Annotated Bibliography

Maha Mapara

3rd November, 2020

1 Approaches to predicting health outcomes

1.1 Arandjelovic, Ognjen.(2015).Prediction of health outcomes using big (health) data. 2015. 10. 1109

This paper proposes a model of disease progression by predicting the probability of a specific hospital admission following the patient history. The premise of the model is that future development of a patient's states can be predicted by the presence of a diagnosis of a condition at a point in the past. Electronic medical records from hospital databases are used as the data. The model used is a mathematical model similar to Markov process models. The results show that the model performs with a 91% accuracy to predict the next hospital admission and with a 82% accuracy for long term admission predictions.

As my project is not focused on disease progression, this source is not useful to me. However, it helped me understand how I should be phrasing my project, because earlier I was phrasing my proposal as predicting health outcomes for diseases associated with modifiable risk factors.

1.2 Jenkins, D.A., Sperrin, M., Martin, G.P. et al. Dynamic models to predict health outcomes: current status and methodological challenges. Diagn Progn Res 2, 23 (2018). https://doi.org/10.1186/s41512-018-0045-2

The focus of this paper are clinical prediction models that evolve over time in response to observed changes in a patient. The authors conducted a literature review in this topic area to understand the current dynamic prediction modelling and identify the methodological challenges. They concluded that dynamic prediction models are better than static clinical prediction models; but there are challenges in validating dynamic models and choosing hyper parameters.

This paper is not useful for my project as I am not predicting health outcomes. But it discusses dynamic and static modelling which is something to consider for future iterations of my work as a dynamic model will be better way of predicting the presence of a disease in a population based on changing factors like food access, demographics, diet etc.

1.3 Venkatesh, R., Balasubramanian, C., Kaliappan. (2019). Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique. Journal of Medical Systems.43.10.1007/s10916-019-1398-y.

This paper discusses how machine learning techniques can be used to to predict the future health status of a disease like heart disease. It particularly mentions using Naive Bayes to get a probabilistic classification based on Bayes' theorem; and constructing a classifier model to predict future health conditions for heart disease. They showed how this method can predict the future health conditions for different patients, and its use in early disease detection.

Even though my project is not based on predicting disease in current patients, this paper highlights an interesting machine learning approach that can be applicable on my data set to predict the prevalence of disease in a population. The related works mentioned in the paper are also relevant to my work as I am using multiple data sources like that in the works mentioned. Approaches using neural networks, support vector machines, k-means clustering, and ensemble classification were also examined. Also, the section on implementing the Naive Bayes algorithm and results is helpful as it can guide my implementation of the method.

2 Approaches for obesity prediction

2.1 Dunstan, Jocelyn Aguirre et al. (2019). Predicting nationwide obesity from food sales using machine learning. Health Informatics Journal. 26. 146045821984595.

This paper discusses using country-level food sale data from 79 countries to predict obesity prevalence. The authors used three machine learning algorithms: Support Vector Machines (SVM), Random Forests and extreme Gradient Boosting. The random forest model had the best performance. They were also able to conclude that the food category most likely to predict obesity is baked goods and flours, followed by cheese and carbonated drinks. A brief description on their exploratory analysis using principal component analysis was also provided.

This is a useful paper for my project as I am looking at food access and quick service restaurant data to see how it contributes to obesity. I would like to look at the available data on food sales on a state- and county-level in the US. Moreover, the methods described here are consistent with some blogs and papers I have read, and will help me in deciding which algorithm will suit my project the best.

2.2 Singh, B., Tawfik, H. (2020). Machine Learning Approach for the Early Prediction of the Risk of Overweight and Obesity in Young People. 10.1007/978-3-030-50423-639.

This paper is looking at early prediction for obesity as a way of seeing how interventions like physical activity and better diet can prevent it; and consequently prevent development of diseases like type II diabetes and heart disease. The paper does a review of machine learning techniques that have previously been used to tackle obesity. Some of these methods are Naive Bayes, Random Trees/Forests, C4.5 Decisions Trees, Bayes Net, Artificial Neural Network (ANN), Genetic Algorithms, K-Nearest Neighbors (KNN).

The authors used UK's Millennium Cohort Study data and classified someone as at risk of obesity or not through information on BMI, age, socio-economic conditions, height etc. Multiple machine learning algorithms were implemented to see which provides the best classification results. These algorithms are: KNN, J48 pruned tree, Random forest, Bagging, SVM, Multilayer Perceptron (MLP), and Voting. They overcome data imbalance by using Synthetic Minority Oversampling Technique (SMOTE), and concluded that future work should include longitudinal and cross sectional data.

This paper is very helpful for my project as I am looking at predicting which groups/communities are at risk of obesity. The researchers work of consolidating multiple algorithms, reviewing previous work and conducting their own research will help me decide which algorithms I should use; and gives me more resources to look at.

2.3 Chatterjee, A.,Gerdes, M., Martinez, S. (2020). Identification of Risk Factors Associated with Obesity and Overweight-A Machine Learning Overview. Sensors.

This paper provides a review of the different machine learning methods and their implementation to analyze existing obesity/overweight related data. The goal is to use these methods to identify potential risk factors of obesity. First, 13 research papers regarding this topic were reviewed and the models used in these papers were presented. Second, the researchers used regression and classification analysis to visualize the change in lifestyle factors (tobacco consumption, sweet beverage use, fast food etc) as they relate to obesity. The ML algorithms they worked with are SVM, Naive Bayes, Decision Tree, Logistic Regression, KNN, Random Forest, and Linear Regression.

This paper looked at data sources that I am looking at like fast food restaurants, economic conditions and exercise. The review of related work cited two papers listed above, and others I had not previously seen. This will help expand my understanding of related works and ML algorithms that can be implemented in my project. Moreover, the paper describes the process of data pre-processing, training, and post-processing in a detailed manner. This makes it a good framework for my project work.

2.4 Jindal, K., Baliyan, N. Rana, P. (2018). Obesity Prediction Using Ensemble Machine Learning Approaches: Proceedings of the 5th ICACNI 2017, Volume 2.

This is a chapter from Recent Findings in Intelligent Computing Techniques. It describes using an ensemble prediction model to predict the obesity value. The ensemble model leverages the generalized linear model, random forest and partial least squares. The chapter gives some background to metrics used for obesity prediction like BMI and body fat percentage, an outline for the ensemble approach, and the methodology. The conclusion is that the model predicts obesity values with 89.68% accuracy.

This was somewhat helpful as it mentioned the ensemble approach; but the majority of the chapter was dedicated to basic techniques on exploring the data, and using R and python.

2.5 Dugan, T., Mukhopadhyay, S., Carroll, A., Downs, S. (2015). Machine Learning Techniques for Prediction of Early Childhood Obesity. Applied Clinical Informatics. 6. 506-520. 10.4338/ACI-2015-03-RA-0036.

This paper used data collected prior to a child's second birthday to predict childhood obesity (after 2 years of age) from a clinical decision support system. Six machine learning methods were used: Random Tree, Random Forest, J48, ID3, Naive Bayes, and Bayes Net. The ID3 model had the best accuracy at 85%. The structure of the tree brings into focus the best predictors of childhood obesity; one of these predictors is whether a child is overweight before 24 months.

ID3 and J48 trees are not algorithms I am familiar with, so learning about these and their implementation is useful for my project. This was also the first time I saw the ID3 perform better than other methods, so I will research on whether this model is a good fit for my data. Additionally, the description of the different analyses and model performance methods was a helpful review that will inform my analyses.