Analysis of health insurance from NFHS dataset

Deep Chordia
Computer Science
BITS Pilani, Hyderabad Campus
Hyderabad, India
f20202073@hyderabad.bits-pilani.ac.in

Moksh Papneja
Computer Science
BITS Pilani, Hyderabad Campus
Hyderabad, India
f20202074@hyderabad.bits-pilani.ac.in

Aaditya Mahesh Rathi
Computer Science
BITS Pilani, Hyderabad Campus
Hyderabad, India
f20202191@hyderabad.bits-pilani.ac.in

Sriram Balasubramanian
Computer Science
BITS Pilani, Hyderabad Campus
Hyderabad, India
f20200002@hyderabad.bits-pilani.ac.in

***Abstract*—This study aims to explore and analyze the NFHS dataset, and predicts the likelihood of a person opting to take health insurance based on the Household data collected across India. We then use data analysis techniques to identify patterns and obtain essential information from this data.**

***Keywords*—PCA, Apriori, DBSCAN, Naive Bayes, Decision Tree, t-SNE**

I. Introduction

The National Family Health Survey (NFHS) is a high-scale, multi-round survey conducted in a representative sample of households throughout India. This study delves into the problem statement of "Health Insurance Analysis" using data mining methods, to figure patterns and impressions in the data. It also tries to answer the question regarding "Who has health insurance and what is the lifestyle of this person?" The dataset used is the IAKR dataset collected through the years 2019-2021.

II. Preprocessing the data

A. Selecting columns required for the analysis

We analyze the raw data and use domain knowledge to prune the unnecessary attributes and the remaining valuable attributes like respondent occupation, disease history, body mass index etc were taken into account. The list of all the 34 columns used for the analysis can be found from the .MAP file provided and the colab link attached with this paper.

B. Elimination of missing values

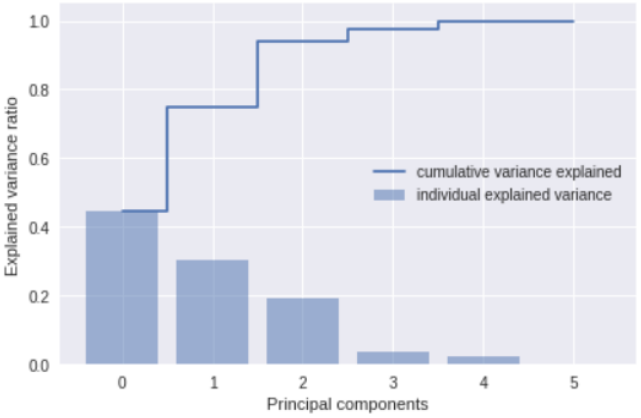
There are data instances where a respondent might not provide any answer to one or more questions or they are missing due to unforeseen circumstances. The .MAP files provide a correspondence for representing such missing or unknown values. The missing value used is different for each different attribute and these columns are individually processed for removal of these values. The missing data is dealt with by either dropping the columns or eliminating the rows (if the no. of missing values are considerably low).

C. Standardizing the data

The columns containing continuous attributes like body mass index, Rohrer's index etc., have different ranges of values, hence it becomes difficult to analyze these attributes. We use data transformation techniques to bring all the continuous values to same scale. We reduce these values by performing z-score normalization, which means we reduce the data values such that the mean of the column data is zero, and the standard deviation is one.

D. Dimensionality Reduction (PCA)

The dataset obtained after all the above preprocessing is high dimensional. We thus apply data reduction technique PCA on the continuous float type attributes. The project performs Principal Component Analysis (PCA) as a method for reducing our dimensions, with the variance of the dimensions to be obtained to be limited to 90%.



E. Data Aggregation

Since it was only required to know if a person had contracted a serious disease in the past, and not the specific disease that he/she contracted, the columns containing various diseases were aggregated into a single column. Thus we create a new feature named 'Disease' which specifies whether a person had or has any serious illness.

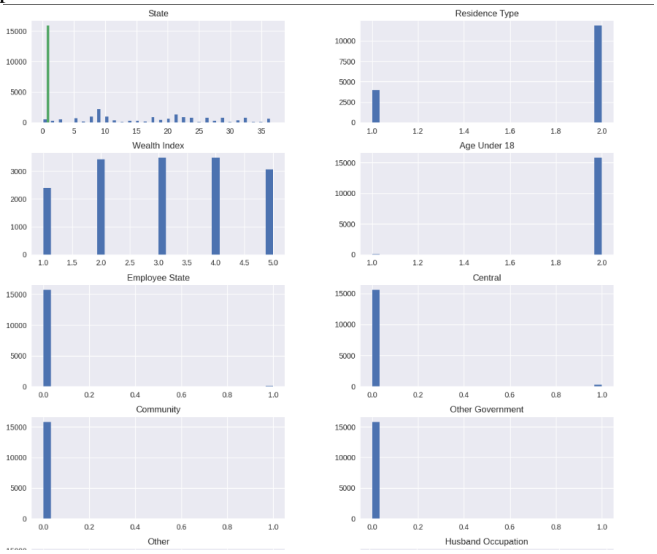
III. Data Visualization

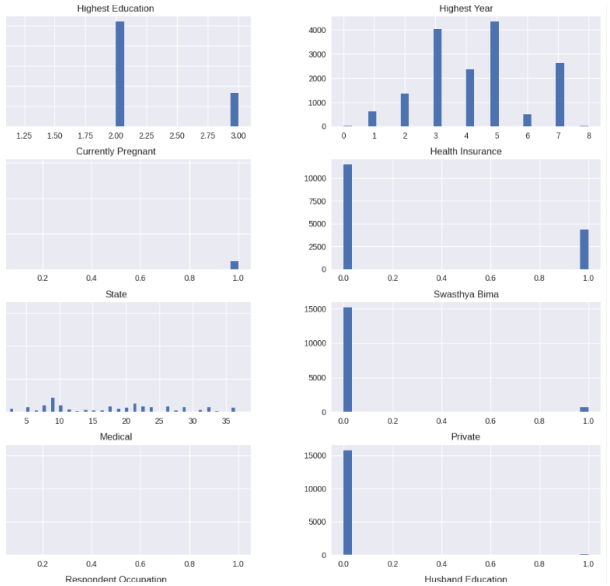
Data Visualization is an important step in data mining. since it helps us to visually find any relations among the data attributes and uncover any underlying pattern. It also helps us to visually remove the outliers present in the dataset, making it more robust.

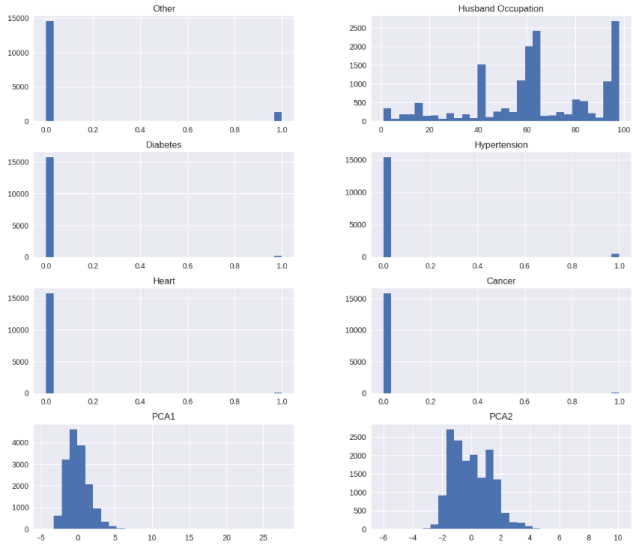
A. Bar Plots

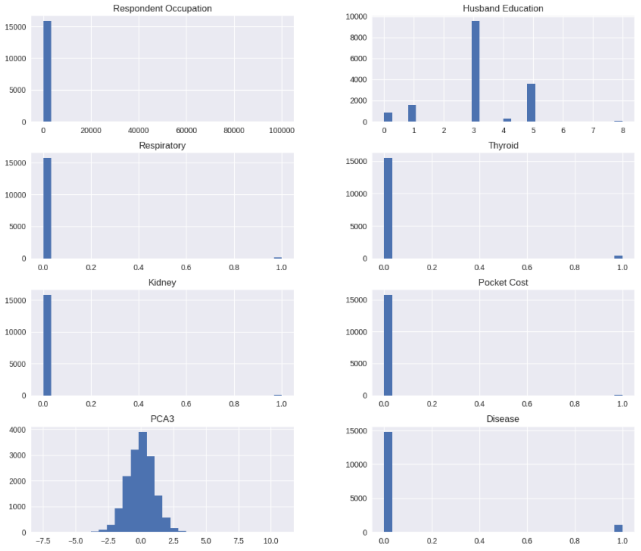
We plot histogram plots for each of the attributes present in the dataset. These plots show the distribution of data

points and values across various data attributes.



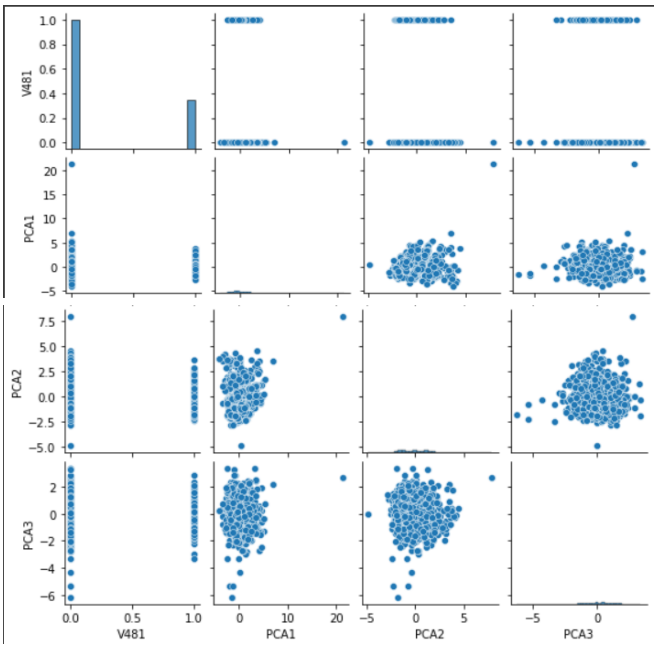






B. *Pair plots*

Pair plots are done only for PCA attributes considering these are the only attributes which are continuous and hence would show some logical variation. We also consider the Health Insurance binary attribute to show the spread of values concentrated only along '0' and '1 and thus providing an inaccurate method of visualizing the spread on entire dataset. Later in the data analysis part we use techniques like t-SNE to reduce the dimensionality of dataset to 2, thus allowing us to visualize the data in a 2 dimensional space.



We see that the pair plots of the PCA attributes show well formed clusters of the data points with some outliers. These outliers are removed later during data analysis.

C. *Correlation Matrix Heat Map*

The correlation between these PCA attributes is also plotted using a correlation heat map, to show the relation between the various principal components.

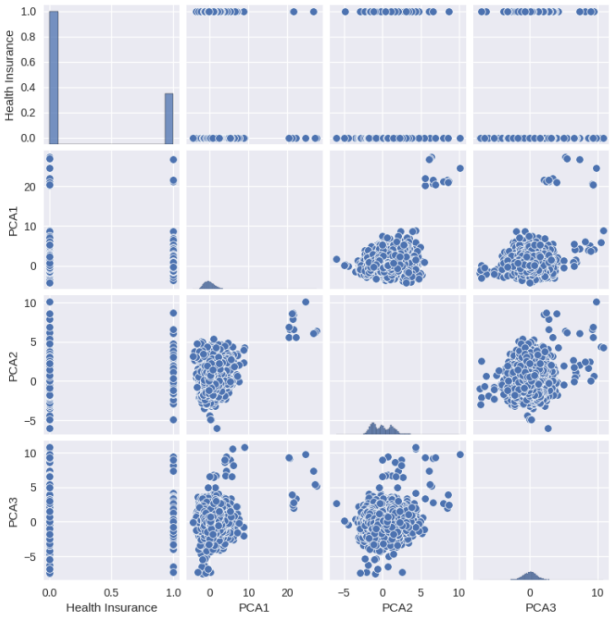
From this heatmap, it is ascertained that the attributes Body Mass Index and Rohrer's Index are very closely related, and follow similar trends. Hemoglobin levels and anemia levels are highly correlated too. It can also be seen that the attributes which are same have a complete correlation.

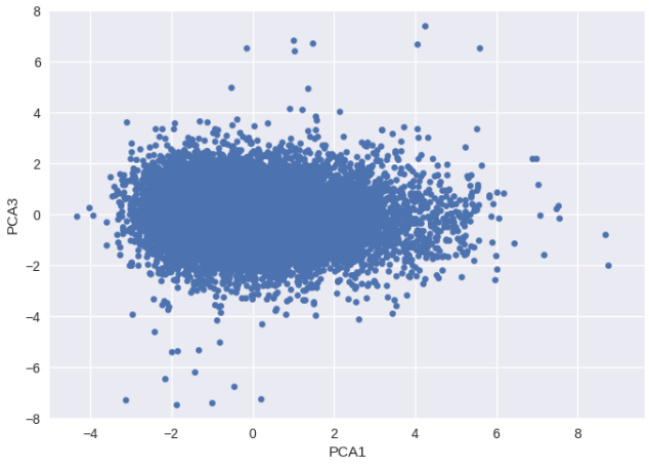
IV. Data Analysis

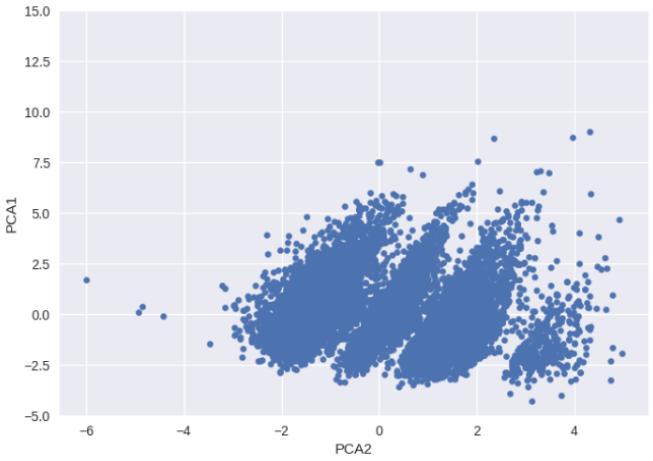
Data analysis is the automated extraction of useful information or patterns using the existing data. To gain useful insights and to predict outcomes for new data, we apply various data analysis techniques to our processed data. These techniques fall under two broad categories namely Descriptive Data Mining and Predictive Data Mining. Under Descriptive Data Mining we have methods like Association Rule Mining and Clustering, which are used to find the similarity between the given data attributes thus drawing any valuable insights. The Predictive Data Mining category consists of methods like Classification, Regression and Outlier Analysis. Now we use some of the above mentioned techniques to analyze our preprocessed data.
Note: We split the dataset into train split and test split before applying any data analysis techniques on the dataset.

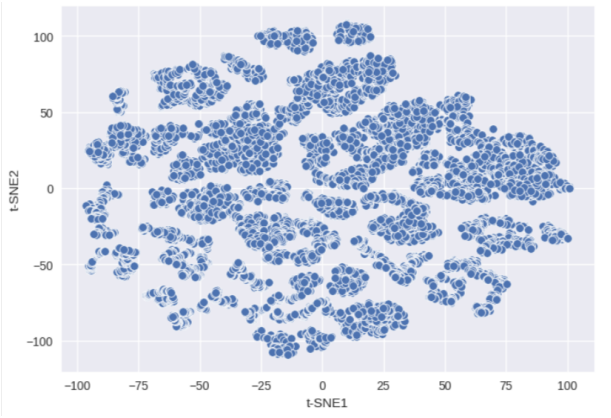
A. Clustering and Outlier Analysis

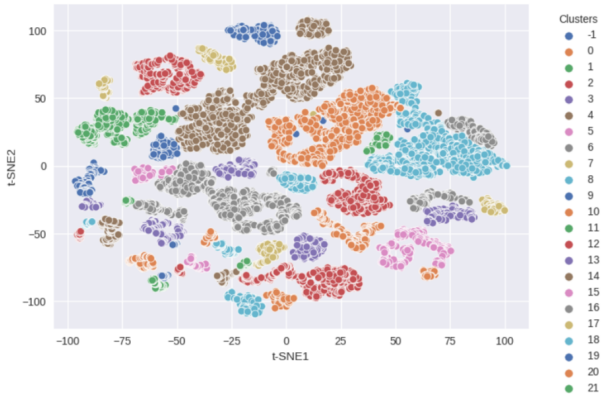
As stated earlier, we see that the pair plots from PCA cluster the data and show the outliers with respect to each principal component. First, outliers are removed from the pair plots of the PCA attributes because they do not belong to the same cluster formed by the attributes. This kind of outlier removal is done visually based on the data distribution and threshold values are chosen using logically. Then, we perform DBSCAN to cluster the data and then remove more outliers that do not align with the distribution of the data. We used t-SNE,which reduces the dimensions still preserving the pairwise similarity between the points. t-SNE calculates the pairwise similarity which are transformed into probabilities using Gaussian kernel. We reduced our 31 dimensional data to 2 dimensions following which we apply DBSCAN clustering algorithm on this reduced data to show better cluster formation on lower dimensional dataset.











B. Association Rule Mining using Apriori Algorithm

Association rule mining was done to generate interesting rules among the attributes of the data and how they relate to each other. The attributes of the dataset were considered as distinct items of a record, and were clubbed to create records for each entry of the dataset.

A genuine effort to produce limited no. of rules for each analysis is done, so as in order to have pertaining rules only for the best cases possible.Some rules are left out due to their low confidence,lift or support.The Apriori algorithm is implemented using

The project implements the Apriori algorithm using the 'mlxtend' library and also includes a from scratch implementation after analysis.

The library specification includes data to be fed in a binary format in form of a record data.Hence, the dataset is turned into this format,such that values to each column become the column themselves and the values to these columns are 0.0 and 1.0 that depict if that particular item is true for the particular record. This helps in forming a "record"-like dataset which can be fed for the Apriori algorithm. The columns are made descriptive for clear analysis of rules and frequent items.

The analysis takes 3 approaches here that represent the logic and aim of rules that are desired:

1) Insurance-Specific low occurrence Rules

Here, columns pertaining to `'Has Health Insurance'` become important attributes.

Our analysis determines the need to have a lower support count in order for evaluating rules pertaining to health insurance, considering only 0.5% of the data points and records commit to having health insurance.

Here, a lower support count and higher confidence is preferred to find rules that would determine the 'Health Insurance' parameter.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
frozenset({'V106_Higher', 'Husband/Partner-Working', 'V024_Punjab', 'V716_Respondent-Not working and did not work in last 12 months'})	frozenset({'V190_Richest', 'S456E_No Health Insurance'})	0.006372641807054073	0.19136854060193073	0.0061202599533093357	0.9603960396039604	5.018568028909716

Support Count taken>=0.004 and filtered sets with health insurance as part.

```
Rules_Insurance=filtered_rules[ (rules_mlxtend['lift'] >= 1) & (rules_mlxtend['confidence'] >= 0.9) ]

Rules_Insurance
```

```
<ipython-input-77-6621b8ef944f>:1: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  Rules_Insurance=filtered_rules[ (rules_mlxtend['lift'] >= 1) & (rules_mlxtend['confidence'] >= 0.9) ]
```

index	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
2072	frozenset({'S456E_Has Health Insurance'})	frozenset({'Husband/Partner-Working'})	0.005363114392075209 5	0.9933749763362012	0.005363114392075209 5	1.0	1.006666207317073
23068	frozenset({'V716_Respondent-Not working and did not work in last 12 months', 'S456E_Has Health Insurance'})	frozenset({'Husband/Partner-Working'})	0.004038109659915452	0.9933749763362012	0.004038109659915452	1.0	1.006666207317073

Confidence taken>=0.9 for rules along with lift>=0.9: give us rules with 1 confidence and >1 lift.

```
value='S456E_Has Health Insurance'
filtered_itemsets = frequent_itemsets[frequent_itemsets['itemsets'].apply(lambda x: value in x)]
filtered_itemsets
```

index	support	itemsets
109	0.005363114392075209 5	frozenset({'S456E_Has Health Insurance'})
1089	0.004038109659915452	frozenset({'S456E_Has Health Insurance', 'V716_Respondent-Not working and did not work in last 12 months'})
1149	0.005363114392075209 5	frozenset({'S456E_Has Health Insurance', 'Husband/Partner-Working'})
4651	0.004038109659915452	frozenset({'S456E_Has Health Insurance', 'Husband/Partner-Working', 'V716_Respondent-Not working and did not work in last 12 months'})

These rules infer that for all points (confidence=1) in the data, if they have health insurance their partner/husband must be working.

Similarly, if the respondent doesn't work in that case it is definitive that if he/she has health insurance, their partner must be working.

2) Insurance-Non-Specific low-occurrence Rules

There also may exist rules for certain values that may occur less due to data biasness. Hence, keeping support count low and producing rules that are not specifically filtered based on insurance , we can also find other rules that may exist in our data.

Considering the high sparsity of the data, a huge value of such rules may exist , however, to demonstrate the analysis provides one such rule that is specific for that support,list and confidence values.

Support >=0.004 , Lift>=5 and Confidence >=0.95
We get a rule as shown:

The rule, here indicates with confidence >=0.95 and hence highly reliable , that if the respondent has studied until higher education,their partner is working, they stay in punjab but however the respondent does not work then it is very highly likely that the family belongs the richest group of their state but does not claim health insurance.

3) High Occurrence Rules

Now, the analysis aims to find general rules that may exist and also have higher occurrences. These are rules with higher support counts and considerably high confidence as well.

Support count>=0.04 and Confidence >=0.72

```
RULES=rules_all[ (rules_all['confidence'] >= 0.72) ]
RULES
```

index	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
14	frozenset({'V106_Higher', 'V190_Richest', 'V025_Urban'})	frozenset({'V729_Husband/Partner-Higher'})	0.05621805792163543 6	0.22638652280901003	0.04151681494100574 5	0.7384960718294052	3.262102631667849
15	frozenset({'V106_Higher', 'V025_Urban', 'V729_Husband/Partner-Higher'})	frozenset({'V190_Richest'})	0.05684901255599722 5	0.19294592718783518	0.04151681494100574 5	0.7302996870366626	3.784996541158759
16	frozenset({'V190_Richest', 'V025_Urban', 'V729_Husband/Partner-Higher'})	frozenset({'V106_Higher'})	0.05514543504322039	0.2093507476812417	0.04151681494100574 5	0.7528604118993135	3.5961677722098315
70	frozenset({'V106_Higher', 'V025_Urban', 'V729_Husband/Partner-Higher', 'S456E_No Health Insurance'})	frozenset({'V190_Richest'})	0.05596567606789072	0.19294592718783518	0.04069657391633542	0.7271702367531003	3.768777332341363
71	frozenset({'V190_Richest', 'V025_Urban', 'V729_Husband/Partner-Higher', 'S456E_No Health Insurance'})	frozenset({'V106_Higher'})	0.05432519401855006 5	0.2093507476812417	0.04069657391633542	0.7491289198606272	3.5783436560792894
72	frozenset({'V106_Higher', 'V190_Richest', 'V025_Urban', 'S456E_No Health Insurance'})	frozenset({'V729_Husband/Partner-Higher'})	0.05539781689696511	0.22638652280901003	0.04069657391633542	0.734624145785877	3.244999466711361
74	frozenset({'V190_Richest', 'V025_Urban', 'V729_Husband/Partner-Higher'})	frozenset({'V106_Higher', 'S456E_No Health Insurance'})	0.05514543504322039	0.2075840747050287	0.04069657391633542	0.7379862700228833	3.5551198764719385
76	frozenset({'V106_Higher', 'V190_Richest', 'V025_Urban'})	frozenset({'V729_Husband/Partner-Higher', 'S456E_No Health	0.05621805792163543 6	0.22455675436936084	0.04069657391633542	0.7239057239057239	3.223709417865088

There rules infer,

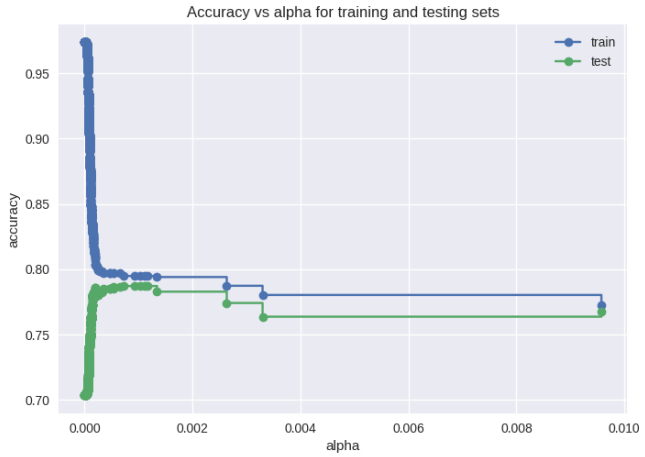
- If the respondent has attained higher education and belongs to the richest class of citizens and live in an urban dwelling then they are likely to have a partner that has also attained higher education.
- If both respondent and partner have attained higher education and are in urban areas they are very likely to belong to the richest class. Similarly, vice-versa.
- The attainment of health insurance doesn't matter in a person belonging to the richest class or having higher education.

All the 3 cases, give us rules pertaining to our project aim, as well as the in general rules that are followed over the whole dataset.

Inferences drawn per each case, shows how association rule mining using Apriori helps in data analysis.

C. *Decision Tree Classification*

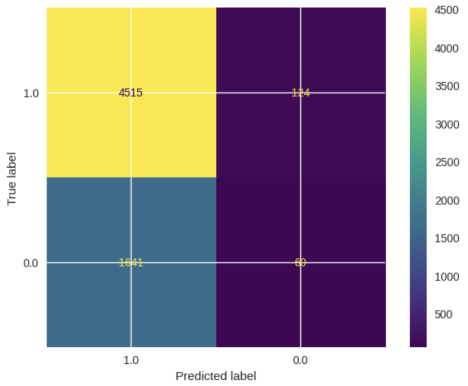
Since most of the data is categorical in nature (both binary and mutli class), and the final outcome to be ascertained i.e., if a person is eligible to obtain health insurance based on his data is binary in nature, classification techniques can be employed. One such classification technique is using Decision Tree Classifier. In this classification technique the data is divided into smaller subparts by taking a decision at each node based on the value of an evaluation function like Gini Index. The goal of this evaluation function is to find the attribute that results in the purest subgroup when used to split the data. When we trained our model on all the attributes our accuracy of the model was coming to just 62%. But after removing respondent's occupation and husband's occupation from the decision tree and state we increased accuracy of our model by 7%. The accuracy obtained from this classification model was 69%. A plot that shows the variation of alpha values and the accuracy of the decision tree classifier with pruning is shown below. Pruning was done to avoid overfitting and improve generalization.



D. *Naive Bayes Classification*

Naive Bayes is a classification technique which using the assumption that each of the data attributes are independent (and thus the name 'Naive'). The newly formed continuous PCA attributes were removed, so as to run a Naive Bayes classification model on discrete attributes in an attempt to get a model with possibly improved accuracy over the decision tree classifier. The dataset was split so as to have 80% of the data for training the classification model, and 20% for testing the accuracy of the model. Upon performing Naive Bayes classification, the created model showed an accuracy of 70.75% on the testing data split.

The confusion matrix for our prediction is as follows:



The 1 in the correlation matrix represents the person has insurance while 0 represents the person does not have insurance.

```
map = {
    'Health Insurance':'V481',
    'Respondent Occupation':'V716',
    'Husband Occupation':'V704',
    'Pocket Cost':'S456E',
    'BMI':'V445',
    'Rohrer Index':'V446',
    'Respondent Weight':'V437',
    'Respondent Height':'V438',
    'Age Under 18':'V452A',
    'Haemoglobin':'V453',
    'Currently Pregnant':'V454',
    'Anemia Level':'V457',
    'State':'V024',
    'Husband Education':'V729',
    'Highest Education':'V106',
    'Highest Year':'V107',
    'Diabetes':'S728A',
    'Hypertension':'S728B',
    'Respiratory':'S728C',
    'Thyroid':'S728D',
    'Heart':'S728E',
    'Cancer':'S728F',
    'Kidney':'S728G',
    'Residence Type':'V025',
    'Wealth Index':'V190',
    'Employee State':'V481A',
    'Central':'V481B',
    'State':'V481C',
    'Swasthya Bima':'V481D',
    'Community':'V481E',
    'Other Government':'V481F',
    'Medical':'V481G',
    'Private':'V481H',
    'Other':'V481X'
}
```

Conclusion

The project aims to produce various algorithms to be run on a real-life dataset like the NFHS data in order to predict if a person is in possession of and using their Health insurance given his/her various features as well as information about their surroundings.

We realize that we can predict this with a decent probability as well as formulate relations between insurance and other features.

Overall, we look at the fact that techniques like naive-bayes and decision trees can predict if a person has insurance with a 70% accuracy.

In order to gain more insight into the data, we also look at the rules that can be mined using the dataset, which provides relations between education and economic status in relation to insurance.

Appendix

The link to the code represented as a python notebook (.ipynb) on Google Colab:

https://colab.research.google.com/drive/1kZBiPwhy98u_AODahm5NTr9hJzXuY9h

https://docs.google.com/spreadsheets/d/1wHry8DM6HfgP_ocNix5nPbkn1DR52GqL_RVSEOSK-zo/edit#gid=1431401673