Heinz 95-845: Project Proposal

Wei Ji wji1@andrew.cmu.edu
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

Shuyi Zheng shuyiz1@andrew.cmu.edu
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

1. Proposal Details

1.1 What is your proposed analysis? What are the likely outcomes?

Our analysis will focus on constructing Machine Learning models to predict the individual-level Chronic Kidney Disease (CKD) on the dataset *National Health and Nutrition Examination Survey* 2013-2014(NHANES). The likely outcome is to predict one's kidney health status, namely no CKD or CKD stages 1-4, based on his or her demographic, dietary, questionnaires (living habits like physical activities), examination, and laboratory data.

1.2 Why is your proposed analysis important?

According to Coresh et al. (2007), CKD is now recognized as a common condition that elevates the risk of cardiovascular disease as well as kidney failure and other complications. The prevalence of CKD in the United States has increased significantly, partially due to the the increasing prevalence of diabetes and hypertension. Therefore automated tool for early prediction of CKD will help doctors verify and accelerate their diagnosis, decrease financial burden for patients for they might go through less clinical tests, and also provide a healthy life style guideline for everyone.

1.3 How will your analysis contribute to existing work? Provide references.

To our best knowledge, there is no previous research in predicting CKD specifically using NHANES data set. There are two innovative points in our analysis: 1. we incorporate more features as model inputs. In other related studies, like Vijayarani et al. (2015), Charleonnan (Anusorn), mostly used features are clinical and laboratory data. Since NHANES provides us a wider scope of data in health and nutritional status including diet, physical activity, etc, we can be more knowledgeable in evaluating individual CKD risk . 2. We provide more refined prediction labels on CKD. Previous research usually use CKD and non-CKD as output label, like Vijayarani et al. (2015), Charleonnan (Anusorn). Since it's of clinical value to classify different stages of CKD, our analysis is of contribution in this respect.

1.4 Describe the data. Please also define Y outcome(s), U treatment, V covariates, W population as applicable.

NHANES provides an ongoing survey and assessment for the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. We define the following list by referring to existing study and our analytic interest.

- **Y outcome**: Discretize *Estimated Glomerular Filtration Rate* as 1,2,3,4 stages of CKD or non-CKD. (Levey et al. (1999))

Variable selected (initial screening): BMI, age, gender, blood pressure, glucose, Cholesterol, Albumin, Creatinine, white/red blood cells, sodium, Potassium, living habit etc. We prefer not to identify treatment and covariates at this time. However, after further exploring the effect of the variables on the outcome, we might be interested in focusing on the effect of specific variables.

1.5 What evaluation measures are appropriate for the analysis? Which measures will you use?

We plan to report the confusion matrix and corresponding metrics, such as accuracy, precision, recall, and plot ROC curve and Precision-Recall curves to evaluate the ML models.

1.6 What study design, pre-processing, and machine learning methods do you intend to use? Justify that the analysis is of appropriate size for a course project.

This analysis provides us a good chance to practice the ML pipeline we learned in class to a specific domain in the health-care industry, and encourages us to explore some extension to the ML methods we learned so far. First, we need to get equipped with domain knowledge to understand the NHANES data set and the prediction task better. Second, data manipulation and conversion are necessary as data pre-processing, which might include missing data imputation, data type conversion, outliers removal, etc. Third, several different classification ML models will be trained and evaluated. The Machine Learning Methods we intend to use to tackle different stages along the pipeline are listed as follow.

Stage	Machine Learning Methods
Data Pre-processing, Exploratory Data Analysis	Missing Value Imputation, PCA (generate new features)
Classifiers	Logistic Regression, Naive Bayes Classifier, SVM, Decision Trees, Boosted Trees, Random Forest, Neural Networks, Nearest Neighbor, etc.
Hyper-parameter Tuning, Model Training	Cross-Validation, Regularization

1.7 What are possible limitations of the study?

Firstly, the data set we use is limited to 2013-2014, which cannot fully capture accumulative effects of some clinical features and living habits. Second, part of NHANES data are based on subjective response to interview questionnaires, which can be biased. Third, NHANES over-samples on persons elder than 60, African Americans, and Hispanics.

References

Coresh, Josef, et al. "Prevalence of chronic kidney disease in the United States." *Jama* 298.17 (2007): 2038-2047.

Vijayarani, S., and S. Dhayanand. "Data mining classification algorithms for kidney disease prediction." *International Journal on Cybernetics and Informatics (IJCI)* (2015).

Charleonnan, Anusorn, et al. "Predictive analytics for chronic kidney disease using machine learning techniques." Management and Innovation Technology International Conference (MITicon), 2016. IEEE, 2016.

Levey, Andrew S., et al. "A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation." Annals of internal medicine 130.6 (1999): 461-470.

National Center for Health Statistics. National Health and Nutrition Examination Survey: Questionnaires, datasets, and related documentation. Available from: https://wwwn.cdc.gov/nchs/nhanes/Default.aspx.