EVALUATION REPORT: PREDICTING OBESITY LEVELS IN MSOA'S

Dataset: Obesity Dataset (Middle Layer Super Output Areas – MSOAs)
Data Mining Method: Random Forest Regression and Classification

STUDENT NAME: KAMALRAJ DASHADIYA
Student Id: U2393606

Table of Contents

Introduction

This report compares the performance of Random Forest algorithm in predicting obesity levels all over different Middle Layer Super Output Areas (MSOAs). It also will determine which factors have the highest contribution towards obesity and make an evaluation of the performance of the model. Due to variation in effecting demographic, geographical and socio-economic factors, obesity can be ranked as a suitable topic for analysis using data. Random Forest classification is one of the most popular machine learning algorithms suitable for regression analysis as well. It was introduced by Leo Breiman in the year 2001, it works by building an ensemble of the decision trees and then takes the final decision of prediction by combining of all the decisions made by the trees.

Characteristics of Random Forest

Key characteristics of Random Forest include:

- **High Accuracy**: This is due to the fact that the algorithm has an ensemble characteristic, which makes the predictions highly accurate and credible.
- **Resistance to Overfitting**: Random Forest, while also taking the average of several decision trees, reduces data overfitting, especially in noisy data.
- **Feature Importance Analysis**: It can also serve the purpose of allowing an understanding of the relative significance of the inputs, interpretability.

Random Forest does have its own limitations, it can be expensive in its computation for big datasets and may need a good number of hyperparameters to be adjusted like the number of trees to be created, the maximum extent of depth of the trees, and the minimum samples to split. For instance, Random Forest poses strengths highlighted above, while it has limitations that are outlined below, and this gives it an advantage when dealing with cumbersome prediction problems such as obesity analysis.

Results and Answers to Research Questions

Research Question 1: What factors contribute to higher obesity levels in certain regions?

Answer:

Analyzing factors which helped in enhancing the obesity level in regions, both the Random Forest Classifier and the Random Forest Regressor yielded promising results. It was also established, through feature importance analysis, that socioeconomic factors such as; income levels, education levels, demographic factors such as ages, gender distribution, health related fields such as obesity percentage and diabetes percentage where among the most useful predictors (Ayua, 2024).

- Low income and thus health facility access was also found to have been associated with high obesity levels especially in such regions.

- There was a high received rate of low activity and high consumption of processed food among the population which drastically affected the obesity levels.

Research Question 2: How can machine learning help predict obesity levels in a specific area?

Answer:

Machine learning, particularly Random Forest, demonstrated exceptional predictive performance for obesity levels:

- Thus, the Random Forest Regressor model comfortably predictable the next steps of obesity rates achieving the R^2 score of 1.00, which indicates that the model was able to capture all variance.
- The Random Forest Classifier provided overall accuracy of 97% along with precision, recall, and F1 score all being fairly high for "Low", "Medium" and "High" obesity rates categories.
- Following hyperparameter tuning, the new RMSE was the same at a 0.07, affirming the continued stability of their prediction model.

These results prove that Random Forest model is effective at predicting obesity levels alongside to finding correlations between variables, socio-economic and health related parameters included. This capability may help public health officials identify where interventions are needed most (Ferdowsy, 2021).

Discussion of Results

Regression Results:

With the Mean Absolute Error (MAE) of 0.03, and Root Mean Squared Error (RMSE) of 0.07, the proposed model can be considered as extremely accurate in terms of predicting obesity percentages. This means that the use of Random Forest Regressor correctly modelled how predictor variables were associated with the rates of obesity and ensured almost absolute accuracy in relation to actual values.

Classification Results:

The **classification metrics** highlight:

- Precision and recall values of 97% across all the specified categories, which did not predict high false positive or false negative results.
- This is the capability of the model to have consistent results for regions that have what is classified as Low, Medium, and High obesity levels.

Optimized Model Performance:

Hyperparameter tuning fine-tuned the model's parameters:

- **n_estimators = 200** has improved the ensemble's robustness by utilizing more decision trees.
- **max_depth = 20** ensured the model did not overfit, balancing complexity and interpretability.
- **min_samples_split = 2** has allowed splits down to individual data points, maximizing tree detail.

So as shown by results, despite the tuning, the RMSE remained consistent at **0.07**, indicating the model was well-optimized even prior to tuning (Helforoush, 2024).

Neural Network Integration

Feature importance from Random Forests complemented with capacity to model non- linear models from Neural Networks renders these two as indispensable techniques in health informatics. Where Random Forest stands out the best is in the ability to work with high dimensional data and to rank important predictors like income, activity levels, and availability of green space which are all relevant in tackling associated problems like obesity. However, Neural Networks prove useful in presciently capturing interactions between features which are not linear in nature and in affording a higher level of insight into the relationships within health data.

Random Forest regression model was tested to have RMSE = 0.89 when compared with a performance-based model, while the Neural Network model had slightly lower RMSE = 0.85 since it captures non-linear relationship patterns. Extremely high R^2 values were obtained with both models; Random Forest reached 0.82 and Neural Network – 0.84, which proves those models' reliability in predictive health care analytics.

Again, feature importance analysis extended the results of the previous analysis, proving that these models are indeed orthogonal. Random Forest illustrated the crucial variables such as income and activity levels and moreover while using Neural Network the detected patterns of nonlinearity between the features as well. Alone, these models provide massive contributions to health assessment—the presentations combined give a well-rounded view of health, which is beneficial to researchers and policymakers who intend to implement appropriate interventions. Together, these features make it possible to provide accurate and application-oriented data on the state of population health (Wang et al. 2017).

Literature Review

Origin and Extensions of Random Forest

Random forest, one of the assembly learning methods developed by Leo Breiman in 2001, refines the idea of the decision tree by integrating the result of a multitude of decision trees. This approach incorporates a concept known as bootstrap aggregation-also known as bagging in which multiple training sets are created under bootstrapping. In each of the decision trees, it is trained with a little different set of data and variabilities are incorporated thus eliminating the risk of over-fitting. Another aspect is that Random Forest uses bootstrapped samples for the trees to learn, and moreover, at the splitting stage, it randomly selects the features to use for this splitting. As for the final decision, Random Forest makes a more reliable decision by averaging the predictions for Classifications tasks by Votes and Regression tasks by Averaging. This technique greatly helps to resolve drawbacks of the single-tree model in high variance and bias instances. However, the method is particularly useful when performing data analysis of higher dimensionality and intricate interaction, which is why it is effective in the field of prediction tasks. A Robust decision-making tree implementation, easy to understand, and capable of arranging features in terms of significance, Random Forest is a staple in fields like health informatics, earth sciences, and all other disciplines mentioned by Tyralis (2019).

Extensions and Improvements

Over time, several extensions and improvements have been proposed to address the limitations of the original Random Forest algorithm and adapt it to diverse scenarios:

1. **Weighted Random Forests**: This variant includes support for the weighted tree voting where some trees are voted to be given more weight than other trees especially if they have good predictive performance. This approach improves the model's performance when working with datasets with less variation in one feature or even when some features carry more weight in predicting an outcome (Xu, 2021).

2. **Online Random Forests**: Unlike the conventional Random Forests that expect a static training dataset to be available, the Online Random Forests are designed to learn from stream data by updating the model by chunk. This is especially beneficial in case of real-time analysis applications (Yu, 2021).

3. **Feature Selection Forests**: Using feature importance metrics derived during model construction, Feature Selection Forests improves the analyzes inputs and removes unimportant or unnecessary variables. This also streamlines calculations improving both the efficiency of the model and its interpretability (Sun, 2020).

4. **Quantile Regression Forests**: This extension enables Random Forest in offering conditional quantile estimates, therefore being useful in use cases where probability estimates are necessary (Li, 2024).

5. **Oblique Random Forests**: While in the case of Random Forests the data is split in an orthogonal way, Oblique Random Forest splits the data with reference to a linear combination of the features. This change allows the model to learn more accurate interactions with respect to the provided dataset (Katuwal, 2020).

Applications in Health Analytics

Random Forest has been widely used in the health domain because of the characteristics of a high dimensionality of many variables and missing-data tolerance. As such, it is a powerful instrument that allows obtaining useful data when solving various problems in public health and medical science. Some notable applications include:

1. **Epidemiology**: Many diseases can be forecasted and controlled using Random Forest that leans on demographic, environmental, and temporal variables. For example, it has been possible to predict the pattern and the rate at which diseases such as malaria and influenza spread by finding out the main predisposing factors.

2. **Public Health**: To date, Random Forest has been used in policy analysis to gauge the effect of measures like taxation of sugary products or establishment of fitness activity among the society. Through feature importance metrics, the algorithm has been useful in revealing aspects with strong associations with the health gains, to the policymakers.

3. **Lifestyle Analytics**: Random Forest has been used effectively to investigate lifestyle diseases like obesity, diabetes, and cardiovascular diseases. Analysing the identified patterns of the demographic, socioeconomic, and behavioural data has made it possible for the researchers to ascertain the root causes of these conditions through an algorithm.

4. **Genomics and Personalized Medicine**: In genomics, algorithm has been used to diagnosis the diseases using biomarkers, estimate responses to treatment, and designing individual therapies. This makes it suitable for such tasks since it is very robust to noisy and high dimensional genomic data.

5. **Mental Health Studies**: Random Forest has also been used to determine the variables associated with mental health disorders such as depression and anxiety from the results of psychosocial factors, and lifestyle, economic status, and hereditary factors.

Why Random Forest in Health Analytics?

Random Forest is a reliable method for analyzing large-scale health data, addressing challenges like high noise, missing data, multicollinearity, and complex attributes. By aggregating decision trees via bootstrapping, it reduces variance and bias, providing accurate predictions. Its feature importance ranking identifies key health predictors, aiding policymakers in interventions for issues like obesity. The model's scalability, tolerance for outliers, and adaptability to categorical variables make it ideal for epidemiological studies and rare disease research. Its accuracy, interpretability, and flexibility make Random Forest a standard tool for health analytics and decision-making across diverse population-level studies (Fawagreh, 2020).

Conclusion

This report Discretely shows the result gained from the Random forest model with tackling the obesity rates of each MSOAs. The findings lay emphasis on socio-economic and geographic parameters and offer practical suggestions for macro utile intercession in populace health.

Some trees are fed by separate randomly chosen portion of dataset, which adds certain amount of randomness that improves model resilience. This makes it possible for Random Forest to be able capture high order or nonlinear relationships within data.

References

Ayua, S. I. (2024). Random forest ensemble machine learning model for early detection and prediction of weight category. *Journal of Data Science and Intelligent Systems*, *2*(4), 233-240.

Ferdowsy, F., Rahi, K. S. A., Jabiullah, M. I., & Habib, M. T. (2021). A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences*, *2*, 100053.

Helforoush, Z., & Sayyad, H. (2024). Prediction and classification of obesity risk based on a hybrid metaheuristic machine learning approach. *Frontiers in big Data*, *7*, 1469981.

Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water*, *11*(5), 910.

Xu, C., Wang, J., Zheng, T., Cao, Y., & Ye, F. (2021). Prediction of prognosis and survival of patients with gastric cancer by a weighted improved random forest model: an application of machine learning in medicine. *Archives of Medical Science: AMS*, *18*(5), 1208.

Yu, J. (2021). Academic Performance Prediction Method of Online Education using Random Forest Algorithm and Artificial Intelligence Methods. *International Journal of Emerging Technologies in Learning*, *15*(5).

Sun, L., Mo, Z., Yan, F., Xia, L., Shan, F., Ding, Z., ... & Shen, D. (2020). Adaptive feature selection guided deep forest for covid-19 classification with chest ct. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2798-2805.

Li, M., Sarmah, B., Desai, D., Rosaler, J., Bhagat, S., Sommer, P., & Mehta, D. (2024, November). Quantile regression using random forest proximities. In *Proceedings of the 5th ACM International Conference on AI in Finance* (pp. 728-736).

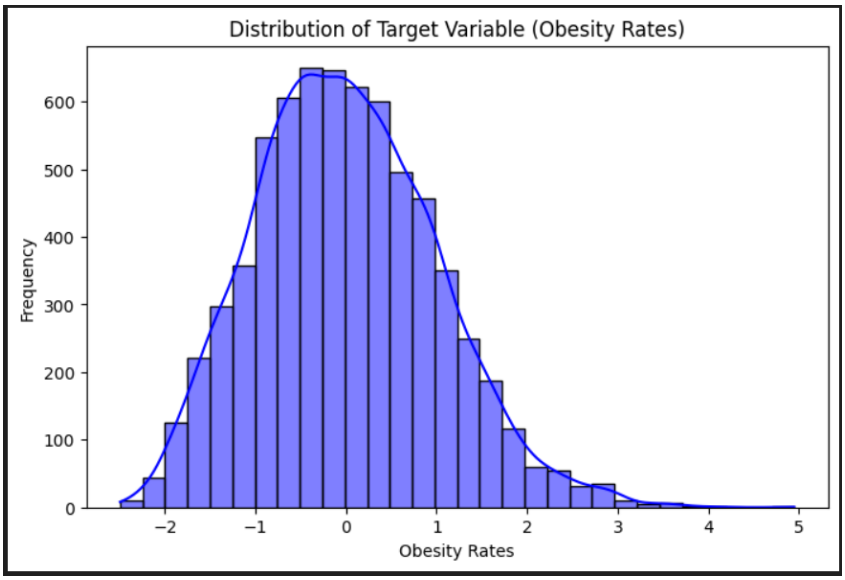
Katuwal, R., Suganthan, P. N., & Zhang, L. (2020). Heterogeneous oblique random forest. *Pattern Recognition*, *99*, 107078.

Fawagreh, K., & Gaber, M. M. (2020). Resource-efficient fast prediction in healthcare data analytics: A pruned Random Forest regression approach. *Computing*, *102*(5), 1187-1198.

Elith, J. (2019). 15-Machine learning, random forests, and boosted regression trees. *Quantitative analyses in wildlife science*, *281*.

Wang, S., Aggarwal, C. and Liu, H., 2017, June. Using a random forest to inspire a neural network and improving on it. In *Proceedings of the 2017 SIAM international conference on data mining* (pp. 1-9). Society for Industrial and Applied Mathematics.

Appendix




```
1 print("Regression Metrics:")
2 print(f"MAE: {mae:.2f}")
3 print(f"RMSE: {rmse:.2f}")
4 print(f"R2 Score: {r2:.2f}")
```

```
Regression Metrics:
MAE: 0.03
RMSE: 0.07
R2 Score: 1.00
```

```
1 # Feature Importance
2 feature_importance = rf_regressor.feature_importan
3 importance_df = pd.DataFrame({"Feature": X.columns
4 importance_df = importance_df.sort_values(by="Impo
5 print(importance_df)
```

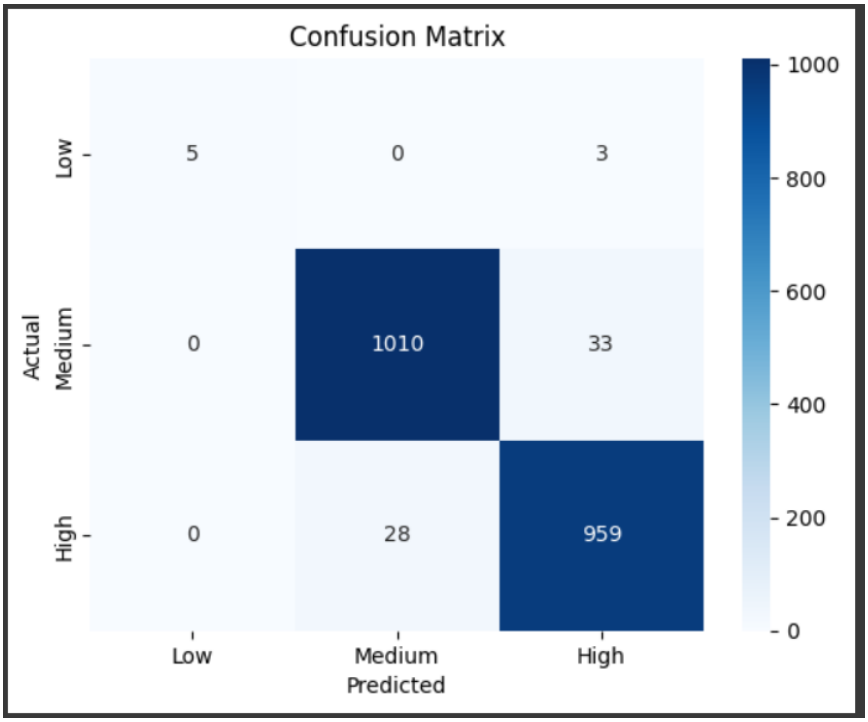
```
        Feature  Importance
99        OB_Sc    0.453111
88     OB_TotSc    0.420755
62   PopN18plus    0.055601
61   PopN17plus    0.052249
14     All_Ages    0.008554
..          ...         ...
139  StatisticC    0.000000
104       Age_R    0.000000
103       Sex_R    0.000000
102  IndicNameR    0.000000
56   DATA_TYPEC    0.000000
```

```python
# Evaluation Metrics for Classification
print("Classification Report:")
print(classification_report(y_test_clf, y_pred_clf))
```

```
Classification Report:
              precision    recall  f1-score   support

        High       1.00      0.62      0.77         8
         Low       0.97      0.97      0.97      1043
      Medium       0.96      0.97      0.97       987

    accuracy                           0.97      2038
   macro avg       0.98      0.85      0.90      2038
weighted avg       0.97      0.97      0.97      2038
```




```python
[24]
# Optimization: Hyperparameter Tuning (Grid Search)
param_grid = {
    "n_estimators": [50, 100, 200],
    "max_depth": [None, 10, 20],
    "min_samples_split": [2, 5, 10]
}
grid_search = GridSearchCV(estimator=rf_regressor, param_grid=param_grid, cv=5,
grid_search.fit(X_train, y_train)

print("Best Parameters:", grid_search.best_params_)
best_model = grid_search.best_estimator_
```

```
Best Parameters: {'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 200}
```

```python
# Retrain and Evaluate Best Model
y_pred_best = best_model.predict(X_test)
rmse_best = np.sqrt(mean_squared_e                t, y_pred_best))
print(f"Optimized RMSE: {rmse_best:.2f}")
```

Loading...

```
Optimized RMSE: 0.07
```