Daniel Leech 937285

Implementation Summary

Health-Related Dataset and Selected Technique

The chosen dataset depicts medical expenditure data and the main factors behind it in various central cities and suburbs within Japan. This specific dataset has been chosen because of medical expenditure’s momentous importance regarding quality of care given to patients as well as resource allocation. Because of this, an effective artificial intelligence (AI) technique utilising the data can potentially lead to vastly beneficial outcomes to the medical industry. The selected AI technique to be implemented is multivariate linear regression which attempts to model the relationships between multiple dependent and independent variables with the goal of making predictions based on the data provided. By applying multivariate linear regression appropriately to the chosen dataset, medical expenditure can be accurately predicted using the plethora of factors available, with the overall aim of providing advantageous forecasts that outline areas where costs can be reduced as well as services that can be optimised to improve patient outcomes.

A screenshot of a computer

Description automatically generated

Figure 1 - Dataset Sample

This figure provides brief insight into the dataset containing overall medical expenses and some of the main factors causing it within various central cities and suburbs within Japan. It was published in 2022 which ensures the data is of recent origin and not outdated otherwise the effectiveness of the implementation could be diminished significantly.

Data Preparation

In order to appropriately and effectively implement the AI technique into the chosen dataset, various methods of data preparation have been chosen to adapt the dataset and make the implementation have a meaningful outcome to healthcare systems.



Figure 2 - Transposing Rows and Columns

When analysing the dataset that has been chosen, the rows and columns initially need to be transposed and aligned in order to mirror the format typically seen within datasets, to enable the AI technique to be implemented effectively. The above code takes the original dataset (medicalexpenditure) and applies the transpose attribute (.T), swapping the rows and columns without harming the data, which is then written to a new CSV file (medicalexpenditure2) to preserve the original. This new file is then used to create the data frame that the AI technique will utilise.

A screenshot of a computer

Description automatically generated

Figure 3 - Transposed Dataset Sample

The above figure represents the dataset after the transposition has occurred, with rows displaying each area’s specific expenditure data and columns containing each of the fields related to expenditure as well as its overall value. This ensures that multivariate linear regression can be performed on the dataset successfully and effectively.

A black background with green lines

Description automatically generated

Figure 4 - Identifying Redundant Data

The next step of data preparation is the identification and omission of redundant and unnecessary data through outputting various general statistics seen in the code above. This outlines any missing or unnecessary data throughout the dataset which can then be removed. As the chosen dataset does not contain any redundant data, specific values did not have to be manipulated or omitted.

A computer screen shot of a code

Description automatically generated

Figure 5 - Identification and Removal of Duplicate and Null Values

Another way in which the dataset is prepared is through the recognition and potential removal of duplicate or null values. The above code outputs the number of duplicated values and subsequently removes them within a new data frame, ensuring the data is accurate. Additionally, null values are also identified and columns containing only null values are dropped from the dataset, ensuring they are not used when the implementation occurs. Each row of the dataset also contains a handful of null values which are then removed, confirming that the dataset is ready for implementation of the AI technique.

Implementation and Parameter Setting

Due to the effective data preparation that has taken place, parameters can be carefully selected from the dataset in order to implement multivariate linear regression and accurately predict medical expenditure.

A screen shot of a computer

Description automatically generated

Figure 6 - Parameter Setting

In order for the AI to make accurate predictions about medical expenditure, a plethora of the most relevant factors have been selected as parameters. Medical expenses for inpatients and outpatients are indispensable parameters because of them making up the vast majority of the overall expenditure. Inpatient expenses include costs such as specialist and surgical fees and outpatient expenses include consultation and medication fees, therefore proving their importance in the context of predicting overall expenditure. Number of doctors, beds and nurses are other vital factors influencing expenditure due to their correlation with labour costs and hospital capacities, which further reflect overall expenses in a specific area. Finally, average income in each area is another crucial parameter when predicting medical expenses as higher levels of income typically leads to higher medical expenditure because of more comprehensive healthcare plans being purchased as well as non-urgent medical services being more accessible.

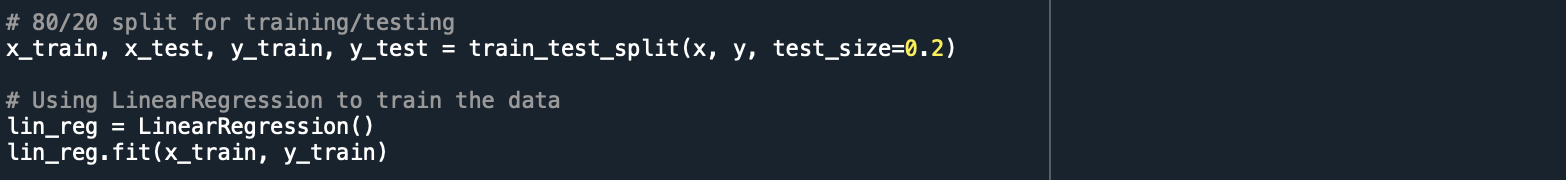


Figure 7 - Implementation

The above figure represents the implementation using the parameters that have been selected. The dataset is split to direct 80% to training the AI and 20% to testing its effectiveness. This ratio has been chosen due to the number of entries within the dataset in an attempt to allocate as much data as possible to the AI while still retaining enough to test it reliably. The “LinearRegression” class from “sklearn.linear\_model” is then utilised in combination with the “.fit” method to train the AI using the defined parameters and training data allocated which can then be used to predict the medical expenditure.

Results

Now that the AI has been trained using the chosen data from the dataset, it can utilise multivariate linear regression in order to predict medical expenditure and its accuracy can be tested and evaluated.

A computer code with green text

Description automatically generated

Figure 8 - Outputting AI Technique’s Effectiveness

The above code is used to determine the effectiveness of the multivariate regression. This is done by outputting the coefficient of determination as well as plotting a graph to illustrate its accuracy when compared to the test values.

A graph with red and blue dots

Description automatically generated

Figure 9 - Illustrating Results of the Implementation

The above figure represents the accuracy of the multivariate regression in numerical and graph formats. The coefficient of determination being 0.95 outlines that the parameters selected are vastly successful at predicting medical expenditure however this number may be subject to marginal changes with each iteration of the implementation due to the AI predicting slightly different values each time. The graph further demonstrates the effectiveness of the multivariate regression as the predicted and test values have high levels of alignment and all predicted values are plotted near where the x-axis equals the y-axis, indicating their overall accuracy. These results evoke the success of the algorithm at predicting expenditure, therefore presenting the various benefits it could have on healthcare systems such as optimising costs and resource allocation.

Comparing implemented AI technique with other possible techniques

When comparing the implemented multivariate linear regression to another technique that could also utilise the expenditure dataset, a decision tree regressor may be effective at identifying relationships between factors of expenditure and region. This could result in meaningful outcomes to healthcare services through the identification of regions with inefficient and unsuitable expenses which could then be optimised. If this technique was implemented, data preparation would be mirrored as the data tree regressor could utilise the numerical data supplied in the dataset, and implementation would involve defining the relevant factors and target variable, splitting the data into train and test sets and plotting the tree using a defined depth. The decision tree could then be used to make predictions about the data and statistics such as accuracy could also be calculated to determine its effectiveness.

References

Yuna Seo; Takaharu Takikawa (2022). Regional variation in national healthcare expenditure and health system performance in central cities and suburbs in Japan [Dataset]. <http://doi.org/10.5061/dryad.h18931znw>

*sklearn.linear\_model.LinearRegression — scikit-learn 1.4.2 documentation*. (2024). [Online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>

*Sklearn.Tree.DecisionTreeRegressor — scikit-learn 1.4.2 documentation*. (2024). [Online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>