"The Impact of Smoking on Medical Insurance Charges: A Machine Learning Approach Using Random Forest Regression"

This project explores how smoking status affects average medical insurance charges, even when controlling for age and BMI. Using a dataset of 1,338 individuals, I built a machine learning model to predict insurance costs based on whether someone smokes, their age, and their BMI. I used a Random Forest Regressor, which is a type of ensemble learning algorithm known for its accuracy and ability to handle complex relationships between features. To ensure fair testing, I applied 5-fold cross-validation and evaluated the model using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R² score.

The dataset came from an open-source Kaggle repository and includes personal and lifestyle information like age, BMI, smoking status, and insurance charges. Categorical variables were encoded for the model, and minimal cleaning was required due to the dataset’s high quality. The target variable was insurance charges. While the dataset did not include clinical or income information, it was still effective for testing the relationship between smoking and medical costs.

Results showed that smoking status was the most influential factor in predicting insurance charges. Smokers were often charged thousands of dollars more than non-smokers, even when they had similar age and BMI. The Random Forest model achieved strong performance, with an R² score of around 0.86, indicating it could explain a large portion of the variation in charges. This supports the hypothesis that smoking significantly drives up healthcare costs, beyond other health indicators.

AI can help address this issue by identifying individuals at risk due to smoking and guiding insurance companies to create fairer, more personalized plans. Health apps powered by machine learning can also support behavior change by giving users real-time feedback and health tracking. In the future, more advanced models and deeper datasets could improve prediction accuracy and help policymakers create more effective public health strategies.

**References:**

1. "Medical Cost Personal Dataset" – Kaggle. https://www.kaggle.com/datasets/mirichoi0218/insurance
2. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
3. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
4. Centers for Disease Control and Prevention (CDC). "Health Effects of Cigarette Smoking." <https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/effects_cig_smoking/index.htm>