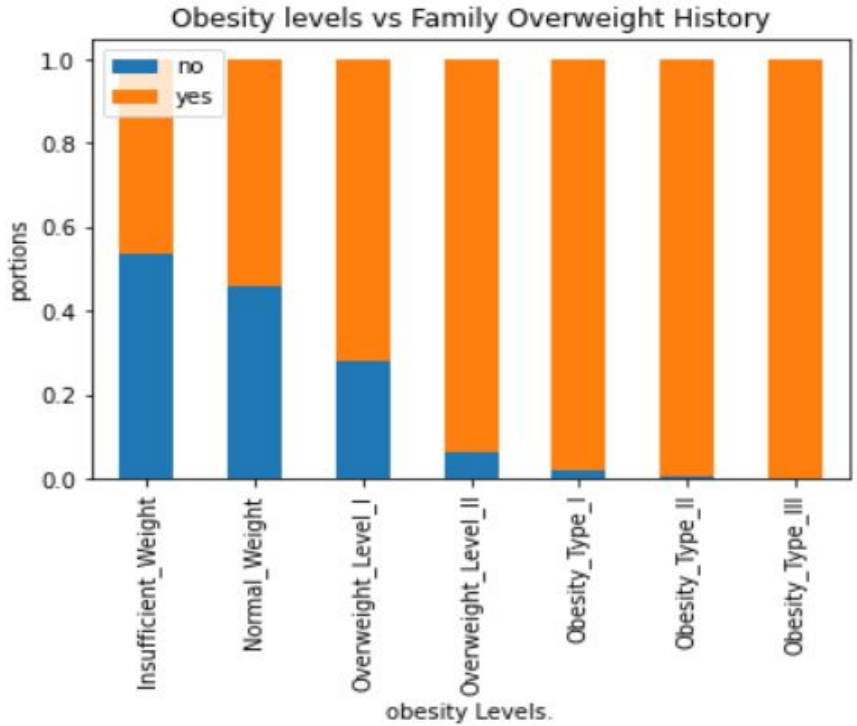
Predicting Obesities Based on Lifestyles and Family History

Tianqi Tang
Brown University
October 15th, 2020
https://github.com/TianqiT/data1030-fa20-project

Introduction

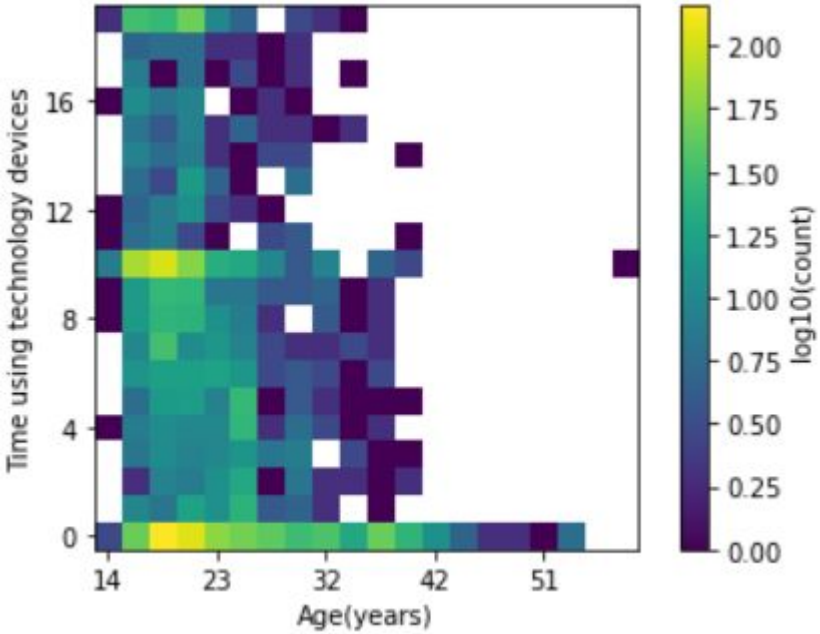
- **Why this project:**
  - Obesity is a major health issue.
  - It is essential to study how obesity is associated with different factors.
- **Data:** From UCI ML Repository.
  - 23% collected from web users.
  - 77% generated using Weka tool and SMOTE filter.
  - Target: Obesity Levels.
- **Goal**: Classify obesity levels based on 16 features.

How family members affect obesity?



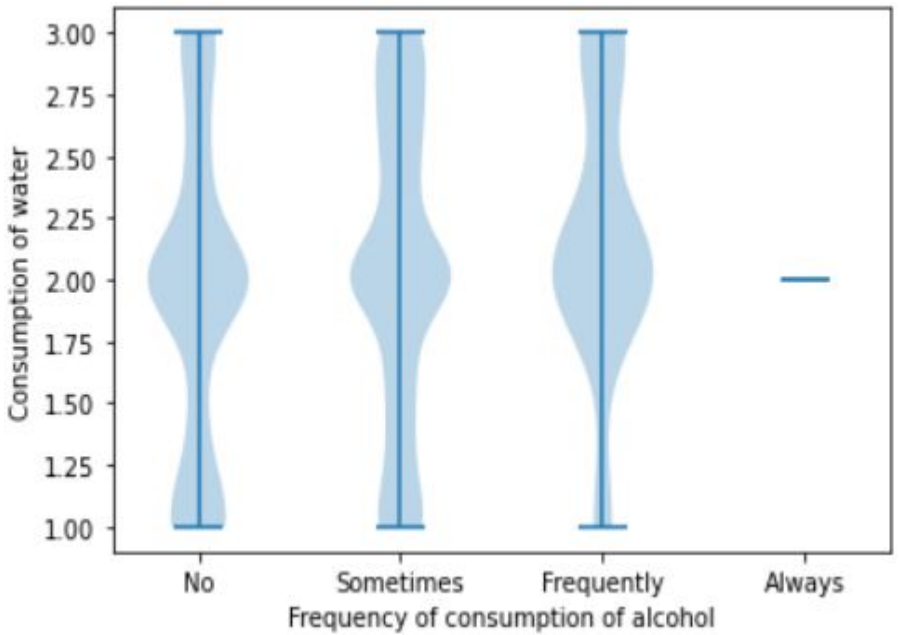
- Positive correlation between obesity level and family overweight history.

Age vs Technological Devices Usage



- Data collected mostly from young people.
- Older people don't tend to use technological devices much.
- **Caveat:** Many features like devices usage were categorical then preprocessed to continuous.

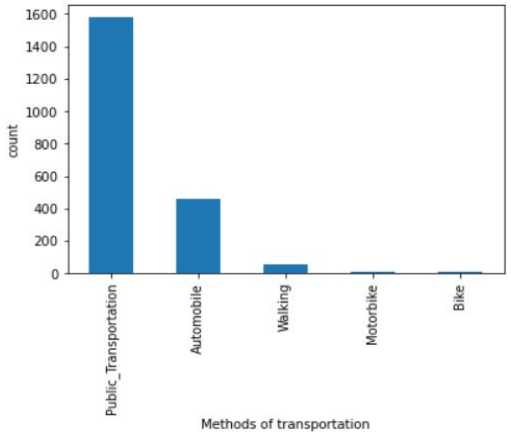
Alcohol vs Water Intake



- Severe data imbalance. Only one person who always drinks alcohol!
- Few people who frequently drink alcohol also drink little water.

Data Splitting

- 2111 points, 16 features, IID, no missing values, no group structures.
- 64%, 16% and 20% for training, validating and testing.
- Some have imbalances, need stratified split.
  - Only 1 person who always drinks alcohol. Stratified split no use.
  - Stratify around the second most imbalanced feature: MTRANS.



Data Preprocessing

- Categorical features with ordered structures → **OrdinalEncoder**
- Categorical features without ordered structures → **One-HotEncoder**
- Continuous features → **StandardScaler**
- Target variable → **LabelEncoder**

16 Features → 25 Features

References

- Data Set: Palechor, F. M., de la Hoz Manotas, A. (2019)., *Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico*. Data in Brief, 104344.

  https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+

- Related Work: De-La-Hoz-Correa, E., Mendoza-Palechor, F. E., De-La-Hoz-Manotas, A., Morales-Ortega, R. C. Beatriz Adriana, S. H. (2019). *Obesity Level Estimation Software based on Decision Trees*, Journal of Computer Science, 15(1), 67-77. https://doi.org/10.3844/jcssp.2019.67.77