Mini Project Report

On

“Medical Cost & Insurance Approval Prediction”

Submitted By:
Anuj Parwal

Course Name: Machine Learning

RCOEM
Shri Ramdeobaba College of
Engineering and Management, Nagpur

Department of Computer Science and Engineering

(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

Shri Ramdeobaba College of Engineering & Management, Nagpur 440013

(An Autonomous Institute affiliated to Rashtrasant Tukdoji Maharaj Nagpur University Nagpur)

October 2025

Problem Statement: Medical Cost & Insurance Approval Prediction

Introduction :

The healthcare industry relies heavily on accurate cost estimation and fair insurance approval decisions. Medical insurance companies face challenges in determining the appropriate insurance charges for individuals and deciding whether an applicant should be approved for coverage. These decisions depend on multiple factors such as age, gender, BMI, smoking habits, region, and number of dependents.

Traditional manual or rule-based systems often struggle to handle the complexity and nonlinearity of these factors. To overcome these challenges, a data-driven, machine learning–based approach can automate and optimize both cost prediction and approval processes, ensuring fair, consistent, and explainable decisions-making in the healthcare insurance sector.

This project aims to develop a machine learning–based dual-model system that performs:

- **Regression** to predict medical insurance charges based on factors like age, BMI, smoking habits, etc.
- **Classification** to determine whether an applicant should be approved for insurance based on the predicted or actual medical cost and other demographic factors.

Objectives:

- **To predict the medical insurance charges** of an individual using regression algorithms based on demographic and health-related attributes.
- **To classify insurance approval status** (approved/rejected) based on the predicted cost and applicant details.
- **To perform comprehensive data preprocessing and EDA** to identify trends, correlations, and data quality issues such as missing values and outliers.
- **To build and tune machine learning models** (using Random Forest and Linear Regression) with cross-validation and hyperparameter optimization for maximum accuracy and reliability.

Methodology:

Workflow Steps:

1. **Data Collection:**
   Import the medical insurance dataset containing attributes like age, gender, BMI, region, children, smoker, and charges.
2. **Data Preprocessing:**
   - Handle missing values, duplicates, and outliers.
   - Encode categorical features using Label and One-Hot Encoding.
   - Apply feature scaling using StandardScaler.
3. **Exploratory Data Analysis (EDA):**
   - Perform univariate, bivariate, and multivariate analysis.
   - Visualize correlations using heatmaps, pairplots, and histograms.

4. **Model Development:**

   - **Regression:** Predict insurance charges using Linear Regression and Random Forest Regressor.

   - **Classification:** Predict insurance approval using Logistic Regression and Random Forest Classifier.

5. **Hyperparameter Tuning & Cross-Validation:**

   - Apply RandomizedSearchCV for parameter optimization.

   - Perform 5-Fold Cross-Validation for model stability.

6. **Evaluation & Visualization:**

   - Use metrics like R^2, MAE, and RMSE for regression; Accuracy, Precision, Recall, and F1-score for classification.

   - Plot feature importance, residuals, and confusion matrix.

7. **Deployment:**

   - Save models using Joblib and build a Streamlit-based web interface for live prediction.

Technology Stack:

1.Programming Language - Python.

2.Libraries Used - Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Joblib, Streamlit.

3.Machine Learning Models - Linear Regression, Random Forest Classifier.

4.Database - CSV-based dataset.

5.Deployment Tool - Streamlit Web Application.

6.Model Optimization - RandomizedSearchCV, Cross-Validation.

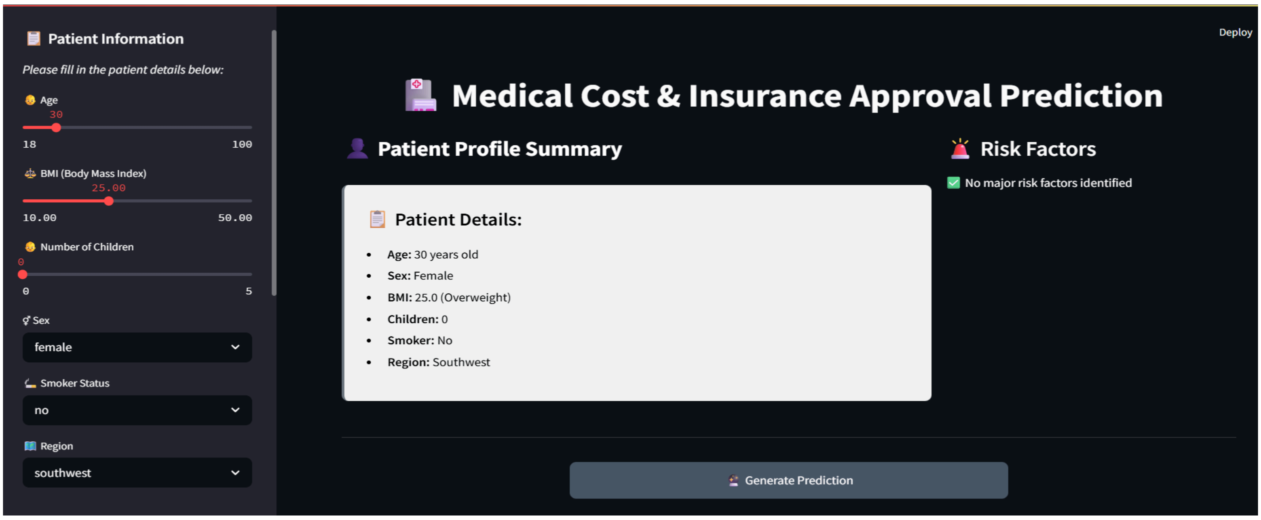
Result:

1.Key Outcomes -

- **Regression Model (Medical Cost Prediction):**
  - R² Score: ~0.78 (explains 78% variance).
  - Top features: *Smoking status*, *BMI*, *Age.*
- **Classification Model (Insurance Approval):**
  - Accuracy: ~91% (after hyperparameter tuning).
  - Stable performance across all 5 CV folds.
  - Best Parameters identified using RandomizedSearchCV.

2.Visual(Screenshot)-

Streamlit Web App Interface for live predictions:



Patient Information

Please fill in the patient details below:

Age
30
18 100

BMI (Body Mass Index)
25.00
10.00 50.00

Number of Children
0
0 5

Sex
female

Smoker Status
no

Region
southwest

Medical Cost & Insurance Approval Prediction

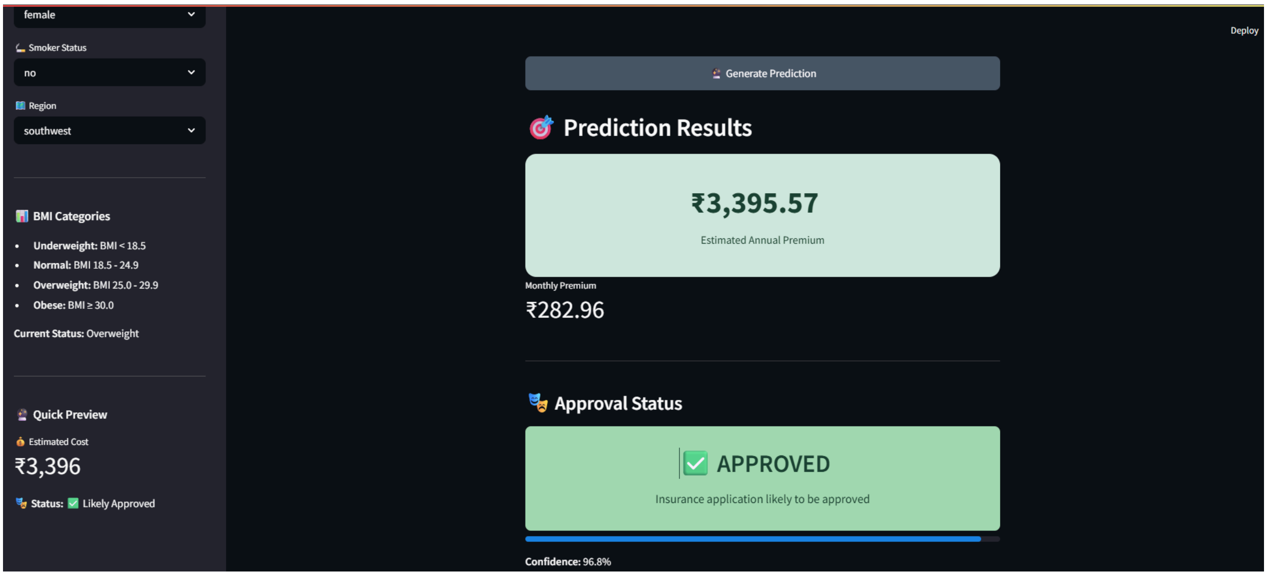
Patient Profile Summary

Patient Details:

- **Age:** 30 years old
- **Sex:** Female
- **BMI:** 25.0 (Overweight)
- **Children:** 0
- **Smoker:** No
- **Region:** Southwest

Risk Factors

✓ No major risk factors identified



female

Smoker Status

no

Region

southwest

📊 **BMI Categories**

- **Underweight:** BMI < 18.5
- **Normal:** BMI 18.5 - 24.9
- **Overweight:** BMI 25.0 - 29.9
- **Obese:** BMI ≥ 30.0

Current Status: Overweight

🔮 **Quick Preview**

💰 Estimated Cost

₹3,396

🎭 **Status:** ✅ Likely Approved

🎯 **Prediction Results**

₹3,395.57

Estimated Annual Premium

Monthly Premium

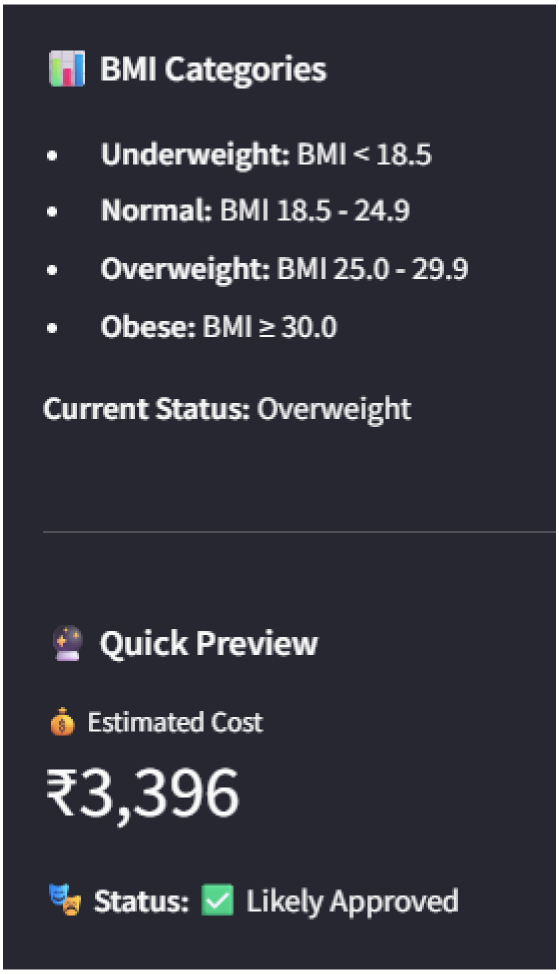
₹282.96

🎭 **Approval Status**

✅ APPROVED

Insurance application likely to be approved

Confidence: 96.8%



📊 **BMI Categories**

- **Underweight:** BMI < 18.5
- **Normal:** BMI 18.5 - 24.9
- **Overweight:** BMI 25.0 - 29.9
- **Obese:** BMI ≥ 30.0

Current Status: Overweight

🔮 **Quick Preview**

💰 Estimated Cost

₹3,396

🎭 **Status:** ✅ Likely Approved

Conclusion:

The project successfully developed an end-to-end machine learning solution for predicting medical insurance charges and automating approval decisions.

Through systematic EDA, preprocessing, and modeling, the solution achieved strong performance in both regression and classification tasks.

Advanced techniques like RandomizedSearchCV, Cross-Validation, and Feature Importance Analysis ensured model reliability and interpretability.

The deployment of the model via Streamlit demonstrates its potential for real-world insurance applications, enabling faster, data-driven, and fair decision-making in healthcare cost estimation and policy approval.