# IMPLEMENTATION SUMMARY

The aim of this report is to present a detailed implementation summary of the AI technique that was applied to a selected health-related open-source dataset. This summary will cover the following key points

1. Dataset presentation and documentation
2. Discussion of selected AI technique(s)
3. Data preparation discussion
4. Insights from data visualization
5. Implementation of the AI technique

## 1.0 Dataset presentation and documentation

The selected dataset for this task was obtained from Kaggle (Available at: https://www.kaggle.com/datasets/mirichoi0218/insurance). Kaggle is an open-source online community for diverse datasets. The selected health-related dataset is titled 'Medical Cost Personal Datasets'. This is a dataset that contains previous records of the health charges paid by individuals from different demographic and health backgrounds. The dataset contains 7 variables which are described viz;

1. Age: this is the age of individuals measured in years
2. Sex: This variable indicates the gender of an individual (Male and Female)
3. BMI: This variable measures the body mass index of individuals. It was measured on a continuous scale
4. Children: the specifies the number of children covered by the individual's insurance plan
5. Smoker: it indicates whether the individual smokes (yes or no)
6. Region: This variable represents the geographical region where the individual resides
7. Charges: this is the dependent variable. It is the medical cost or insurance charges associated with the individuals.

The table below provides a further description of the dataset

| Variable | Description | Data Type |
|---|---|---|
| Age | Age of the patient in years | Numeric |
| Sex | Gender of the patient: 0 - female, 1 - male | Object |
| BMI | Body weight in kg/m²: <br>1 - Underweight (Below 18.5) <br>2 - Healthy Weight (18.5 - 24.9) <br>3 - Overweight (25.0 - 29.9) <br>4 - Obesity (30.0 and above) | Float |
| Children | Number of Children | Numeric |
| Smoker | Number of Smokers | Numeric |
| Region | Regions of Patients: <br>1 - Southwest <br>2 - Southeast <br>3 - Northwest <br>4 - Northeast | Object |
| Charges | Charges of each Patient | Numeri |

***Table 1: Dataset variable description***

**2.0 Discussion of Selected AI Techniques**

One single factor has influenced the choice of the selected AI techniques used for this task. The factor is the type of problem that this task aims to solve.

The aim of this task is to develop a predictive model which will be capable of accurately predicting medical charges for individuals based on their demographic characteristics and other factors.

Firstly, it was identified that this is a supervised machine learning problem because of the presence of input features and target variable (Singh, Gupta and Garg, 2022). Supervised machine learning models are models that are trained using labelled data where the algorithm learns to map input data to the correct output (Rashidi et al., 2019). Having recognized that this is a supervised machine learning model problem, it is important to decide if it is a regression problem or classification problem. Regression problems in supervised machine learning are problems that involves the prediction of a continuous output variable based on one or more input variables (Sundari et al., 2022). The output variable in the problem at hand is *'charges'* which is measured on a continuous scale of measurement. Therefore, this informs the decision that this is a regression supervised machine learning problem.

There are various types of machine learning models that come under regression models. Examples are linear regression, Support vector regression, random forest regressor, Naïve Bayes and so on (Sundari et al., 2022, Guo and Chang, 2022, Prasad et al., 2022).

This implementation has chosen to adopt Linear regression and Random forest regressor to fit the medical health insurance charges dataset. The choice of using two models was driven by the passion to be able to compare the two models so as to be able to select the most accurate model.

The evaluation metric used for selecting the best model is the r2_score.

## 3.0 Data Preparation

The data preparation techniques involve the transformation of raw data into a suitable format for analysis. It involves handling missing values, encoding categorical variables, scaling, removing outliers and so on (Erol *et al.*, 2022).

### *Handling missing values*

```
In [4]: df.shape
Out[4]: (1338, 7)
```

The loaded dataset contains 1338 observations or data points. Checking for the presence of missing values as displayed in the next screenshot showed that the dataset has no missing values.

```
In [77]: df.isnull().sum()
Out[77]: Age                   0
         Sex                   0
         Body weight in kg/m2  0
         Children              0
         Smoker                0
         Regions               0
         Charges               0
         dtype: int64
```

.

### *Encoding categorical variables*

The categorical variables contained in the dataset are *sex, object* and *region.* Each of these variables was assigned values as shown below

```
In [11]: df['Sex'] = df['Sex'].replace({'female': 0, 'male': 1})
         df['Smoker'] = df['Smoker'].replace({'no': 0, 'yes': 1})
         df['Regions'] = df['Regions'].replace({'southwest': 1, 'southeast': 2, 'northwest': 3, 'northeast': 4 })
```

***Outlier detection***

To detect outliers from the numeric columns, a boxplot was employed. While there was no case of outlier detected from the *"Age"*, the *"Body weight in kg/m2"* showed the presence of no serious cases of outlier, and the *"Charges"* showed a significant presence of outlier.

The process taken to remove the outlier is shown below

```
In [14]: Q1 = np.percentile(df['Charges'], 25)
         Q1
Out[14]: 4740.28715

In [15]: Q3 = np.percentile(df['Charges'], 75)
         Q3
Out[15]: 16639.912515

In [16]: IQR = Q3-Q1
         IQR
Out[16]: 11899.625365

In [17]: lowerbound = Q1 - 1.5 * IQR
         lowerbound
Out[17]: -13109.1508975

In [18]: upperbound = Q3 + 1.5 * IQR
         upperbound
Out[18]: 34489.350562499996

In [20]: df_clean = df[(df['Charges'] >= lowerbound) & (df['Charges'] <= upperbound)]
         df_clean.head(2)
```



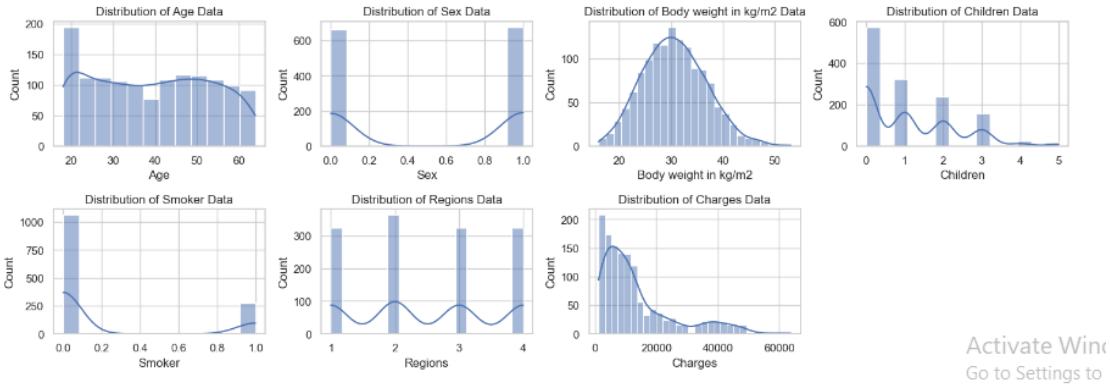
The cleaned data was stored in *df_clean* and the new shape of the data was;

```
In [21]: df_clean.shape
Out[21]: (1199, 7)
```

## 4.0 Insights from Data Visualization

The visualization included univariate bar charts to show the distribution of each variable, some bivariate plots to show the relationship between the variables and correlation heatmap.

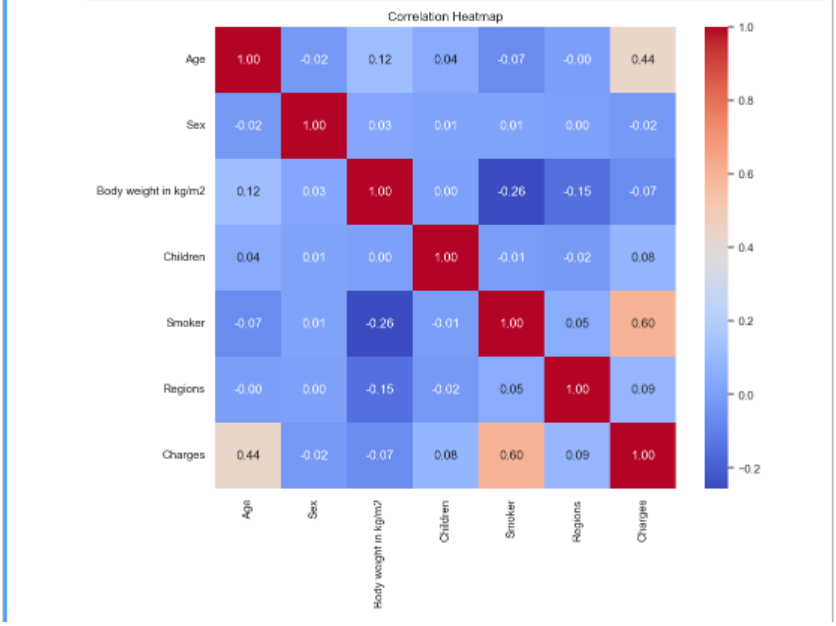
```
In [28]: plt.figure(figsize=(15,10))
         for i,col in enumerate(df_clean.columns,1):
             plt.subplot(4,4,i)
             plt.title(f"Distribution of {col} Data")
             sns.histplot(df[col],kde=True)
             plt.tight_layout()
             plt.plot()
```






```
In [45]: corr_matrix = df_clean.corr()

         # Plot the heatmap
         plt.figure(figsize=(10, 8))
         sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f', square=True)
         plt.title('Correlation Heatmap')
         plt.show()
```




## 5.0 Implementation of AI technique

The implementation process involves the following steps

1. Declaring the dependent variable y and the independent variables X: X is a dataframe containing all the variables except *"charges"* while y contains *"charges"* alone

2. Splitting into train test: the split ratio adopted was 80% for the training dataset while 20% was used for the test dataset
3. Model fitting: Each of LinearRegression and RandomForestClassifier was fitted and the prediction was made on the testing dataset. The r2_score was computed for each model and the result of the r2_score is displayed in the table below

**Note:** A higher R2_score implies a better model

| **Model** | **r2_score** |
|---|---|
| LinearRegression | 78.3% |
| RandomForestClassifier | 86.6% |

With RandomForestClassifier having the highest accuracy, the model became our chosen model for saving.

4. Model saving: the model was saved with 'pickle'; a python framework for exporting the model.
5. Model deployment: Deploying the model on the local computer gave an avenue to be able to use the model's application. The model was deployed by utilizing 'streamlit' robust strength. The screenshot below shows the usage of the web app.

By successfully building a model that can predict medical health charges with 81.7% accuracy, health facility users can know how much they are likely to be charged for health services even before visiting a health facility, thus providing an AI-based platform to make life easier.

## References


Erol, G. *et al.* (2022) 'Analyzing the effect of data preprocessing techniques using machine learning algorithms on the diagnosis of COVID-19,' *Concurrency and Computation*, 34(28). https://doi.org/10.1002/cpe.7393.

Guo, C.-Y. and Chang, K.-H. (2022) 'A novel algorithm to estimate the significance level of a feature interaction using the extreme gradient boosting machine,' *International Journal of Environmental Research and Public Health/International Journal of Environmental Research and Public Health*, 19(4), p. 2338. https://doi.org/10.3390/ijerph19042338.

Prasad, J. *et al.* (2022) 'Relevant-Based Feature Ranking (RBFR) method for text classification based on machine learning algorithm,' *Journal of Nanomaterials*, 2022, pp. 1–12. https://doi.org/10.1155/2022/9238968.

Rashidi, H.H. *et al.* (2019) 'Artificial intelligence and Machine Learning in Pathology: The present Landscape of Supervised Methods,' *Academic Pathology*, 6, p. 2374289519873088. https://doi.org/10.1177/2374289519873088.

Singh, T.P., Gupta, S. and Garg, M. (2022) 'Machine Learning: A Review on Supervised Classification Algorithms and their Applications to Optical Character Recognition in Indic Scripts,' *ECS Transactions*, 107(1), pp. 6233–6250. https://doi.org/10.1149/10701.6233ecst.



Singh, T.P., Gupta, S. and Garg, M. (no date) 'Machine Learning: A Review on Supervised Classification Algorithms and their Applications to Optical Character Recognition in Indic Scripts,' *ECS Transactions*, 107(1), pp. 6233–6250. https://doi.org/10.1149/10701.6233ecst.

Sundari, V. *et al.* (2022) 'Crop recommendation and yield prediction using machine learning algorithms,' *World Journal of Advanced Research and Reviews*, 14(3), pp. 452–459. https://doi.org/10.30574/wjarr.2022.14.3.0581.