Predicting Obesity Levels Based on Lifestyle Factors

Presented By
Sunera Athukorala
Steph Adey
Laura Liu
John Robertson

BACKGROUND OF THE STUDY | OBESITY EPIDEMIC

Overweight and obesity refers to excess body weight. It is a risk factor for many chronic conditions and is associated with higher rates of death.

31.3% of Australian adults live with obesity.

In 2022, 1 in 8 people in the world were living with obesity.

Obesity is a complex issue with many causes. It's caused when extra calories are stored in the body as fat.

DATA COLLECTION & PROCESSING

1. Data sourced from UC Irvine Machine Learning Repository using ‘unimlrepo’
2. Data processing, cleaning and feature engineering to apply defined conversions and handle missing values
3. Data Transformation converting categorical variables to numerical and scaling to ensure comparable
4. Split the data into training and testing sets for machine learning model training and evaluation
5. Support Vector Machine (SVM)
6. Logistic Regression model with the use of GridSearchCV
7. Neural Network
8. Model comparison

FEATURE ENGINEERING

Data Conversion Functions

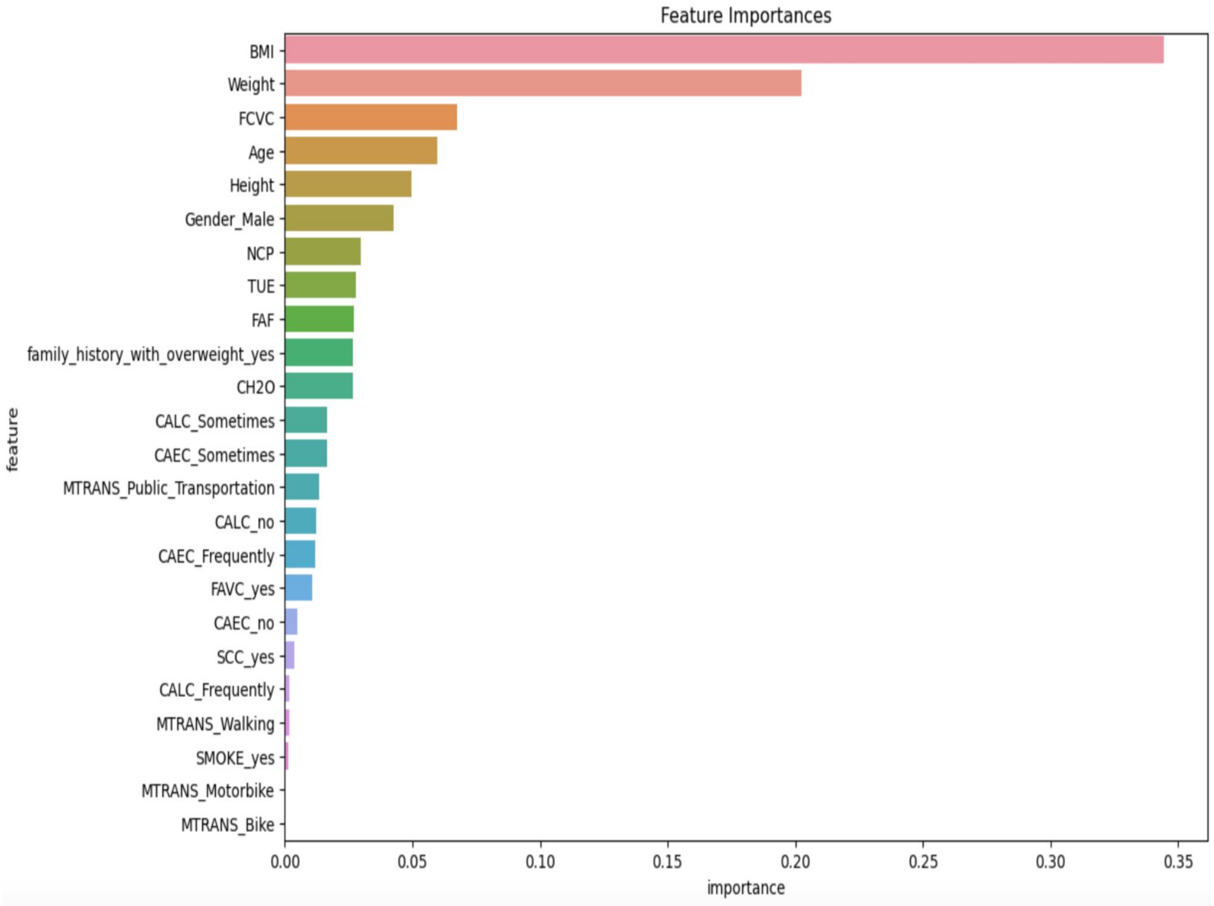
- **Binary Conversion**: Transformed categorical variables with 'yes' and 'no' responses into binary values (1 for 'yes' and 0 for 'no').
- **Weight Replacement**: Mapped descriptive weight categories (e.g., 'insufficient', 'normal', 'overweight', 'obesity') to numerical values (0 to 3).
- **Integer and Categorical Buckets**: Grouped continuous variables into discrete buckets to simplify data representation.
- **Encoding Techniques**: Used pd.get_dummies to encode categorical variables into numerical values.

Data Preprocessing Steps

- **Rounding Values**: Applied rounding to certain features like age and water consumption to reduce precision and standardise the data.
- **Scaling Features**: Standardised features using StandardScaler to ensure all features have a mean of 0 and a standard deviation of 1.
- **Data Splitting**: Divided the dataset into training and testing sets, ensuring an appropriate split for model training and evaluation.

Feature Selection and Model Preparation

- **Feature Selection**: Dropped irrelevant features such as 'BMI', 'Height', and 'Weight' from the dataset to focus on the most impactful features.
- **Target Variable Transformation**: Converted the target variable 'NObeyesdad' to a more intuitive 'Obesity_Level' and cast it as an integer type.



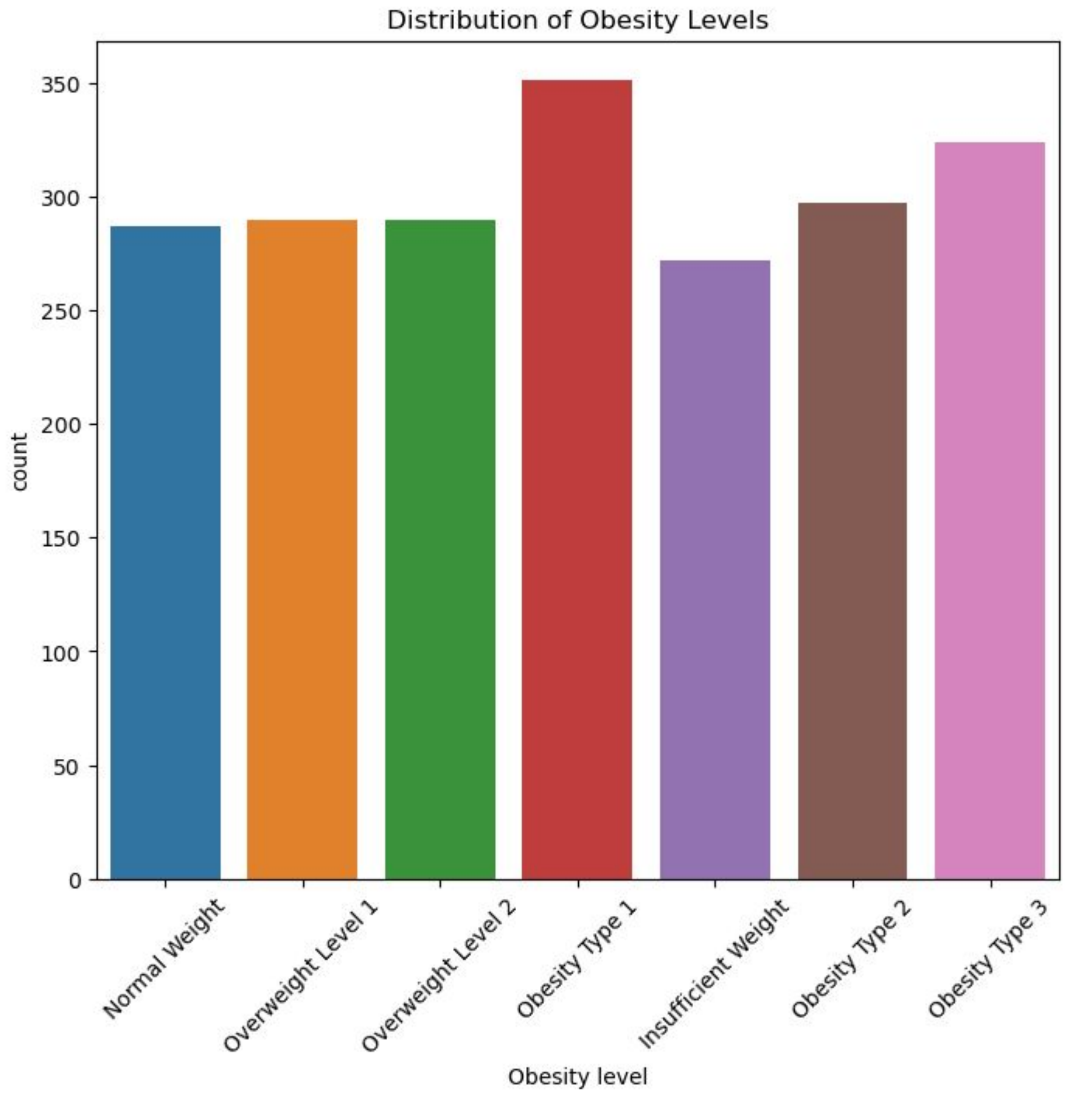
EXPLORATORY DATA ANALYSIS

Distribution of obesity levels:

- This bar chart shows the distribution of obesity levels across six categories.
- "Obesity Type I" has the highest count, suggesting that mild obesity is quite prevalent.
- "Obesity Type II" and "Obesity Type III" also have substantial counts, indicating the presence of severe obesity.

Conclusions:

- High Prevalence of Obesity: Shows a high prevalence of obesity, with "Obesity Type I" being the most common.
- Balanced Distribution in Overweight Categories: The counts for "Overweight Level I" and "Overweight Level II" are relatively balanced, indicating that overweight individuals are fairly common.



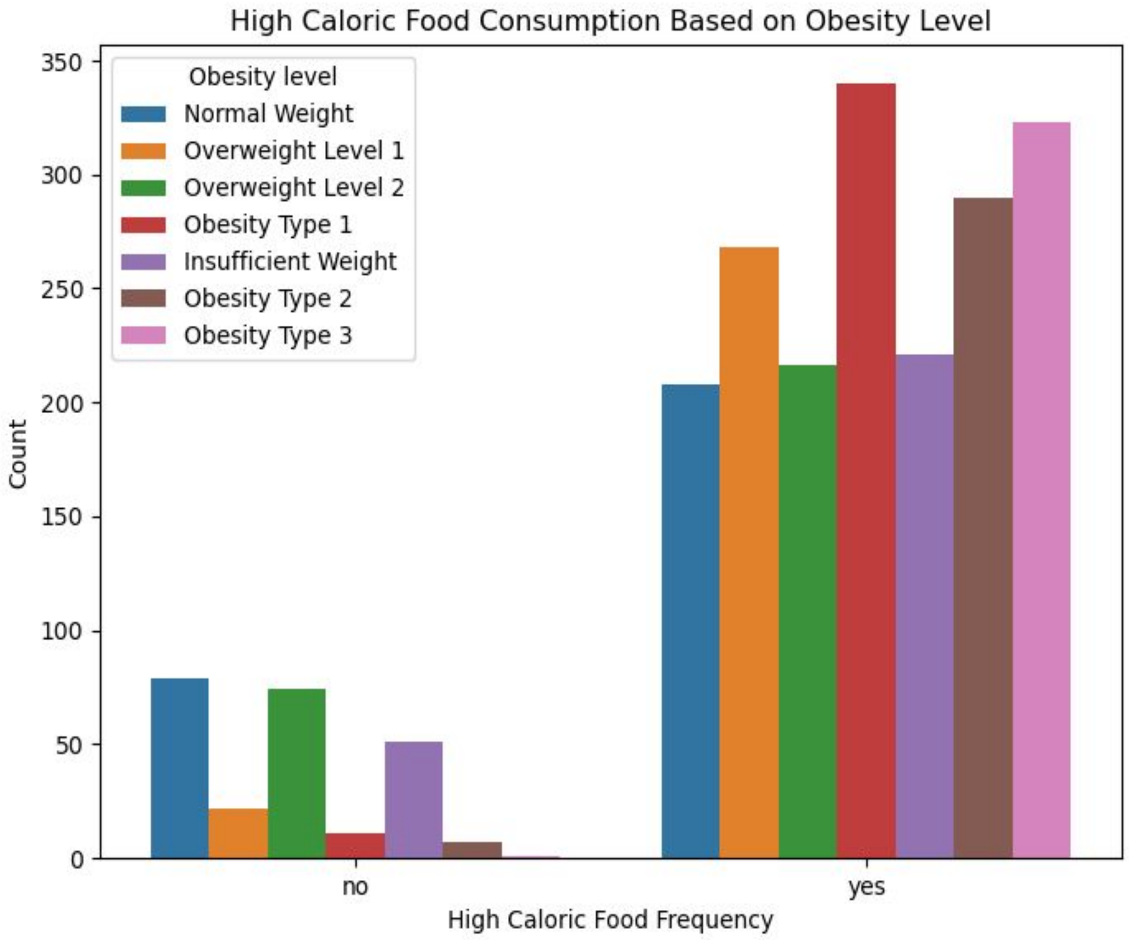
EXPLORATORY DATA ANALYSIS

High Calorie Food Consumption based on obesity levels:

- The bar plot above shows the relationship between high caloric food consumption and obesity levels.
- High Caloric Food Frequency ("**no**"):
  - The majority of individuals who do not frequently consume high caloric food fall into the "Normal Weight" category.
- High Caloric Food Frequency ("**yes**"):
  -A significant portion of individuals who frequently consume high caloric food fall into the "Obesity Type I" category

Conclusions:

- There is a strong association between frequent high caloric food consumption and higher obesity levels. The majority of individuals in the "**Obesity Type I**", "**Obesity Type II**", and "**Obesity Type III**" categories frequently consume high caloric food.



MODEL SELECTION, TRAINING & EVALUATION

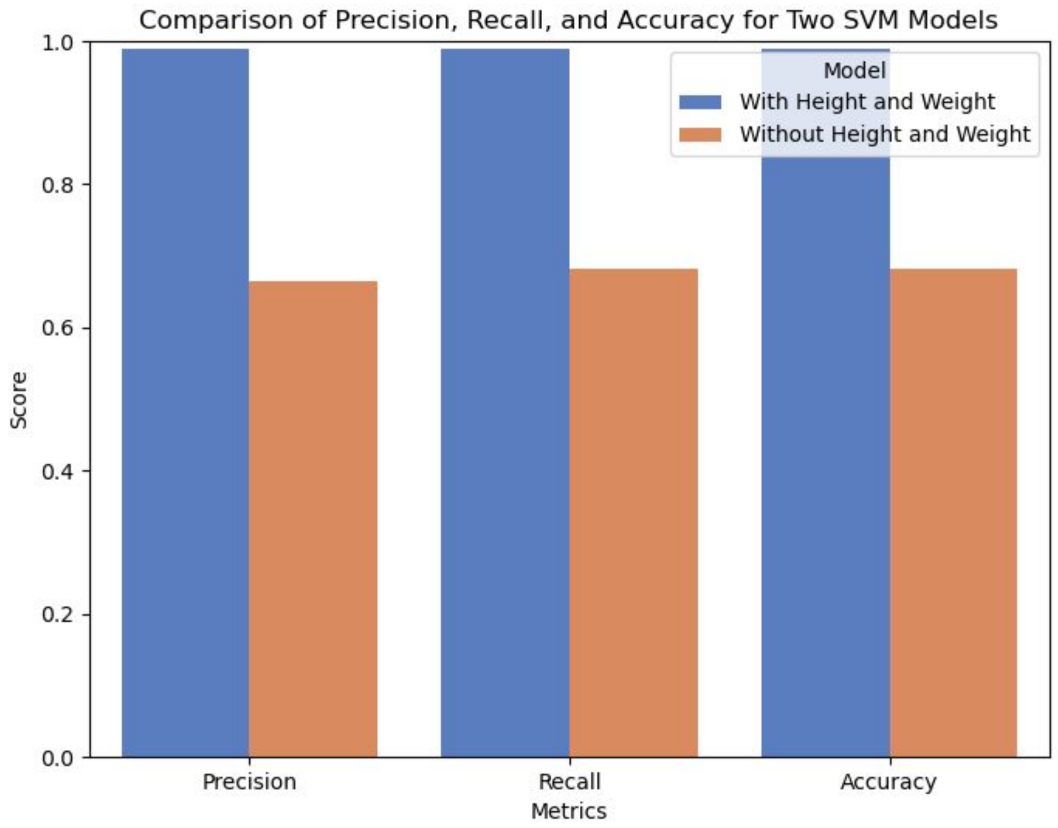
Support Vector Machine (SVM) for Predicting Obesity Levels:

Modelling Technique:

- SVM is a supervised machine learning algorithm used for classification.
- It works by finding the hyperplane that best separates the data into different classes
- The features are various attributes related to individuals' lifestyle and physical characteristics, which are used to predict obesity levels such as age, height, weight and e.t.c

Model Accuracy:

- **High Accuracy with Height and Weight**: The model that includes height and weight as features achieves a very high accuracy of 99%. This suggests that these features are significant predictors of obesity levels in the dataset.
- **Lower Accuracy without Height and Weight**: When height and weight are excluded from the features, the model's accuracy drops significantly to 68%. This indicates that without these key physical lifestyle factors, the model's ability to accurately classify obesity levels is compromised



MODEL SELECTION, TRAINING & EVALUATION

Logistic Regression for Predicting Obesity Levels:

Training and Predictions

- **Training**: Logistic Regression model trained with 1000 iterations and fixed random state for reproducibility.
- **Predictions**: Model made predictions on the test set to evaluate generalisation.

Initial Evaluation

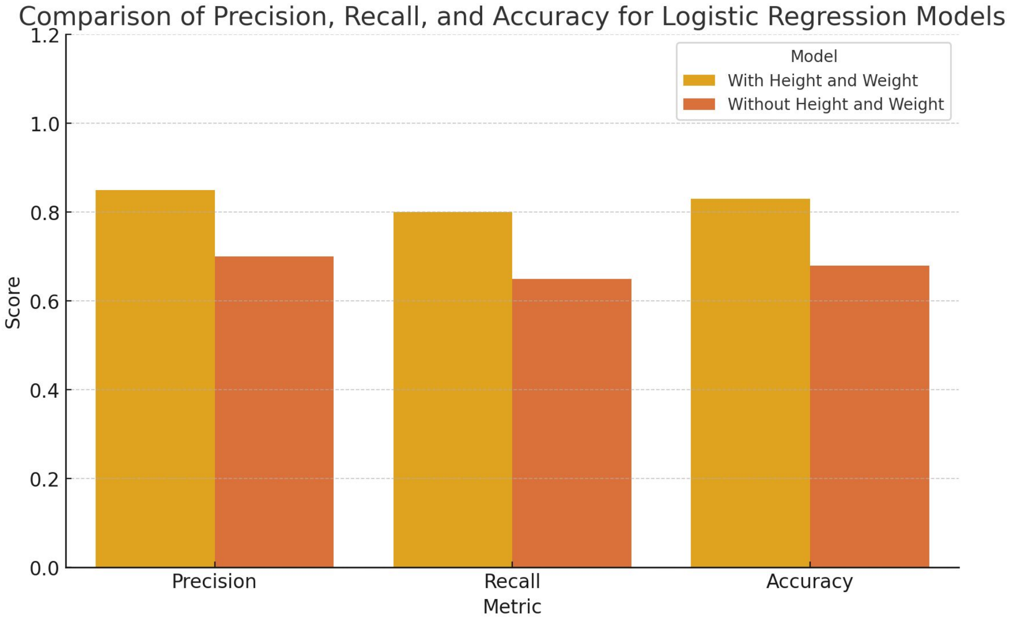
- **Accuracy**: Indicates satisfactory correct classification rate.
- **Precision**: Reasonable proportion of correct positive predictions.
- **Recall**: Effective identification of positive instances.
- **F1 Score**: Balanced performance between precision and recall.

Model Optimisation

- **Grid Search**: Hyperparameter tuning conducted to find best values for regularisation strength (C) and solver.
- **Best Parameters**: Identified through cross-validation for improved accuracy.

Optimised Model Evaluation

- **Improved Metrics**: Optimised model showed better accuracy, precision, recall, and F1 score.



MODEL SELECTION, TRAINING & EVALUATION

Neural Network (NN) for Predicting Obesity Levels:

Training and Predictions

- **Training**: Two NN Models were trained in parallel, one *without* and one *with* BMI, height & weight.
- **Predictions**: Two sets of predictions were made:
  - *Normal Weight* or *Overweight/Obese*
  - *Overweight* or *Obese*

Initial Evaluation

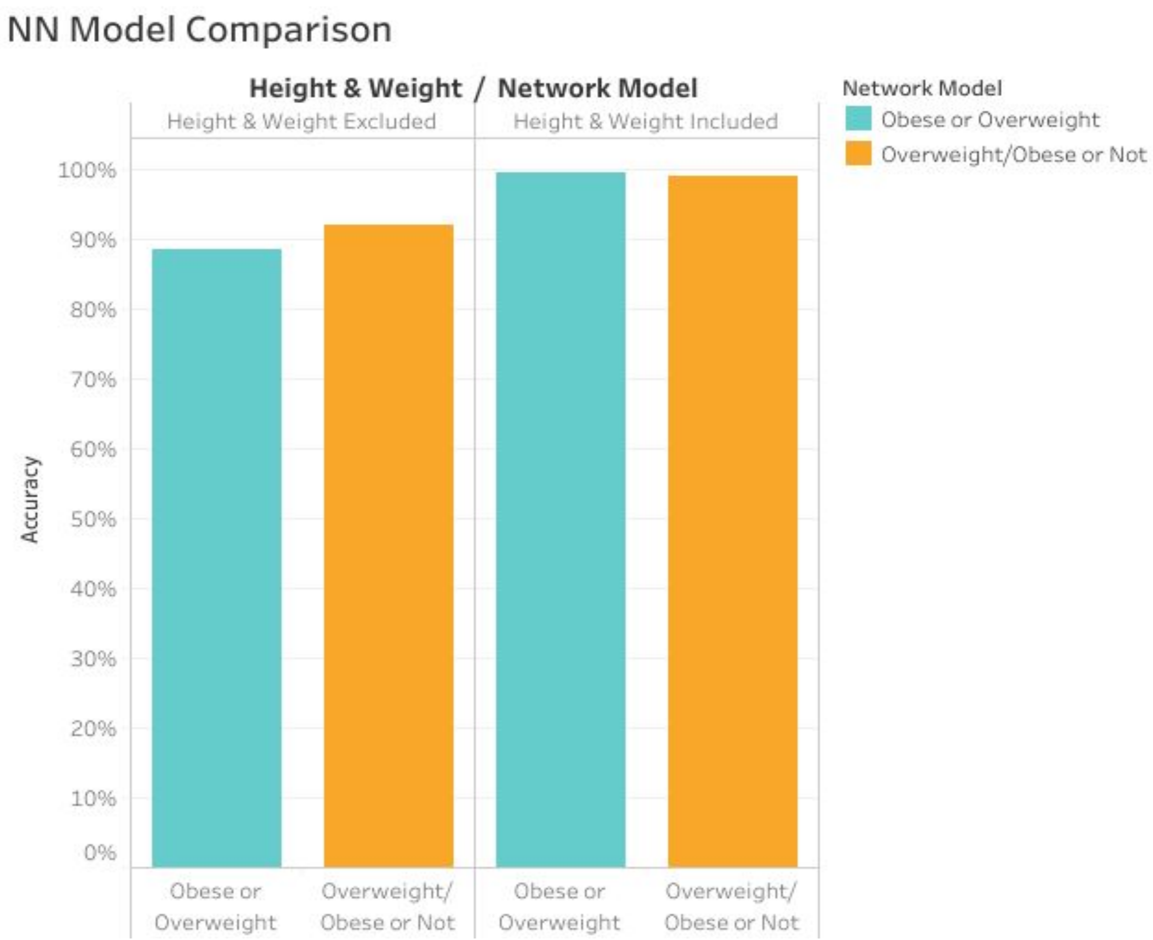
- **Accuracy**: How well a model can predict outcomes.
- **Loss**: The difference between the model's predictions and true outcomes.

Model Optimisation

- **Activation Functions**: ReLu, Softmax or SeLu
- **Neurons**: 25 - 40, in increments of 5
- **Hidden Layers**: 2 - 3

Optimised Model Evaluation

- **Optimised Outcome**: NN Models with similar accuracies yielded similar loss values. The model with the highest accuracy score was identified as the most optimal for each model iteration.



MODEL SELECTION, TRAINING & EVALUATION

Logistic Regression Modelling

Variables to optimise:

- ***C*** - Inverse of regularisation strength.
- ***Solver*** - Algorithm to use in the optimisation problem.

Evaluated against:

✓ Accuracy score
✓ Precision
✓ F1 Score
✓ Recall Score

Support Vector Machine (SVM)

Evaluated against:

✓ Accuracy score
✓ Precision
✓ Recall Score

Effective models:
✓ ***Maximise accuracy***
✓ ***Maximise Precision / Reduce Loss***

Neural Network (NN)

Variables to optimise:

- ***Activation Function*** - Function applied to neuron output in layer
- ***Neurons*** - Applies weights & biases using function
- ***Hidden layers*** - Layers where computation is done

Evaluated against:

✓ Accuracy score
✓ Loss

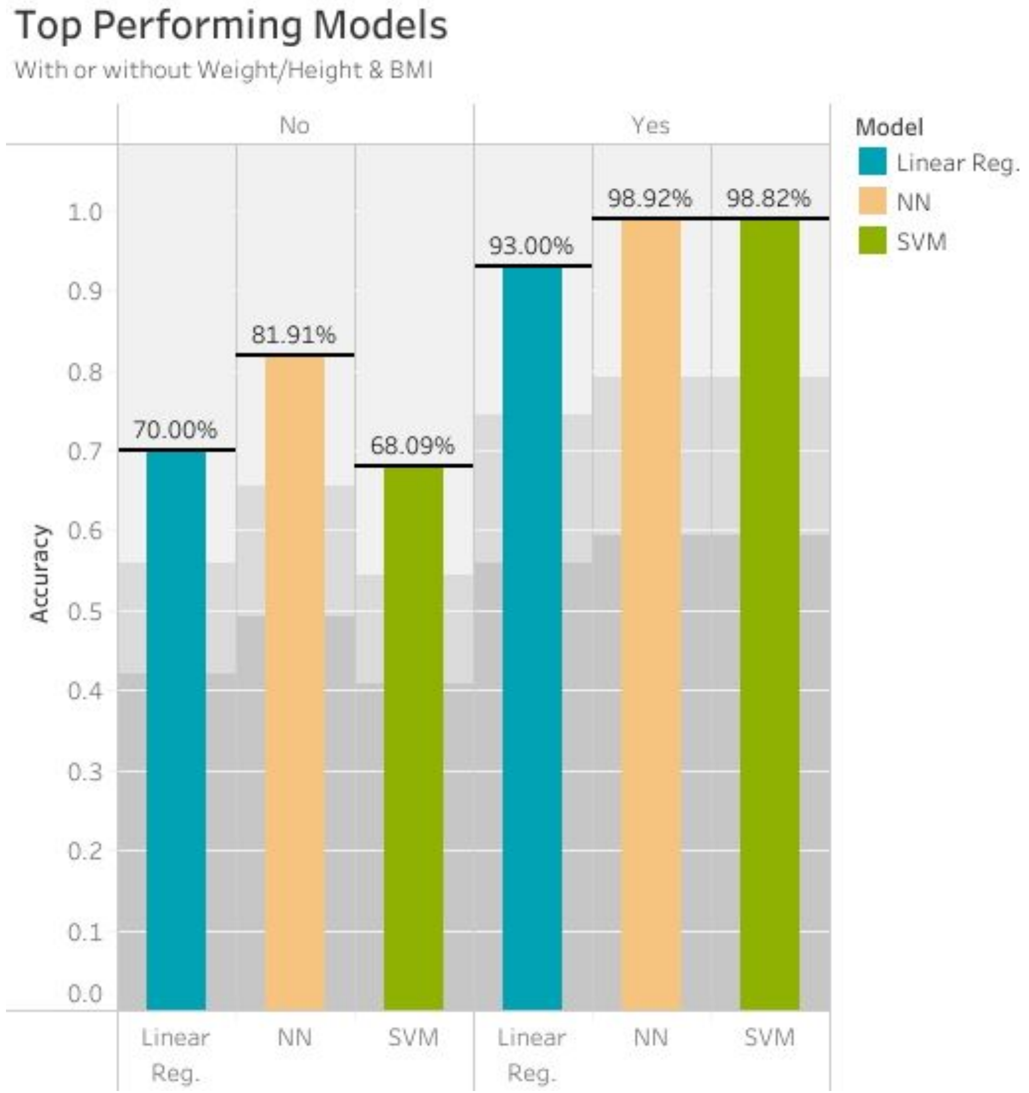
RESULTS & COMPARISON

- All models performed incredibly well, >93%
- Height, weight & BMI are physical indicators directly linked to BMI
- Physical indicators dominate all other lifestyle factors in predicting obesity

Including Height, Weight & BMI

Excluding Height, Weight & BMI

- All models performed well.
- Neural Network model accurate above 81%
- Lifestyle factors can be obtained easily through patient questionnaires
- Omitting medical indicators reduces the stigma or anxiety some patients face from health classification.

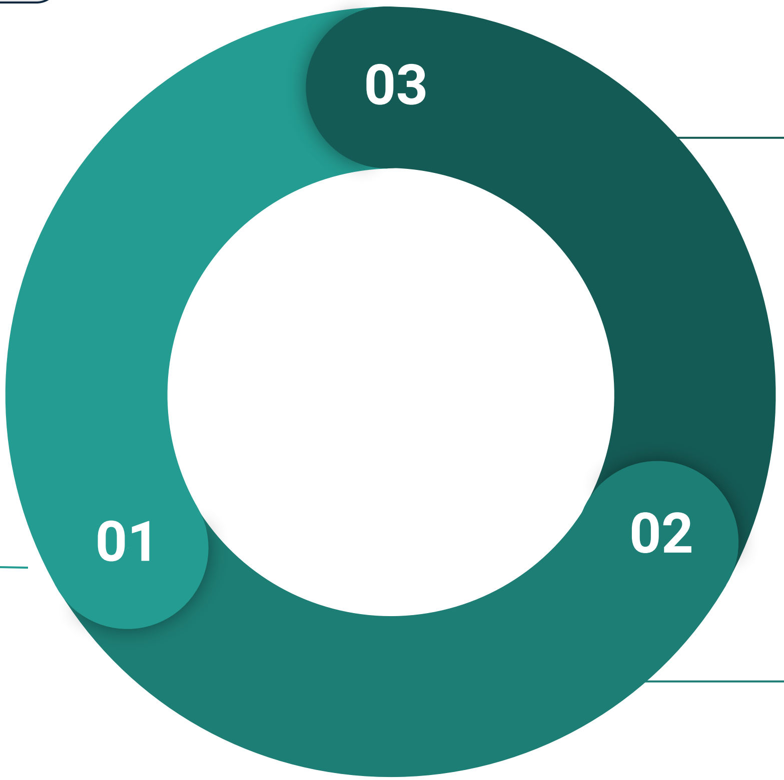


MODEL APPLICATION

Community Deployment

Identify & Examine

- ➢ Questionnaire on lifestyle choices
- ➢ In-person or remote
- ➢ Reduces examination time
- ➢ Removes travel requirements
- ➢ Improves access



Assess Outcome

- ➢ Reassess at end of treatment regime
- ➢ Complete questionnaire
- ➢ Measure progress
- ➢ Suggest treatments based on outcome

Classify & Designate

- ➢ Medical treatment regime
- ➢ Social programs
- ➢ Living assistance
- ➢ In-home education
- ➢ Improve access to equipment & amenities

Thank you for your attendance

Presented By
Sunera Athukorala
Steph Adey
Laura Liu
John Robertson