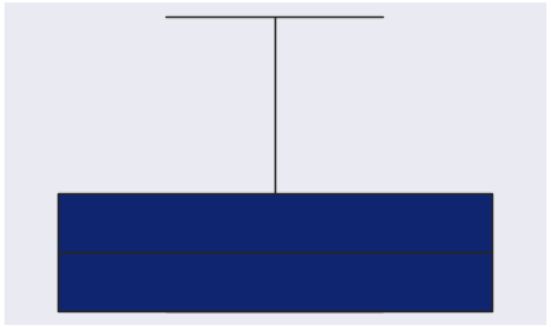


Distribution Graph

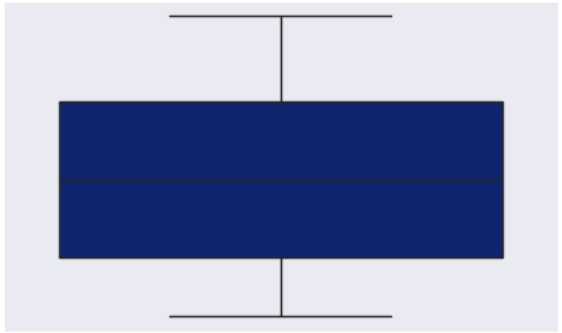


Box plot | by feature

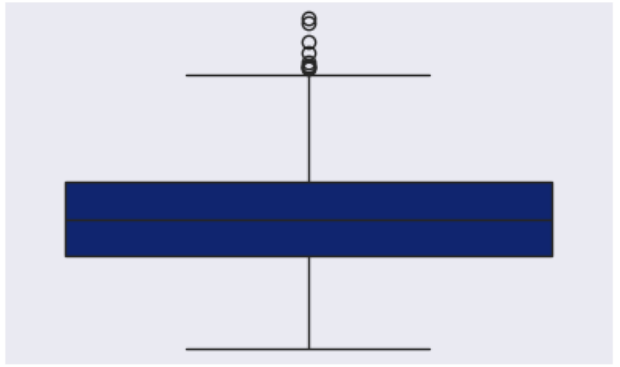
Children



Age



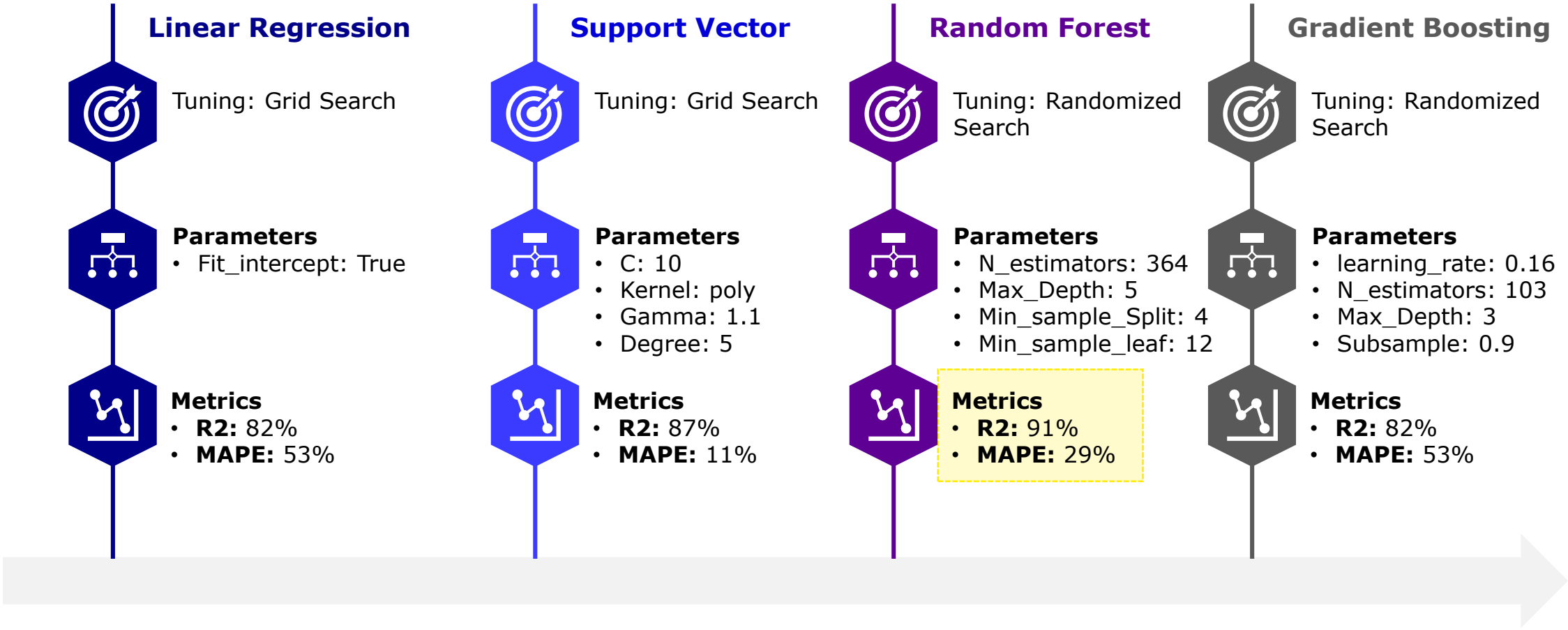
BMI



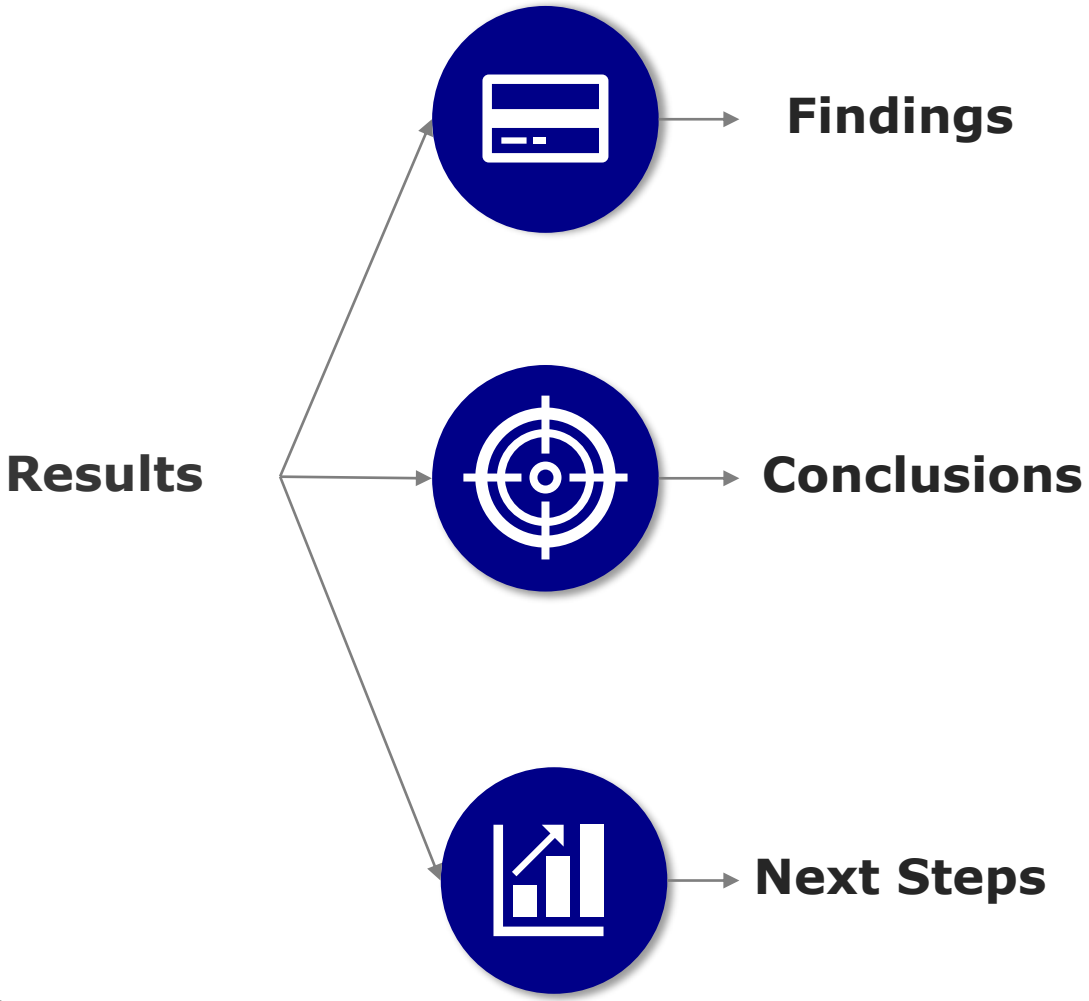
03

Model Selection and Results

MODEL SELECTION | RANDOM FOREST PERFORMED THE BEST IN R2 SCORE; OPTIMAL FOR CLAIM PREDICTION



DATA SOLUTIONS FOR OPTIMAL MANGAMENT | USING MACHINE LEARNING TO FIND THE OPTIMAL INSURANCE PREMIUM



- Costs are influenced mainly by **BMI, Age and Number of Children of the patient**
- **Male smokers from the southeast** have the most expenisve costs
- **Male non-smokers from the southeast** have the least expensive costs

- **Random Forest** is the model that best captures the underlying relationships in the data to predict cost. **It achieved a R2 Score of 91%**
- Given that the dataset uses few features for prediction, **R2 is a realistic indicator of goodness of fit**

- **Deploy the model** and monitor to achieve a better **distribution of siniestrality and a fair pricing scheme for the customers**
- The provisions of the insurer will have more accurate estimates, **positively impacting P&L results**

Note: