# 

| Module Title: | Machine Learning |
| --- | --- |
| Assessment Title: | Machine Learning Repeat |
| **Lecturer Name:** | ***Muhammad Iqbal*** |
| **Student Full Name:** | ***Barbara* *Weltson*** |
| **Student Number:** | sba24096 |
| **Assessment Due Date:** | 27/07/2025 |
| **Date of Submission:** | 25/07/2025 |

# 

[**Introduction 1**](#_ve65avaahqbz)

[**Methodology 2**](#_tx3kmfkmn887)

[Data Source 2](#_kjrg2chsp1cp)

[Data Preparation 2](#_pvh2ikwkz7ou)

[**Model Features 3**](#_bxq1iwii6hpj)

[**Technical Challenges 3**](#_m81nijctkfob)

[**Insights and Analysis 4**](#_d5qmub8l0gee)

[**Conclusion 5**](#_v5iqcyrmt04)

## **Introduction**

The ability to understand insurance cost patterns and predict healthcare expenses is important for making good decisions in healthcare and insurance companies. This project aims to build machine learning models to study UK medical insurance cost data, with a focus on making prediction tools easy to use for insurance professionals and healthcare workers.

The analysis uses Python as the main technology, which allows us to create prediction models that are simple to build and still give clear, useful results for business needs (Python Software Foundation, 2023).

## **Methodology**

### **Data Source**

The analysis uses data from a medical insurance dataset, originally provided as an Excel file but converted to CSV format for easier processing. I changed the file type to CSV because this format loads faster, works better with Python, and avoids compatibility issues with Excel dependencies. The "insurance\_data.csv" file contains 1,338 records, including:

* age: Main person's age (18-64 years)
* gender: Insurance holder's gender (female/male)
* bmi: Body Mass Index values from 15.96 to 53.13
* children: Number of family members covered by health insurance
* smoker: If person smokes (yes/no)
* region: Where person lives across four areas (northeast, southeast, southwest, northwest)
* NoClaimsBonus: Bonus percentages (5%, 10%, 15%, 20%)
* charges: Medical costs paid by health insurance

### **Data Preparation**

To prepare the data for machine learning analysis, I took these steps:

Loading the Data: I loaded the CSV file and checked the structure to find missing values and data problems that needed fixing.

Cleaning Missing Values: I removed incomplete records instead of trying to fill in missing data. This removed 17 rows but made sure we had clean, reliable data for training models. For prediction work like this, data quality is more important than having lots of data.

Creating Model Features: I changed text categories into numbers for machine learning algorithms. For simplicity, I used these codes:

* Gender: Female=0, Male=1
* Smoker: No=0, Yes=1
* Region: Number codes for four areas
* NoClaimsBonus: Changed percentages to decimal numbers

**Setting Target Variables**: I created two prediction approaches to give users complete analysis options:

* Regression: Predict exact insurance cost amounts
* Classification: Put customers into "High Cost" or "Low Cost" groups using average charges

## **Model Features**

The machine learning analysis has two main approaches chosen for insurance industry needs:

**Regression Models**: Predict exact cost amounts using Random Forest and Linear Regression. This approach was chosen because exact cost prediction helps insurance companies with setting prices and financial planning.

**Classification Models**: Put customers into risk groups using Random Forest algorithms. This classification was selected because simple risk assessment helps with quick decisions and customer grouping strategies.

**Multiple Testing Splits**: I tested models using 70-30, 80-20, and 75-25 train-test splits. I included multiple splits because model stability across different data divisions shows reliability for real-world use.

**Model Stability:** Testing across three different train-test splits (70-30, 80-20, 75-25) demonstrated excellent model stability, with performance variations of less than 1% for both regression and classification approaches. This consistency indicates reliable model performance for real-world deployment.

**Model Optimization**: GridSearchCV with cross-validation to find the best model settings. This optimization helps get the best possible performance while avoiding overfitting to training data.

**Feature Importance Analysis**: Shows which factors most strongly predict insurance

**Performance Comparison**: Side-by-side comparison of different models and data splits.

## **Technical Challenges**

Some challenges came up during development, each needing careful consideration:

**Missing Value Treatment**: At first, I tried to use advanced methods for filling in missing data. However, this approach had problems related to adding bias into the prediction models. Instead of using complex filling methods, I chose to remove incomplete records entirely.

**Feature Engineering**: The dataset had mixed data types and percentage values that needed careful preprocessing. This challenge showed the typical gap between raw insurance data (designed for admin purposes) and machine learning datasets (optimized for algorithms). The transformation process involved normalizing categories, converting percentages, and creating meaningful binary classifications.

**Model Selection Strategy**: The original approach tested many algorithms at the same time, creating evaluation problems. Too many model comparisons resulted in unclear performance patterns, while focusing on single algorithms missed important insights. After testing multiple approaches, I developed a focused comparison between Random Forest and Linear Regression that provided clear performance differences while staying easy to understand.

**Cross-Validation Implementation**: Model validation needed careful implementation to ensure reliable performance estimates. Insurance professionals, used to clear business rules, need confidence in prediction stability. Using k-fold cross-validation, multiple train-test splits, and systematic model tuning were essential steps to create models that could maintain trust with this professional audience.

## **Insights and Analysis**

The machine learning analysis reveals several important patterns relevant to insurance cost prediction:

**Smoking Impact**: Smokers show dramatically higher average charges (£32,050 vs £8,434 for non smokers), representing a 284% cost increase. I highlighted this pattern because it directly impacts risk assessment and premium calculation strategies, which aligns with established health research showing smoking as a major risk factor (WHO, 2023).

**Age Progression:** When filtering by age groups, costs increase consistently with age, particularly accelerating after age 50. I emphasized this trend because it validates age based pricing tiers and helps with actuarial modeling, reflecting the well-documented relationship between aging and healthcare costs (NHS, 2023).

**BMI Relationships**: Higher BMI correlates with increased charges, though the effect is less dramatic than smoking. I included this analysis because weight related health factors increasingly influence insurance pricing decisions, consistent with medical evidence linking obesity to increased healthcare utilization (CDC, 2023).

**Model Performance:** The Random Forest regression models achieved strong predictive performance with R² scores ranging from 0.843 to 0.854 across different train-test splits, indicating the models explain approximately 85% of the variation in insurance costs. Classification models performed even better, achieving accuracy scores between 92.7% and 92.8%, demonstrating excellent ability to categorize customers into high and low cost groups.

**Algorithm Comparison:** Random Forest consistently outperformed Linear Regression across all splits, with Random Forest achieving R² scores around 0.85 compared to Linear Regression's 0.75-0.76. This 10-percentage point difference highlights Random Forest's superior ability to capture complex relationships in insurance data.

## **Conclusion**

This machine learning analysis successfully created reliable prediction models for insurance cost estimation using Python tools. The Random Forest algorithm achieved strong performance with 85% accuracy for cost prediction and over 92% accuracy for risk classification. The analysis demonstrates that smoking status, age, and BMI are the most important factors driving insurance costs, providing valuable insights for premium calculation and risk assessment in the insurance industry.

**References**

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), pp.5–32. doi:https://doi.org/10.1023/a:1010933404324.

CDC (2022). *About adult BMI* . [online] Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/healthyweight/assessing/bmi/adult\_bmi/index.html.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*. [online] O’Reilly | Safari. Available at: https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/.

Matplotlib (2024). *Matplotlib: Python Plotting — Matplotlib 3.1.1 Documentation*. [online] Matplotlib.org. Available at: https://matplotlib.org/.

scikit-learn (2019). *scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/.

World Health Organization (2025). *Tobacco*. [online] World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/tobacco.

www.oreilly.com. (n.d.). *Introduction to Machine Learning with Python [Book]*. [online] Available at: https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/.