Project: Insurance Pricing Prediction

This project explores factors influencing insurance costs and uses machine learning to predict charges based on customer demographics.

Objective

- Understand how different factors impact insurance pricing**.**
- Use Machine Learning (Random Forest) to predict insurance costs.
- Visualize key findings using Tableau for business insights.

Dataset

Source: Insurance data

Features:

- **age:** Age of the individual
- **bmi:** Body Mass Index
- **children:** Number of dependents
- **smoker:** Smoking status (yes/no)
- **region:** Residential area (northeast, northwest, southeast, southwest)
- **charges:** Target Variable (Total insurance cost)

Methodology

Data Pre-processing

- Converted categorical variables (smoker, region) into factors**.**
- Checked for missing values (none found).
- Normalized numeric variables (bmi, age) for machine learning.

Machine Learning Model (Random Forest)

- **Model Used:** random Forest
- **Key Parameters:** trees = 1000, mtry = 2
- **Performance Metrics:**
    - **RMSE (Root Mean Squared Error)**: Measures prediction accuracy.
    - **R^2 Score**: Checks how well the model explains variance in charges.

Feature Importance Analysis

- Used impurity-based importance scores from ranger.
- Smoking had the highest impact**,** followed by BMI and age.

Tableau Dashboard

Visualizations Included:

Feature Importance – Shows top predictors of insurance charges.
Actual vs. Predicted Charges – Evaluates ML model accuracy.
Average Charges by Category – Explores variations across age, BMI, and smoking status.
Average Charges by Region – Analyses geographical trends.

Dashboard Link:
https://public.tableau.com/app/profile/gayathri.parupalli/viz/Book1_17422302366810/Dashboard1

Key Takeaways

- **Smokers pay significantly higher insurance premiums** compared to non-smokers.
- **BMI and Age are major cost factors**—higher BMI & older age increase charges.
- **Region has little impact**—insurance costs are mostly standardized across locations.

R CODE:

```
library(tidyverse)
library(tidymodels)
library(corrplot)
library(vip)

# Check structure
glimpse(insurance)

# Summary
summary(insurance)

# Check missing values
sum(is.na(insurance))
# No missing values

# Visualize price distribution
ggplot(insurance, aes(x = charges)) +
  geom_histogram() +
  labs(title = "Distribution of Charges")

#Convert categorical variables as factors
insurance$smoker<- as.factor(insurance$smoker)
insurance$sex<- as.factor(insurance$sex)
insurance$region<- as.factor(insurance$region)

#linear regression to know the best predictors
model<- lm(charges~age+bmi+smoker+children+sex+region,data=insurance)

# Correlation matrix (numeric variables only)
insurance |>
select_if(is.numeric) |>
```

```
cor(use = "complete.obs") |>
  corrplot()

set.seed(123)
split <- initial_split(insurance, prop = 0.8)
train_data <- training(split)
test_data <- testing(split)

recipe <- recipe(charges ~ ., data = train_data) |>
  step_normalize(all_numeric_predictors()) |>
  step_dummy(all_nominal_predictors())

# Define Random Forest(rf) model
rf_model <- rand_forest(mtry=2,mode = "regression",trees = 1000) |>
  set_engine("ranger",importance= "impurity")
summary(rf_model)

# Create workflow
rf_workflow <- workflow() |>
  add_recipe(recipe) |>
  add_model(rf_model)


# Train and evaluate
rf_fit <- rf_workflow |>
fit(train_data)
rf_pred <- predict(rf_fit, test_data) |>
bind_cols(test_data)

metrics(rf_pred, truth = charges, estimate = .pred)

rf_pred |>
  mutate(difference= charges-.pred) |>
  View()

# Extract feature importance scores
importance_values <- rf_fit |>
  extract_fit_engine() |>
  ranger::importance()

# Print feature importance scores
View(importance_values)

importance_df <- data.frame(Feature = names(importance_values),
                            Importance = importance_values) |>
  arrange(desc(Importance))

# Save to CSV
write.csv(importance_df, "feature_importance.csv", row.names = FALSE)
write.csv(rf_pred, "predicted_vs_actual.csv", row.names = FALSE)
```