**Scenario:**

In recent years, the healthcare landscape in the United Kingdom has been significantly impacted by inflation, particularly affecting the cost of medical insurance. As a student tasked with analyzing a dataset on medical insurance costs, your objective is to conduct thorough data cleaning, exploratory analysis, and potentially use relevant features to enhance predictive accuracy. This dataset contains valuable information that includes the age, gender, BMI (Body Mass Index), number of children covered, smoking status, residential region, NoClaimsBonus and medical charges for each insurance beneficiary. The dataset has null values and it is your responsibility to clean before performing tasks for this assignment.

**Dataset Description:**

* age: Age of the primary beneficiary under the insurance policy.
* gender: Gender of the insurance contractor (female or male).
* bmi: Body Mass Index, a numerical indication of relative body weight based on height and weight.
* children: Number of children or dependents covered by the health insurance.
* smoker: Binary variable indicating whether the beneficiary is a smoker (yes/no).
* region: Residential area of the insurance beneficiary in the US (northeast, southeast, southwest, northwest).
* NoClaimsBonus: It is ranged from 5% to 20% with only a maximum 4 possibilities (5%, 10%, 15% and 20%).
* charges: Individual medical costs billed by health insurance.

**Machine Learning**

* Provide a conceptual understanding and logical justification based on the reasoning for the specific choice of machine learning approaches (supervised/ Unsupervised) for the provided data set. (0-20 marks)
* Machine Learning models can be used for Prediction, Classification, and Clustering. You can choose suitable features for the machine learning models based on feature selection methods, such as random forest or any other method. The selection of hyperparameters for the ML models should be performed by using hyperparameter tuning, such as GridSearchCV. Obtain best accuracy using optimal values of the hyperparameters. (0-30 marks)
* You should train and test the Machine learning models in the case of supervised learning for different splits (at least 3 distinct logical splits) and use appropriate metrics for unsupervised learning. Use K-fold cross-validation to provide authenticity of the modelling outcomes. (0-30 marks)
* Exhibit a comparison of ML modelling outcomes using a Table or graph visualisation. Identify the possible similarities and contrast of the Machine Learning modelling outcomes based on chosen metric and discuss their statistical understanding. (0-20 marks)

**Report**

A report is required to provide the details of work performed in all tasks. The report should be based on Introduction, rationale for machine learning models and conclusions. Illustrations should be used to highlight the details of any section.

This should also include evidence to support your data wrangling and analysis through the use of references and citations.

Conclusions, Findings of data set and references/citations in (HARVARD style).

This should be completed in the Jupyter Notebook Markdown.

**Submission Requirements**

All assessment submissions must meet the minimum requirements listed below. Failure to do so may have implications for the mark awarded.

All assessment submissions must:

* 1000 (+/- 10%) words in report (not including code, code comments, titles, references or citations)
* The Jupyter Notebook File Must be saved as “YourName\_ML\_HDip\_CA.ipynb”
* Be submitted by the deadline date specified or be subject to late submission penalties
* Be submitted via Moodle upload
* Use [Harvard Referencing](http://40.115.124.2/sp/subjects/guide.php?subject=harvardref) when citing third party material
* Be the student’s own work.
* Include the CCT assessment cover page.

# 

| 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:** |  |

# 

## **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.

**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