One Medical Passport: Predictive Obstructive Sleep Apnea Analysis

Anu-Ujin Gerelt-Od, Lee Kho, Paula Kiatkamolwong, Sumedha Rai
Center for Data Science, New York University
{ago265, ltk224, kk4158, sr5387}@nyu.edu

1 Introduction

The use of machine learning algorithms in the medical field has gained traction on account of its ability to provide more concrete and accurate results. As there is a sufficiently large amount of health data available, it is becoming common for hospitals to leverage these datasets to reassess current methodologies. For this project, we examined potential predictors of obstructive sleep apnea (OSA). This is a condition that affects the ability to breathe due to upper airway obstruction and often goes undiagnosed, creating the need for a comprehensive and precise pre-screening method.

In the scope of this project, we used various machine learning and deep learning algorithms including logistic regression, random forest, multi-layer perceptron, and K-modes clustering to create a pre-screening tool for OSA with data collected by the One Medical Passport team. We assessed the predictive power of 4 out of the 8 criteria included in STOP-BANG which is the most widely-used OSA screening questionnaire, and also explored other possible predictors that could be used to augment the existing questionnaire. Based on our results, we concluded that there are other factors that may be as strong or even stronger predictors of OSA than those included in STOP-BANG.

2 Related Work

Kreitinger et al. (2020) from UC San Diego evaluated the four most widely used OSA screening questionnaires: STOP-BANG, ESS, NO-OSAS, and No-Apnea. STOP-BANG and NO-OSAS have lower false-positive rates than the other two questionnaires, but suffer from false-negative screens. The team also concluded that sex shows no significant ability for predicting OSA.

Another study by researchers at the Keimyung University School of Medicine (Kim and Cho, 2019) analyzed the usefulness of the 8 criteria of STOP-BANG to establish the best assembly for OSA-screening methods in the Korean population. Their work focused on whether the questionnaire could exclude some of the factors in order to get the same predictive power as the original STOP-BANG. They concluded that STOP-BANG can be simplified to SOPBAG, dropping tiredness and neck circumference, while maintaining comparable screening performance.

Although both of these studies may be inconclusive on which of the STOP-BANG criteria play a significant role in screening OSA in a patient, there is one thing that they agree on: because polysomnography, considered to be the gold standard in diagnosing OSA, is expensive and has restricted availability, it is vital that more research be conducted in order to develop a more accurate pre-screening tool for OSA.

3 Problem Definition and Algorithm

3.1 Company Description

One Medical Passport (1MP) is a SaaS healthcare company for ambulatory surgery centers (ASC) founded in 2000. The company's platform connects over 800 facilities to physician offices and over

2.5 million patients each year, offering solutions for tasks such as booking, pre-admission, pre- and post-op patient engagement, and real-time patient tracking. Through its ASC partners, 1MP has accumulated a large amount of data on patients' information, medical and surgical histories, and medical survey responses. However, the company does not have a data science team and has not had the capabilities to perform significant data analysis. Our team was tasked with performing exploratory analysis on their patient database, specifically with regards to predictors of *obstructive sleep apnea*.

3.2 Obstructive Sleep Apnea

Obstructive sleep apnea (OSA) is a common medical condition in the United States and across the world. It can occur when the upper airway becomes blocked repeatedly during sleep, reducing or completely stopping airflow. This condition affects up to 9% of adults in the United States, although studies have shown that up to 80% of cases go undiagnosed (Kapur et al., 2002). An individual with OSA may feel tired during the day, and if left untreated, the condition may contribute to high blood pressure and other cardiovascular diseases, memory problems, weight gain and other long-term health risks.

In 2008, a screening questionnaire called **STOP-BANG** was created by researchers at the University of Toronto to help diagnose OSA (Chung et al., 2008). STOP-BANG includes the following 8 questions:

- Do you **snore** loudly?
- Do you often feel **tired** during the daytime?
- Has anyone **observed** you stop breathing or choking/gasping during your sleep?
- Do you have high blood **pressure**?
- Is your **BMI** more than 35 kg/m^2?
- Is your **age** older than 50?
- Is your **neck** size large (i.e. shirt collar 40 cm or larger)?
- Is your **gender** male?

Every question on the survey that a patient answers "Yes" to is equivalent to one point. Therefore, a higher score on the STOP-BANG questionnaire indicates a higher risk of having OSA.

3.3 Task

Since its inception, STOP-BANG has become the standard assessment tool used by medical professionals in diagnosing OSA. However, there are two potential issues with the questionnaire as it exists today:

- Because the survey assigns equal weights to each of the 8 predictors, it does not take into account *relative importance* among predictors, and therefore may produce simplistic and potentially misleading results.
- Only a small number of predictors are included in the STOP-BANG questionnaire, and thus it may be *too limited in scope* to effectively evaluate OSA risk levels.

In order to assess and develop solutions for these two possible shortcomings of STOP-BANG, we implemented three classification models (logistic regression, random forest, and multi-layer perceptron) for predicting the presence of OSA in a patient. We trained and tuned these models on 1MP's extensive patient medical record database and then extracted the feature importance from these models. In addition, we performed K-modes clustering to determine which predictors occur together in patients with OSA. In the following sections, we will detail our methodologies for data processing and modeling, our evaluation criteria, and the final results of our analyses.

3.4 Algorithms

3.4.1 Supervised Learning

Logistic Regression *Logistic regression* (or *logit*) performs classification by searching for the hyperplane in k-dimensional space that separates the two classes with minimal logistic loss. Since we are primarily using this model for feature importance (rather than for actual classification), we did not perform any regularization as not to introduce bias into the model. While logistic regression is a widely used model for extracting feature importance, the model assumes a linear relationship between the log odds of the dependent variable and the independent variables, which may be too strict an assumption in certain situations.

Random Forest Another useful and computationally efficient method for obtaining feature importance is *random forest*. Random forest is a bagging learning method that takes the average output of a collection of uncorrelated decision trees. Due to its non-linear nature, random forest can capture more complex dependencies between variables than linear algorithms such as logistic regression can. Random forests must be carefully tuned as to not over- or under-fit the data in training. For our model, we used a forest of 50 trees with max depths of 20.

Multi-layer Perceptron (MLP) *Multi-layer perceptron* (or *MLP*) is a feed-forward neural network model that can learn a non-linear function approximator for classification. MLP's functionalities are similar to that of logistic regression, however it has one or more non-linear hidden layers, which allow the decision boundary to have more flexibility. We performed a grid search to determine the parameters to use in our final model, which resulted in a TanH activation function, Adam optimizer, and 2 hidden layers consisting of 200 neurons each.

3.4.2 Unsupervised Learning

K-modes Within our larger feature importance analysis, we were tasked with exploring how the various medical conditions and physical attributes of a patient clustered together, specifically for patients who were clinically diagnosed with OSA. We performed clustering on all positive instances of OSA in our dataset to determine which patient profiles (in the undiagnosed or incorrectly self-diagnosed cohort) are at higher risk of having or developing OSA. We used *K-modes* clustering for this analysis, which is a variation of K-means, and is suitable with categorical features. Instead of distances and means, K-modes uses dissimilarity measures and modes to determine clusters (Saggio, 2016). We ran our model with 5 clusters using the Huang clustering algorithm (Huang, 1998).

4 Experimental Evaluation

4.1 Data

The data provided to us by 1MP consisted of approximately 21M records and 478 features (excluding unique patient and date of service identifiers) covering patient information and medical history. Majority of the features were binary features, however there were a few numerical features such as `dobyear`, `weight` and `heightfeet`. Each record corresponded to a date of service for a patient and was identified by a unique `q_id` code. Therefore, each unique patient (identified by their `patient_id`) had one or more corresponding records in the database.

Each of 1MP's partner ASC's collects information specific to the type of surgery they specialize in, therefore the majority of the 478 fields in the original dataset was missing (see Figure 1). In addition, the duplicate entries for `patient_id` presented high levels of redundancy in the dataset. For these reasons, we performed the following initial data processing:

- Dropped all records with `NaN` in the `sleepapnea` column (i.e. all unlabeled records)
- For each unique `patient_id`, dropped all records that did not correspond with the most recent date of service
- Dropped all columns with more than 20% missing data
- Randomly imputed remaining missing values assuming a 50/50 split between positive and negative instances

These steps resulted in a remaining 12M records and 56 features. We then removed some highly correlated features and converted `age` and `bmi` from numerical to categorical by ranges, resulting in a final dataset with 53 features. See Figure 5 for descriptions of fields as well as Table 4 for detailed age and BMI ranges (both located in the Appendix).

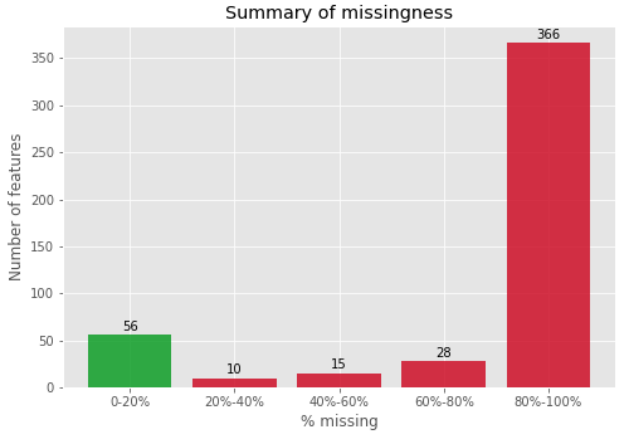

Figure 1: 422 out of the 478 original fields (∼88% of fields) had greater than 20% missing data.

4.2 Methodology

The processed dataset exhibited a large class imbalance, with a 90/10 split between negative and positive instances. Given the size of our dataset, we decided to perform random downsampling on our negative class to obtain a balanced training set (i.e. 50/50 split). Based on this decision, we randomly sampled 99% of our processed data and then performed the downsampling to obtain a balanced training set. The remaining 1% of the processed data was reserved as our test set. Note that we *did not* downsample the test dataset. These steps resulted in 2.4M instances in the train set and 119K instances in the test set.

Evaluation of Classifiers Since OSA has a very high undiagnosed rate and all of our data was self-reported by patients, this implies that a significant number (and potentially a majority) of negative instances in our dataset may be mislabeled as such. For this reason, we focused on capturing as many true positives (i.e. recall) as possible at the expense of high false positive rates (i.e. precision). Therefore, our two main evaluation metrics were *accuracy* and *recall*. The definitions of these metrics are defined as follows:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}} \tag{1}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3}$$

Feature Importance Metrics Each classification model implemented in our analysis determines feature importance using a different importance metric, which are described as follows:

Logistic regression uses the *odds ratio* to calculate feature importance. The regression coefficient of a variable corresponds to the change in log odds and its exponentiated form corresponds to the odds ratio. Used most often in case-control studies, the odds ratio computes the likelihood that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

Random forests calculate feature importance weights using *Gini importance*, which is defined as the total decrease in node impurity (weighted by the probability of reaching that node, which is approximated by the proportion of samples reaching that node) averaged over all trees (Lee, 2020). It should be noted that Gini importance does not capture the direction of the relationship between a feature and the target variable (i.e. does not indicate if they are positively or negatively correlated).

MLP itself does not have a way of outputting feature importance weights, thus we used the *permutation importance* method from ELI5, which measures the decrease in the score (accuracy, recall, or precision) when some features are unavailable. This is done by shuffling the values of certain features at each run, iterating the process for 100 epochs to account for noise in the data. The final feature importance output has a similar format to that of random forest, where each of the features are reported alongisde their weights.

Overall Feature Importance: Importance Points Since each model uses different feature importance metrics, we cannot simply take an average or sum of the weights across each model to get final feature weights. Therefore, we designed an **importance points** system to combine the results of each model into one aggregate ranking. Since each model exhibited similar classification performance, the models are weighted equally in the importance points system. Based on the importance ranking of a feature in a given model, we assigned between 0 and 3 points. Then, the points for each feature are summed across the three models, and the final standings determine whether we consider the feature a primary, secondary, or tertiary predictor. See Table 1 for details on the importance points system.

(a) For each model, we allocate points to features depending on the ranking of that feature.

Ranking	Importance Points
Rank 1-5	3
Rank 6-10	2
Rank 10-15	1
Rank 15+	0

(b) After summing the points for each feature across the models, we assign each feature a predictor level to indicate strength of predictive power.

Predictor Level	Total Imp. Points
Primary Predictor	7-9 points
Secondary Predictor	4-6 points
Tertiary Predictor	3 points

Table 1: Importance points system for aggregate ranking.

4.3 Results

Supervised learning

Our models achieved accuracies between 67.3% and 70.7% and recall scores between 73.5% and 78.6% on the test set. Table 2 contains the detailed performance results.

Table 3 shows the final feature importance ranking based on the importance points system outlined previously. For details on the feature importance weights derived from the three classification models, see Figure 6 in Appendix. Note that `homeo2use` is excluded from the final rankings due to its data leakage effect as Oxygen (O_2) administration is one of the treatments prescribed to patients with OSA (Mehta et al., 2013).

Model	Accuracy	Recall	Precision
Logit	0.70701	0.73520	0.21851
Random Forest	0.70237	0.74902	0.21787
MLP	0.67273	0.78566	0.20653

Table 2: Performance results of the classification models. All three models performed similarly in terms of accuracy, recall, and precision.

Predictor Level	Feature	Importance Points
Primary Predictor	BMI**	9
	Sex**	8
	Depression	8
	High blood pressure**	8
Secondary Predictor	Age**	5
	Asthma	4
	Abnormal heart rhythms	4
	COPD	4
	Arthritis	4
Tertiary Predictor	Acid reflux	3
	High cholesterol	3
	Lower back pain	3
	Neuropathy	3
	Hilatial hernia	3
	Panic	3

Table 3: Final ranking of top 15 OSA predictors. Features with ** indicate predictors included in STOP-BANG.

Unsupervised learning

We used K-modes on the subset of patients with a positive diagnosis of OSA, with the number of clusters set to 5. The resultant clusters had a very high count for *age* and *BMI* for categories 2 and 3 (corresponding to age of 40 years and above and BMI of 25 and higher; Figures 2 and 3). This was true across all the 5 clusters. Looking at the feature *sex* (Figure 4), our first cluster grouped together most of the female patients and we took the positive attributes within this cluster as representative of high-risk factors, specifically for females.

The remaining features were chosen based on their presence across all clusters. A positive feature that shows up in all the clusters with a high count was taken as a likely predictor, while a feature that only showed up in one cluster was dropped. Analyzing our clusters, we were able to conclude that *age* over 40 years, *high BMI* and *high blood pressure* would be likely predictors of the existence of OSA in female patients.

Additionally, *high cholesterol* levels in male patients might also be a high risk factor. Presence of *arthritis* in both females and males, and *chronic low back pain* and *heartburn* (i.e. frequent acid reflux) in males would be some additional features that could be subjectively looked at as potential predictors of OSA based on further medical evidence and specialist opinions. The results and graphs on these features identified as likely predictors and some examples of the features that were dropped are reported in Figure 7 in the the Appendix.

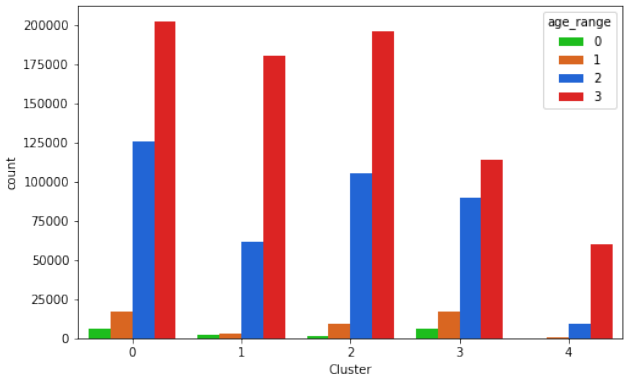

Figure 2: Count of age categories in clusters for diagnosed OSA patients

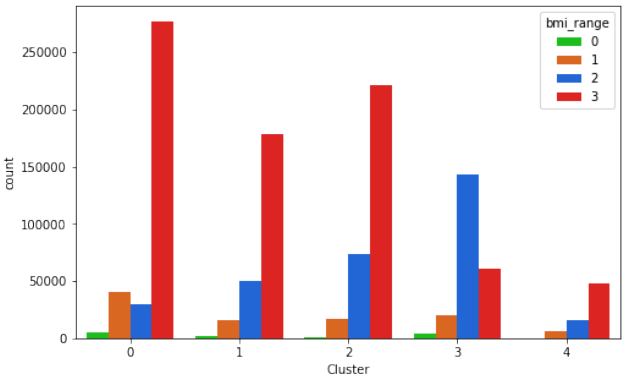

Figure 3: Count of BMI categories in clusters for diagnosed OSA patients

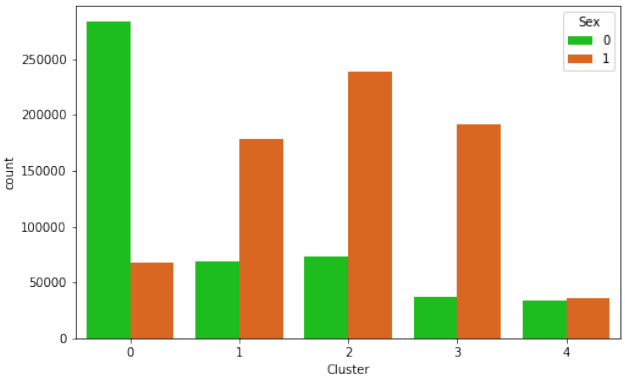

Figure 4: Sex (Female = 0, Male =1)

In addition to shedding light on some of the new potential indicators of OSA based on diagnosed patient profiles, our clustering analysis also validated the existing STOP-BANG criteria since all the features that are a part of the current survey (for which we had analyzable data) showed up as positive attributes for the diagnosed patients as part of our study.

4.4 Discussion

Three of the STOP-BANG criteria (*BMI, sex,* and *high blood pressure*) appear as the primary predictors from our analyses, while the fourth criterion (*age*) is a secondary predictor. This concurs with our hypothesis, as we expected features from STOP-BANG to appear as the top important features. On other hand, *depression* (which is not included as a criterion) is considered as an important feature across multiple models, thus it is ranked as a primary predictor. In addition to depression, multiple disorders that are not included in the STOP-BANG questionnaire (e.g. asthma, abnormal heart rhythms) also appear as secondary predictors.

We conducted further research from medical studies to see whether a conclusion could be drawn on these additional predictors. One explanation for our models having other disorders as important predictors is that these disorders share medical conditions that overlap with those of OSA. Here is a brief summary of our findings for our primary and secondary predictors:

- **Depression**: Recent studies have found that untreated OSA is associated with multiple psychiatric disorders such as depression and anxiety. Studies show that approximately half of people with OSA have depressive symptoms (Rezaeitalab et al., 2014; Shoib et al., 2017).
- **Asthma**: Asthma patients appear to have an increased risk for OSA than the general population, and there seems to exist a bidirectional relationship where each disorder adversely influences the other one (Dixit, 2018).

- **Abnormal heart rhythms**: Abnormal oxygen saturation level during sleep caused by OSA increases the risk of irregular heartbeats (Kendzerska et al., 2018).
- **Chronic obstructive pulmonary disease (COPD)**: A synergistic relationship between OSA and COPD has been observed in patients with both conditions and is often referred to as *overlap syndrome* (Mieczkowski and Ezzie, 2014).
- **Arthritis**: There is an increased risk of developing OSA in individuals with rheumatoid arthritis (RA). This is partly due to the fact that RA causes underdevelopment of the lower jaw, reduction of the size of the upper airway due to the degeneration of the temporomandibular joints, and narrowing of the spaces between cervical vertebrae that can cause compression on the brain stem and affect the severity of OSA (Shen et al., 2016).

5 Conclusions

Based on the results of our analyses, we conclude that the 4 STOP-BANG criteria included in our analysis are effective indicators for the presence of OSA. However, more research should be done with data pertaining to all 8 of the features in STOP-BANG to determine weights that indicate the strength of predictive ability to make the assessment more robust and accurate.

We also propose that additional features (e.g. depression, asthma, abnormal heart rhythms, etc.) should be considered as potential additions to STOP-BANG to create a more comprehensive assessment. Should these changes make sense from a medical point of view, and if there are no correlation or causation effects that we overlooked, adding these criteria could be helpful in diagnosing a wider range of patients.

One shortcoming of our analysis is the use of different features importance metrics for the three classification models, which does not allow us to directly compare feature importances across models. To rectify this issue, we could try using permutation importance (which is model agnostic) on all models to calculate feature importances. Using permutation importance for all models would allow us to take a simple average or sum of importances, and we would not have to use the importance points system. Thus, we would be able to compare cardinal rather than ordinal values and would retain more information about relative feature importance.

Another aspect of our analysis that we could revise is our data imputation method. For the sake of speed and simplicity, we randomly imputed missing values assuming an equal split between positive and negative instances. However, it might be beneficial to use a more sophisticated (albeit computationally expensive) imputation method such as multiple imputation, which takes into account both the distribution of the known values for each feature as well as the uncertainty of imputed values.

As for future work, we believe that conducting a time series analysis on the medical history of patients will give more insightful results. For patients that were not initially diagnosed with OSA, we can derive the factors that could have potentially led to the prognosis. It would also be a worthwhile exercise to obtain more accurate and complete data, meaning data that is collected from medical professionals rather than self-reported by patients, to perform our analyses on. This would greatly reduce the number of mislabeled instances and overall noise in the data.

6 Lessons Learned

One of the most challenging tasks for us was the preparing the dataset for modeling, without losing important information and getting tangible results. Real world data, and especially medical data, is very sparsely populated and it is difficult to determine which fields to drop without losing key features. After dropping features, we had to experiment with several data imputation techniques to get a dense dataset to run our models on. While there exist a multitude of ways to impute or interpolate data, since we were dealing with a medical dataset, we had to be careful to use methods that would cause minimal changes to the existing dataset in order to keep the original patient profiles intact.

We also found that it is important to have some basic domain-specific knowledge, as it helps to interpret and make qualitative sense of the results from the models. For instance, looking at the outputs in the light of additional medical research, we were able to segregate highly correlated features and drop those that contributed to data leakage.

7 Contributions

All team members contributed to the data cleaning, EDA, modeling, and report writing processes.

8 References

Frances Chung, Balaji Yegneswaran, Pu Liao, Sharon A Chung, Santhira Vairavanathan, Sazzadul Islam, Ali Khajehdehi, and Colin M Shapiro. 2008. Stop questionnaire: a tool to screen patients for obstructive sleep apnea. *Anesthesiology*.

Ramakant Dixit. 2018. Asthma and obstructive sleep apnea: More than an association! *Lung India: Official Organ of Indian Chest Society*, 35(3):191.

Zhexue Huang. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304.

Vishesh Kapur, Kingman P Strohl, Susan Redline, Conrad Iber, George O'connor, and Javier Nieto. 2002. Underdiagnosis of sleep apnea syndrome in us communities. *Sleep and Breathing*, 6(2):49–54.

Tetyana Kendzerska, Andrea S Gershon, Clare Atzema, Paul Dorian, Iqwal Mangat, Gillian Hawker, and Richard S Leung. 2018. Sleep apnea increases the risk of new hospitalized atrial fibrillation: a historical cohort study. *Chest*, 154(6):1330–1339.

Keun Tae Kim and Yong Won Cho. 2019. Real-world stopbang: how useful is stopbang for sleep clinics? *Sleep Breath*, 23:1219–1226.

Kimberly Y. Kreitinger, Macy M. S. Lui, Robert L. Owens, Christopher N. Schmickl, Eduardo Grunvald, Santiago Horgan, Janna R. Raphelson, and Atul Malhotra. 2020. Screening for obstructive sleep apnea in a diverse bariatric surgery population. *Obesity Symposium*, 28:2028–2034.

Ceshine Lee. 2020. Feature importance measures for tree models - part i.

Vanita Mehta, Tajender S Vasu, Barbara Phillips, and Frances Chung. 2013. Update on obstructive sleep apnea and its relation to copd. *Obstructive sleep apnea and oxygen therapy: a systematic review of the literature and meta-analysis.*

Brian Mieczkowski and Michael E Ezzie. 2014. Update on obstructive sleep apnea and its relation to copd. *International Journal of Chronic Obstructive Pulmonary Disease*, 9:349.

Fariborz Rezaeitalab, Fatemeh Moharrari, Soheila Saberi, Hadi Asadpour, and Fariba Rezaeetalab. 2014. The correlation of anxiety and depression with obstructive sleep apnea syndrome. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, 19(3):205.

Alessia Saggio. 2016. Into the world of clustering algorithms: k-means, k-modes and k-prototypes.

Te-Chun Shen, Liang-Wen Hang, Shinn-Jye Liang, Chien-Chung Huang, Cheng-Li Lin, Chih-Yen Tu, Te-Chun Hsia, Chuen-Ming Shih, Wu-Huei Hsu, and Fung-Chang Sung. 2016. Risk of obstructive sleep apnoea in patients with rheumatoid arthritis: a nationwide population-based retrospective cohort study. *BMJ open*, 6(11).

Sheikh Shoib, Javid A Malik, and Shariq Masoodi. 2017. Depression as a manifestation of obstructive sleep apnea. *Journal of neurosciences in rural practice*, 8(3):346.

9 Appendix

Field	Field Description/Question
abheart	Diagnosed Abnormal Heart Rhythm
acidreflux	Frequent acid reflux or heartburn
aids	HIV or AIDS
anemia	Anemia
anemia	Anemia, including sickle cell anemia
arthritis	Arthritis, juvenile rheumatoid or other
arthritis	Arthritis
asthma	Asthma
bipolar	Bipolar Disorder
bleeding	Bleeding or blood clotting disorders
cad	Coronary Artery Disease (CAD)
cancer	Cancer
cardiacother	Other cardiac (heart) conditions
chestpain	Heart related chest pain (angina)
cholesterol	High cholesterol or lipids
copd	Emphysema/COPD
depression	Depression
endoother	Other endocrine conditions
giother	Other GI conditions
heartattack	Heart attack
heartfail	Congestive heart failure (CHF)
heartvalve	Heart valve problems
hemother	Other hematologic conditions
hepatitis	Hepatitis
hiatushernia	Hiatal hernia of the stomach
highblood	High blood pressure
homeo2use	Home oxygen use
kidney	Kidney failure
liver	Liver failure or yellow jaundice
liverother	Other hepatic (liver) conditions
lowback	Chronic low back pain
migraine	Frequent migraines that require treatment
msra	MRSA (Methicillin-Resistant Staphylococcus Aureus)
musculo	Musculoskeletal problems
musother	Other musculoskeletal conditions
neuropathy	Neuropathy
neurother	Other neurologic conditions
panic	Panic/anxiety attacks
pneumonia	Pneumonia within the last 6 weeks
psyother	Other psychiatric conditions
psyother	Other psychiatric/behavioral conditions
pulother	Other pulmonary (lung) conditions
renalother	Other renal (kidney) conditions
schizo	Schizophrenia
seizures	Seizures
sleepapnea	Diagnosed with Sleep apnea
stroke	Stroke (including "ministrokes" or TIAs)
thinners	Recent use of "blood thinners"
thyroid	Thyroid disease
tuber	Tuberculosis (TB)
ulcers	Stomach ulcers
urinaryin	Urinary incontinence
vascular	Vascular disease (low blood flow from narrowed or blocked arteries)

Figure 5: Descriptions of fields that were kept in the final dataset after processing.

Age Range	Age Range Values
Range 0	0-20 years
Range 1	21-40 years
Range 2	41-60 years
Range 3	61+ years

BMI Range	BMI Range Values
Range 0 (Underweight)	<18.5
Range 1 (Normal)	18.5-24.9
Range 2 (Overweight)	25-29.9
Range 3 (Obese)	>30

Table 4: Breakdown of age and BMI ranges used for converting numerical features to categorical features.

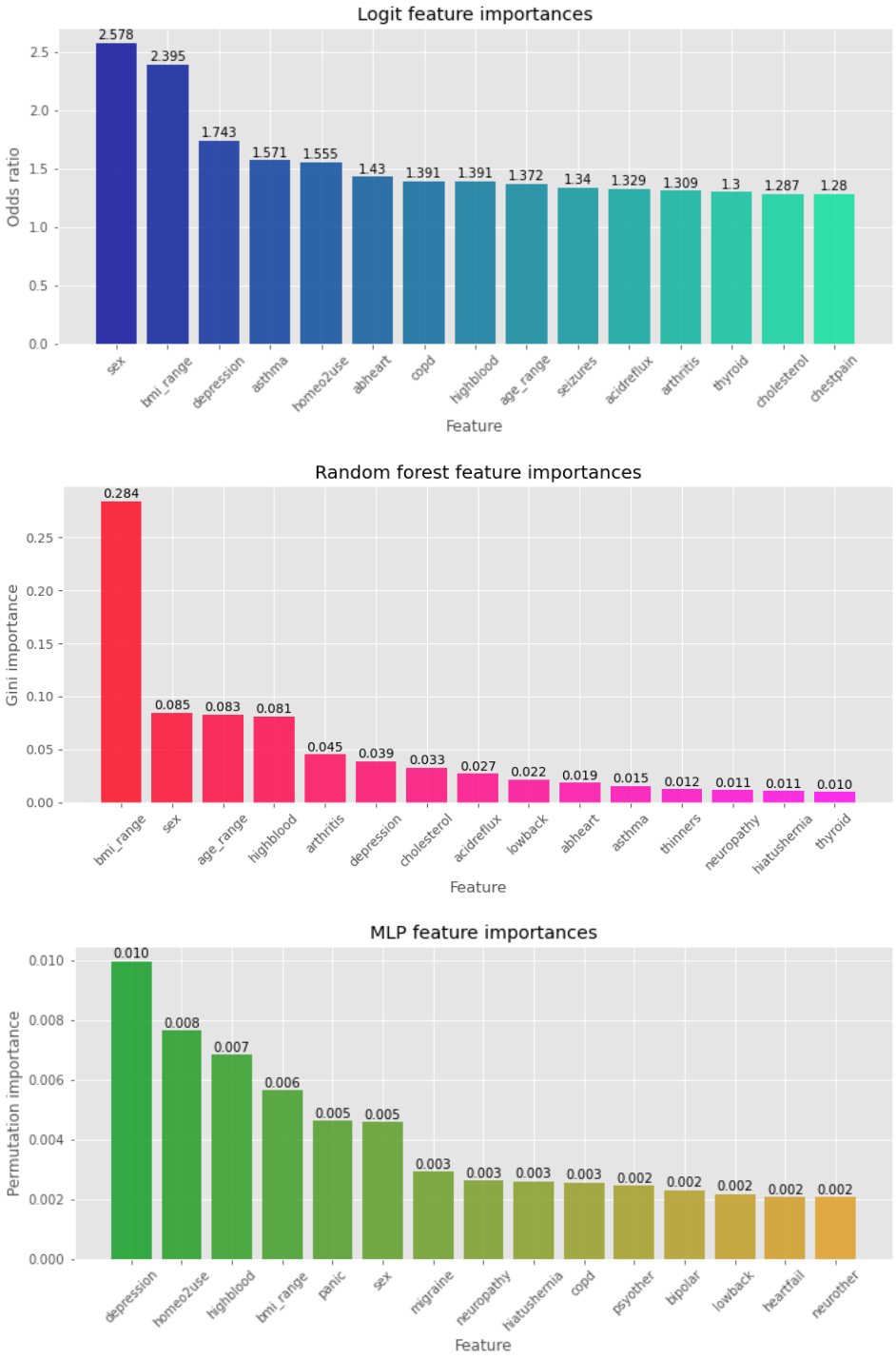

Figure 6: Top 15 features and their importance weights from logistic regression, random forest, and MLP classification models.

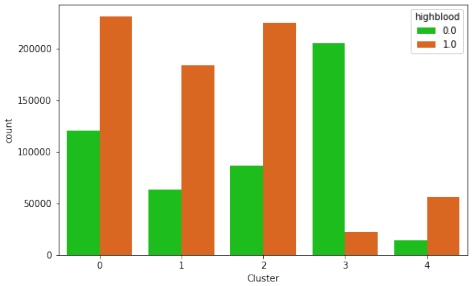

Figure 7 (a): High Blood Pressure

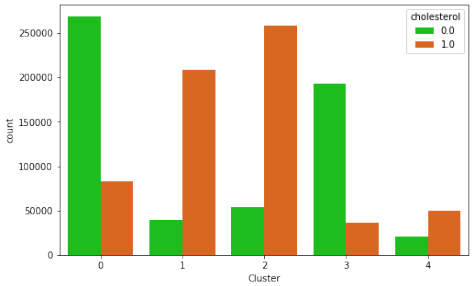

Figure 7 (b): High Cholesterol

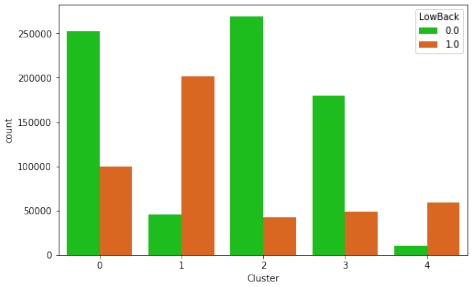

Figure 7 (c): Chronic Low Back Pain

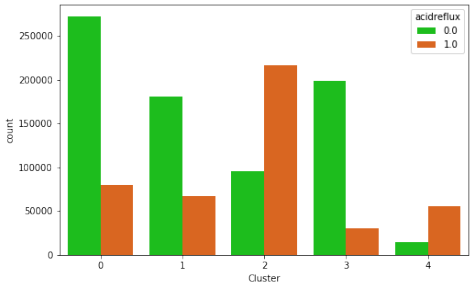

Figure 7 (d): Heartburn (frequent acid reflux)

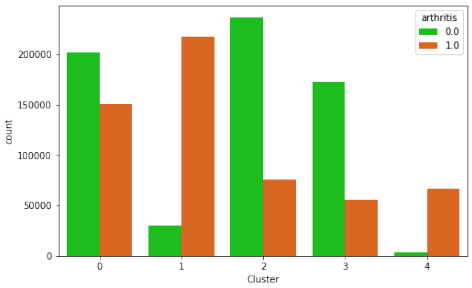

Figure 7 (e): Arthritis

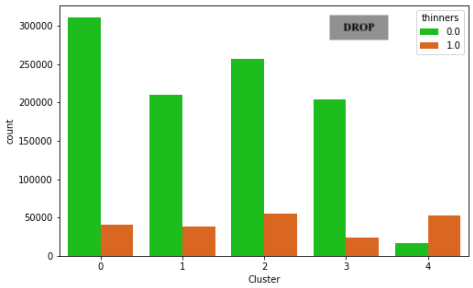

Figure 7 (f): Use of blood thinners

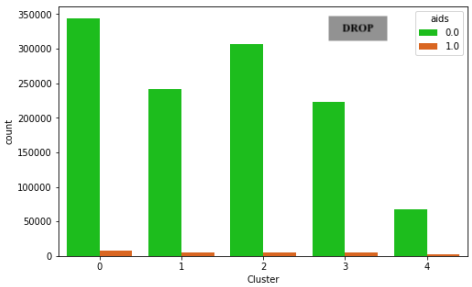

Figure 7 (g): AIDS

Figure 7: Results of the k-modes clustering: Features with counts of their categories across the 5 clusters