Modeling and Inference with Feature Importance for Assessing the Quality of Sleep among Chronic Kidney Disease Patients

Presented by: Surani Matharaarachchi

Joint work with:
Dr. Saman Muthukumarana, Dr. Mike Domaratzki, Dr. Chamil Marasinghe & Dr. Varuni Tennakoon

August, 12 2021

University of Manitoba

Outline

Chronic Kidney Disease (CKD)

- A gradual and irreversible loss of kidney function called chronic kidney failure.
- The treatment focuses on slowing the progress of kidney damage, usually by controlling the underlying cause.
- CKD can progress to end-stage kidney disease (ESKD).
- Decreased sleep quality is typical in CKD patients.
- Increasing attention on finding the factors affecting the sleep quality.

Data

- Data were collected from 101 CKD patients (65 male) in the Colombo South Teaching Hospital, Sri Lanka.
- The data set consists of 12 features (4 numerical and 8 categorical).
- Target variable is “Quality of sleep".

Pittsburgh Sleep Quality Index (PSQI)

- Sleep quality is measured using PSQI.
- This self-administered questionnaire assesses the quality of sleep during the previous month.
- It contains 19 self-rated questions yielding seven components:
  - subjective sleep quality
  - sleep latency
  - sleep efficiency
  - sleep duration
  - sleep disturbance
  - use of sleep medications
  - daytime dysfunction
- Each component is scored from 0 to 3, yielding a global PSQI score between 0 and 21.
- A global PSQI score$>$ 5 indicates that a person is a “poor sleeper", “good sleeper" otherwise.

Detailed Feature Information

Table: Summary Table - Counts (%) are shown unless otherwise specified.

Feature		Total (101)	Good Sleepers (32)	Poor Sleepers (69)	p-value
Age - mean ($\pm SD$)		60.38 ($\pm$10.92)	62.72 ($\pm$9.22)	59.29 ($\pm$11.53)	0.1429
Creatinine (μmol/L) - mean ($\pm SD$)		347.61 ($\pm$253.29)	245.22 ($\pm$87.96)	395.09 ($\pm$289.13)	**0.0051**
Duration of CKD (years) - mean ($\pm SD$)		26.04 ($\pm$25.74)	22.74 ($\pm$21.35)	27.57 ($\pm$27.55)	0.3833
haemoglobin (g/dL) - mean ($\pm SD$)		10.58 ($\pm$2.16)	11.76 ($\pm$1.67)	10.03 ($\pm$2.15)	**0.0001**
CKD stage	Stage 3 Stage 4 Stage 5	32 (31.6%) 31 (30.7%) 38 (37.6%)	9 (28.1%) 16 (50%) 7 (21.9%)	23 (33.3%) 15 (21.7%) 31 (44.9%)	**0.0113**
On haemodialysis	Yes No	21 (20.8%) 80 (79.2%)	0 (0%) 32 (100%)	21 (30.4%) 48 (69.6%)	**0.0012**
Sex	Female Male	36 (35.6%) 65 (64.4%)	14 (43.8%) 18 (56.2%)	22 (31.9%) 47 (68.1%)	0.3497
Employed	Yes No	23 (22.8%) 78 (77.2%)	11 (34.4%) 21 (65.6%)	12 (17.4%) 57 (82.6%)	0.1013
Heart failure	Yes No	4 (4%) 97 (96%)	2 (6.2%) 30 (93.8%)	2 (2.9%) 67 (97.1%)	0.7986
COPD	Yes No	6 (5.9%) 95 (94.1%)	1 (3.1%) 31 (10.3%)	5 (7.2%) 64 (92.8%)	0.7167
GORD	Yes No	12 (11.9%) 89 (88.1%)	4 (12.5) 28 (87.5%)	8 (11.6%) 61 (88.4%)	0.8418
Depression	Yes No	1 (1%) 100 (99%)	1 (3.1%) 31 (96.9%)	0 (0%) 69 (100%)	0.6923

Methodology

- Machine learning concepts were used.
- Decision trees used to identify the impact of each feature to predict sleep quality.
- Compare different classification models.
- Re-sampling techniques were used to avoid the imbalance issue.
- Compare results with four clinically relevant features determined by the physicians.They are creatinine, age, haemodialysis and haemoglobin.

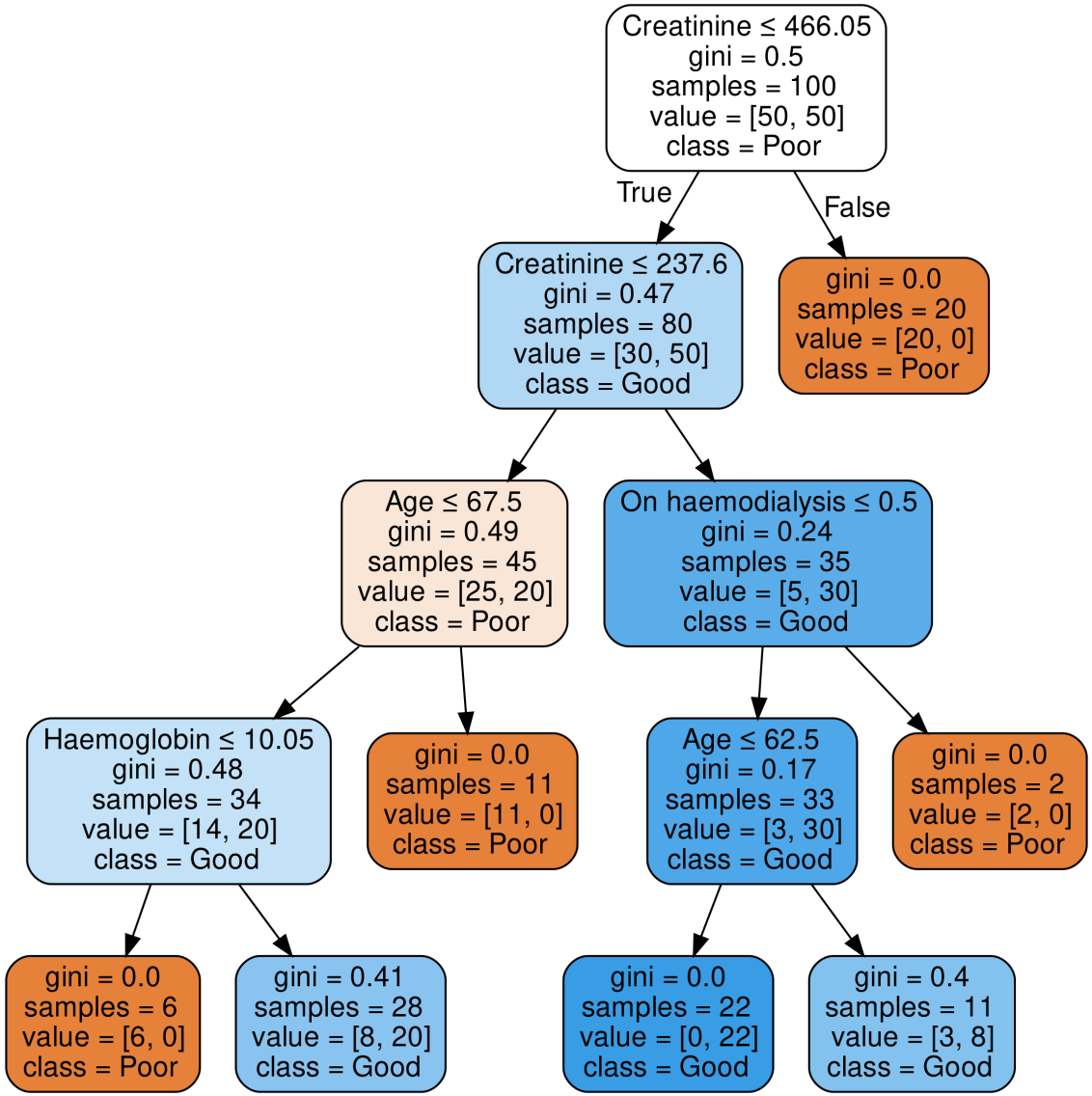

Figure: Decision tree fitted on the random over-sampled CKD data. The maximum allowed depth for the tree was set to 4.

Feature Importance

1. Identifying the most important features affecting sleep quality is a crucial aspect of this study.
2. Consider the feature importance, which was directly obtained from the classification model trained.
3. We fitted each model on 100 random over-sampled training sets and computed the average feature importance and rank features accordingly.
4. The mean of the ranks per each feature was also obtained.

Table: Rankings of average of feature importance for each classification methods

Feature	Decision_tree	RFC	Lgbm_c	Logit	SVM-linear	Mean	Rank of means
Haemoglobin	2	1	2	2	4	2.2	1
Creatinine	1	2	1	7	3	2.8	2
On Haemodialysis	9	6	10	1	1	5.4	3
Sex	5	8	5	3	8	5.8	4
Employed	6	7	7	4	7	6.2	5
CKD Stage	8	5	6	11	2	6.4	6.5
Duration of CKD	4	4	4	9	11	6.4	6.5
Age	3	3	3	12	12	6.6	8
Heart Failure	7	10	10	5	6	7.6	9
GORD	10	9	10	8	9	9.2	11
Depression	11	12	10	6	5	8.8	10
COPD	12	11	10	10	10	10.6	12

Discussion

- Decision tree classification methods are commonly used due to many reasons.
- Decision tree-based models, identified creatinine, hemoglobin, age, and duration of CKD as the most important four features.
- SVM-linear method strongly agreed with the statistical relationship results by identifying hemodialysis, CKD stage, creatinine, and hemoglobin as the most important four features.
- Haemoglobin, creatinine, haemodialysis, and sex are the most affected features identified by all the methods on average.

Discussion Cont.

- Traditional classification algorithms can perform poorly on imbalanced data sets and small sample size [1].
- Smaller sample sizes can reduce the power of the study.
- Manuscript is submitted based on “Modeling and Inference with Feature Importance for Assessing the Quality of Sleep among Chronic Kidney Disease Patients".

References

[1] Hu, Y., D. Guo, Z. Fan, C. Dong, Q. Huang, S. Xie, G. Liu, J. Tan, B. Li, and Q. Xie (2015, 01). An improved algorithm for imbalanced data and small sample size classification. *Journal of Data Analysis and Information Processing 03*, 27–33.

Acknowledgment

I would like to express my special thanks of gratitude to,

- To my supervisors Dr. Saman Muthukumarana & Dr. Mike Domaratzki for their excellent guidance
- To physicians Chamil Marasinghe & Varuni Tennakoon for providing data and the domain knowledge.
- To the department of Statistics and the staff for funding and resources
- To my family and friends for the continuous support

Thank You!

Contact: matharas@myumanitoba.ca