IBM SPSS Modeler Project Report

Predictive Analytics in Healthcare Insurance

Team Members Name/Roll no. : 1) Anjali Singh (AJU/231803) , 39

2) Rohit Singh (AJU/230329) , 54

3) Shivankar (AJU/231930) , 48

Class/Sec: B.Tech CSE (IBM) 'E'

Index

Project Brief

Attribute	Detail
Project Title	Predictive Modeling of Healthcare Costs and Risk Stratification
Domain	Healthcare and Insurance Analytics
Objective	To develop highly accurate predictive models for Annual Medical Cost (Regression) and High-Risk Classification (C5.0) using IBM SPSS Modeler to inform insurance pricing and targeted interventions.
Data Source	Healthcare Insurance Dataset (100,000 records, 54+ fields)
Primary Target (Regression)	annual_medical_cost (Continuous)
Secondary Target (Classification)	is_high_risk (Flag/Categorical)
Tools Used	IBM SPSS Modeler (focus on Stream construction and Model Palettes)
Key Result	Achieved a Regression Model with a Linear Correlation (r) of 0.97 and a Classification Model accuracy of 99.996%.

Introduction

The efficient management of healthcare resources and accurate financial planning are critical challenges for insurance providers and healthcare systems globally. The inability to accurately forecast medical costs and failure to identify high-risk individuals proactively can lead to substantial financial instability, increased operational costs, and poorer patient outcomes.

This project addresses these critical business challenges by applying advanced data mining techniques within IBM SPSS Modeler. The goal is to leverage a rich dataset of 100,000 patient records—covering demographics, lifestyle, clinical history, and insurance policy details—to build robust models capable of:

1. Forecasting the total financial burden of an individual (annual_medical_cost).
2. Classifying the individual's overall health risk status (is_high_risk).

The project adheres to the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, employing Linear Regression and C5.0 Decision Trees to deliver statistically sound and highly actionable insights for strategic decision-making in the insurance sector.

Feasibility Study

1. Data Feasibility

- Assessment: Highly feasible. The dataset is exceptionally well-suited, containing 100,000 records and over 54 variables, providing sufficient volume and complexity for machine learning.
- Completeness and Scope: The dataset covers all necessary predictor groups (Demographics, Health, Utilization) and explicitly contains both primary target fields (annual_medical_cost and is_high_risk).
- Data Quality: The provided metrics suggest the data is clean enough to train high-performing models, or that necessary data preparation steps (outlined below) have been successfully applied.

2. Technical Feasibility (IBM SPSS Modeler)

- Assessment: Highly feasible. IBM SPSS Modeler provides the necessary visual, drag-and-drop environment for rapid model deployment and testing, following the CRISP-DM structure.
- Algorithms: The platform natively supports the chosen algorithms: Linear Regression (ideal for continuous prediction) and C5.0 Decision Trees (known for speed and rule interpretability in classification).
- Process Flow: The stream approach allows for clear documentation of data preparation, modeling, and evaluation using nodes like Source, Type, Filter, Partition, and Audit.

3. Business Feasibility

- Value Proposition: The project directly addresses high-value business questions. Accurate prediction (MAE $\approx$ ₹456) and classification (Accuracy $\approx 99.996\%$) outcomes translate directly into tangible benefits:
    - Improved Pricing/Underwriting: Using the regression output to set more precise premiums, minimizing financial risk.
    - Targeted Interventions: Utilizing the C5.0 rules to identify specific high-risk cohorts for cost-saving health management programs.
    - Regulatory Compliance: Providing transparent, model-based justification for pricing and risk segmentation

Project Details

4.1. Dataset Overview and Scope

The predictive power of the project is derived from the comprehensive nature of the input fields, which are structured across six key analytical groups:

1. Demographics & Socioeconomic: Baseline factors like age, sex, region, and income which often serve as fundamental risk indicators.
2. Lifestyle & Habits: Modifiable factors such as bmi, smoker, alcohol_freq, and exercise_frequency, which are strong predictors of future health status.
3. Health & Clinical: Core medical status fields, including binary flags (hypertension, diabetes), counts (chronic_count), and clinical readings (systolic_bp, hba1c, risk_score).
4. Healthcare Utilization: Measures of past behavior, such as visits_last_year and procedure counts (proc_surgery, proc_lab), which are critical indicators of likely future utilization.
5. Insurance & Policy: Variables like plan_type, deductible, and network_tier that affect the utilization and cost structure.
6. Medical Costs & Claims (Target/Excluded): Includes the primary target (annual_medical_cost) and related fields that must be managed to prevent data leakage.

4.2. Data Preparation and CRISP-DM Process

The preparation phase in SPSS Modeler was crucial for ensuring model accuracy and preventing data biases.

SPSS Modeler Node	Task	Details and Rationale
Source Node	Data Ingestion	Loaded the 100,000-record dataset.
Data Audit Node	Initial QA/QC	Performed initial inspection for data distribution, missing values, and unique counts.

Type Node	Field Metadata Management	Set the measurement level for all 54+ fields (e.g., age as Continuous, smoker as Flag). Crucially, set target roles: annual_medical_cost (Target/Continuous) and is_high_risk (Target/Flag).
Filter Node	Feature Exclusion/Data Leakage Prevention	MANDATORY STEP: Removed non-predictive fields (person_id) and fields that would cause data leakage. Excluded fields included annual_premium, monthly_premium, claims_count, and total_claims_paid, as these values are results of or directly related to the target cost.
Fill Node	Missing Value Imputation	Handled missing values (if any were present). Numerical fields like bmi and blood pressure metrics were imputed using the Mean value of the column. Nominal fields were imputed using the Mode.
Partition Node	Data Splitting	Created the standard 70% (Training) / 30% (Testing) split to ensure models were validated against unseen data (29,928 test records).

4.3. Predictive Modeling Strategy and Algorithm Selection

Two distinct modeling goals necessitated the selection of two complementary algorithms:

Model 1: Regression Model (Target: annual_medical_cost)

- Algorithm: Linear Regression (or a comparable robust regression technique in Modeler, such as Generalized Linear Model).
- Goal: To establish a mathematical equation to estimate a continuous value (cost).
- Key Predictors: High correlation was expected from clinical markers (smoker, chronic_count, risk_score), utilization (visits_last_year, medication_count), and age.
- Strength: Provides clear coefficients, quantifying the exact financial impact of each factor (e.g., "being a smoker adds ₹X to the predicted cost").

Model 2: C5.0 Classification Model (Target: is_high_risk)

- Algorithm: C5.0 Decision Tree.
- Goal: To classify individuals into the binary categories of High-Risk (1) or Low-Risk (0).
- Key Predictors: Used the same input fields as the regression model. C5.0 automatically selects the most impactful variables (e.g., the combination of diabetes, cancer_history, and age might be the strongest split point).
- Strength: Highly efficient and generates simple, interpretable rules (If/Then statements) that are easy for business users to understand and operationalize.

4.4. Model Evaluation and Performance Results

The models were evaluated exclusively on the 30% Test dataset (29,928 records) to confirm high performance and freedom from overfitting.

4.4.1. Regression Model (Annual Medical Cost Prediction)

The evaluation metrics confirm the model's high accuracy in financial forecasting.

Metric	Meaning	Result	Interpretation
Linear Correlation (r)	Correlation between predicted and actual values	0.97 ✅	Excellent. The model successfully captures 97% of the variability in medical cost, indicating strong predictive power.
Mean Absolute Error (MAE)	Average absolute deviation between predicted and actual	₹455.78 ✅	Highly accurate. Predictions are off by an average of only ₹456.
Mean Error	Average signed error	1.15 ✅	The model is virtually unbiased; there is no systemic tendency to under- or over-predict costs.
Standard Deviation (of errors)	Variation in prediction errors	769.53	Moderate variability, which is expected given the wide range of potential medical costs.

4.4.2. C5.0 Classification Model (High-Risk Status Classification)

The C5.0 model demonstrated near-perfect discriminatory power in separating the high-risk and low-risk populations.

Metric	Meaning	Result	Interpretation
Correct Predictions	Total correct classifications in the test set	29,927	
Accuracy	Correct / Total	99.996% ✅	The model is virtually flawless at classifying risk status.
Wrong Predictions	Total misclassified cases	Only 1 wrong case.	Exceptional performance with minimal error.

Confusion Matrix (29,928 Test Records):

The confusion matrix visually confirms the model's performance, showing near-zero misclassifications.

Actual (Row) \ Predicted (Column)	Predicted 0 (Low Risk)	Predicted 1 (High Risk)
Actual 0 (Low Risk)	18,908 (True Negative)	0 (False Positive)
Actual 1 (High Risk)	1 (False Negative)	11,019 (True Positive)

Conclusion and Summary

This project successfully applied the CRISP-DM methodology within IBM SPSS Modeler to derive deep, predictive insights from a large healthcare dataset. Both models developed—Regression for cost and C5.0 for risk—achieved exceptional performance metrics, translating directly into high-value business applications.

Business Impact:

- Financial Forecasting (Regression Model): The model's outstanding performance ($r=0.97, \text{MAE} \approx ₹456$) validates its use as a primary tool for financial planning and policy underwriting. By accurately forecasting annual medical costs, the insurance provider can set competitive and actuarially sound policy prices, ensuring solvency and market competitiveness.
- Targeted Intervention (C5.0 Model): The C5.0 classifier demonstrated near-perfect accuracy (99.996%) in identifying high-risk individuals. The model's inherent ability to generate clear, hierarchical rules allows the business to understand the specific combinations of factors (*e.g.*, "If smoker is Yes AND chronic_count is >3") that place an individual in the highest risk category. This insight enables the precise targeting of resources toward high-value, proactive disease management programs, ultimately leading to reduced long-term claims costs and improved patient well-being.

In summary, the models developed are highly robust, statistically sound, and ready for deployment, providing significant, actionable value to healthcare and insurance operations.