Project Proposal : Classify Insurance Cost Tiers (Idea 1)

Abstract: This project classifies individuals into three tiers of insurance cost (low, medium, high) using demographic and health features. The objective is to build an interpretable multiclass classifier using a reframed version of the Medical Cost dataset. The result will help simulate tiered pricing in real-world insurance systems.

Introduction: Insurance pricing models benefit from tiering applicants based on risk. Instead of predicting exact charges, we categorize individuals into meaningful buckets. This project will present exploratory analysis, tier definition, model building, and error analysis across classes.

Motivation: Multiclass classification provides more actionable insights than simple regression. This method can inform pricing strategies and customer segmentation. The dataset is widely used and offers diverse input features for training a practical ML model.

Problem Formulation: I will transform continuous `charges` into discrete tiers using binning. This is an extension of a regression problem into multiclass classification.I will use Python with scikit-learn and XGBoost. Label creation is based on domain-informed thresholds.

Datasets:

- Source: Kaggle - Medical Cost Personal Dataset
- Type: Tabular
- Size: 1,338 rows × 7 columns
- Quality: High, clean, categorical and numerical mix
- Features: `age`, `sex`, `bmi`, `children`, `smoker`, `region`, `charges`

Resources & Processing: Requires binning target values into 3 cost categories, encoding, and scaling features. Will use scikit-learn for model training and evaluation, and Matplotlib for visualization.

Task & ML Algorithm: Multiclass classification using Random Forest, XGBoost, Logistic Regression

Tools/Libraries: Python, scikit-learn, Pandas, Seaborn, Matplotlib

Performance Metrics: Accuracy, Macro F1-Score, Confusion Matrix

Expected Performance: Expected baseline accuracy of ~65–70%. Confusion matrix will help reveal misclassification patterns between Medium and neighboring classes.

Validation: Train/Test split (80/20), Stratified K-Fold cross-validation

Goal of Final Package: A modular classification script or notebook with pre-set bin definitions, preprocessing functions, model training pipeline, and final performance report.

Workplan:

- Define cost tiers and explore data
- Prepare features and visualize distributions
- Train multiclass models and tune hyperparameters
- Evaluate and interpret results
- Package pipeline and finalize documentation

Project Proposal : Predict High-Cost Patients (Idea 2)

Abstract:

This project aims to classify individuals as high-cost or low-cost patients based on demographic and lifestyle attributes. Using the Medical Cost Personal Dataset, we will frame the regression target charges into a binary label and train classification models to predict high-cost risk. This project is useful for understanding healthcare cost prediction and binary classification pipelines.

Introduction:

Medical costs are rising globally, and insurance providers need models that can estimate patient risk levels based on personal and lifestyle information. This project focuses on predicting whether an individual will belong to the high-cost insurance group using simple machine learning models. The presentation will include data exploration, label transformation, model development, evaluation, and conclusion.

Motivation:

The ability to anticipate high-cost patients allows insurers to adjust policies and premiums proactively. The dataset is well-structured, publicly available, and beginner-friendly. This makes it ideal for exploring binary classification and its applications in healthcare. The goal is to frame the cost prediction task as classification instead of regression for better interpretability and practical use.

Problem Formulation:

This is a new take on an existing dataset, reframing a regression problem as a classification task. No prior work specifically addressed the binary classification of cost using this dataset, although many regression analyses exist. Python and libraries like scikit-learn and XGBoost will be used. I will label patients as high-cost if their charges exceed the median.

Datasets:

Source: Kaggle - Medical Cost Personal Dataset

Type: Tabular

Size: 1,338 rows × 7 columns

Quality: Clean, no missing values

Features: age, sex, bmi, children, smoker, region, charges

Resources & Processing:
I will use Python with Pandas and scikit-learn for data processing. Processing involves label creation, encoding categorical variables, and normalization. Visualizations will be created using Seaborn and Matplotlib.

Task & ML Algorithm:
Binary classification using Logistic Regression, Decision Trees, and Random Forests.

Tools/Libraries:
Python, scikit-learn, XGBoost, Pandas, Seaborn, Matplotlib

Performance Metrics:
Accuracy, Precision, Recall, F1-Score, ROC-AUC

Expected Performance:
Baseline accuracy expected around 75%. Tree-based models are likely to improve performance. SHAP values will help explain predictions.

Validation:
Train/Test split (80/20), cross-validation (5-fold)

Goal of Final Package:
A reusable Python module or notebook that takes input features and returns a binary prediction along with prediction probability and key feature influences.

Workplan:

- Data understanding and label creation
- Preprocessing and feature engineering
- Model training and evaluation
- Hyperparameter tuning and visualization
- Documentation and packaging