An Analysis of 'Risk prediction in life insurance industry using supervised learning algorithms'

Nimalan Subramanian

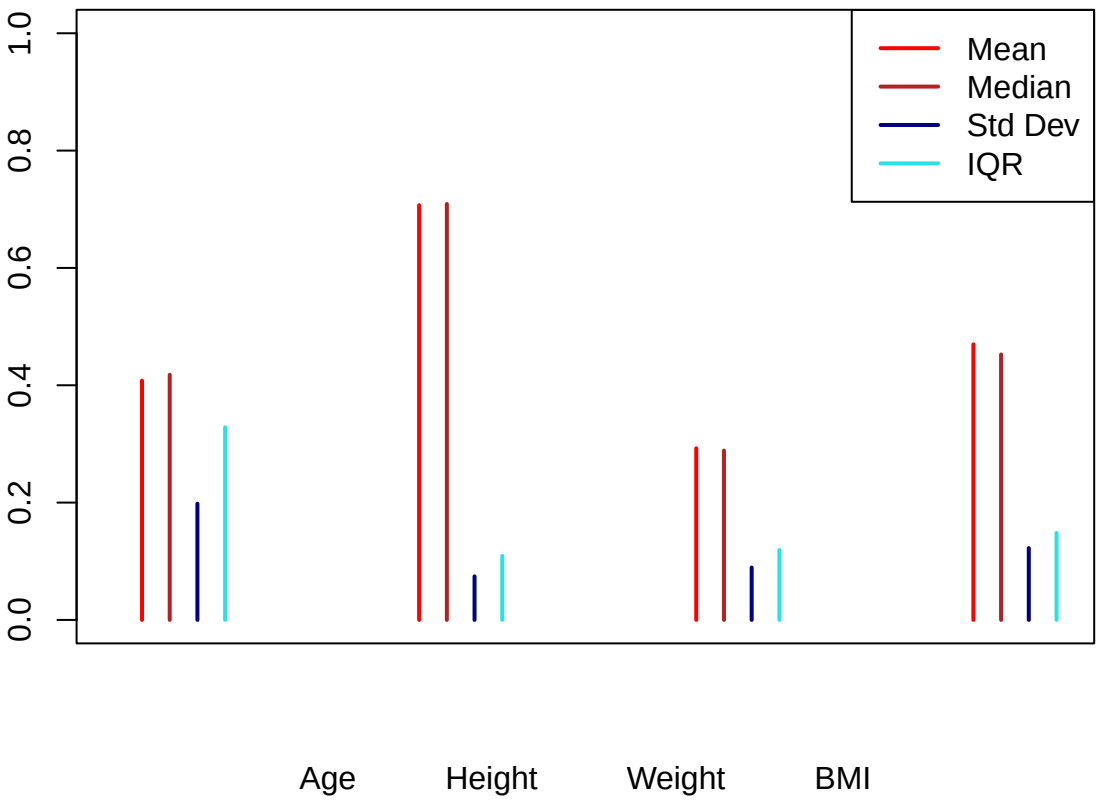
May 03, 2024

Introduction

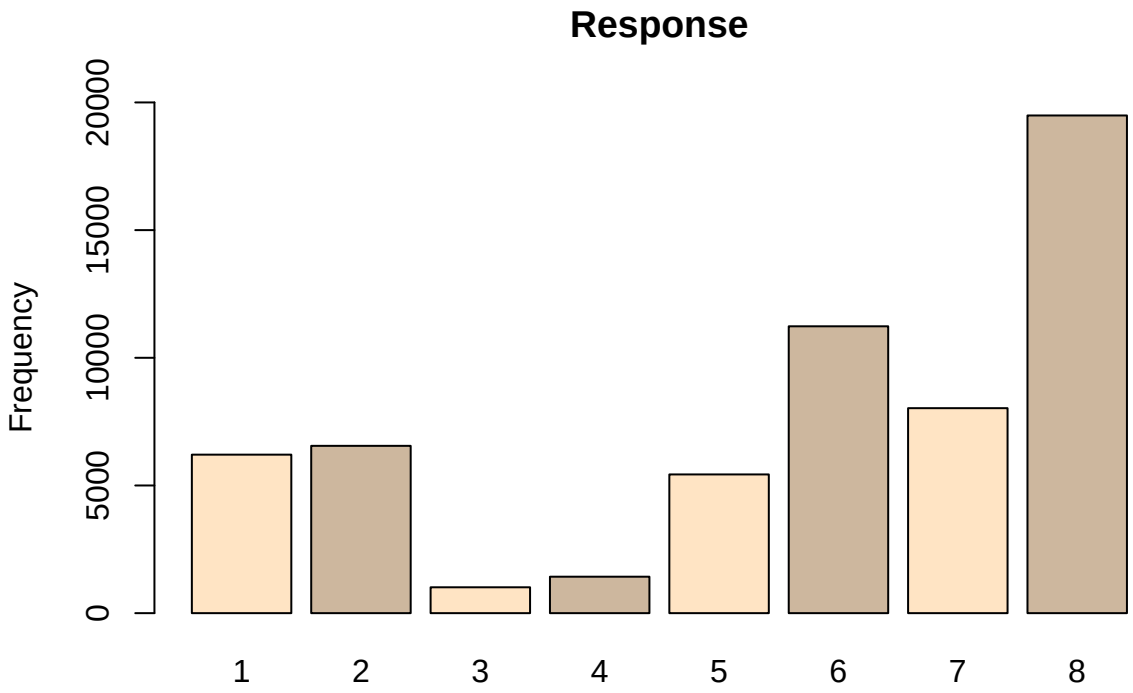
In the field of life insurance, risk assessment is a crucial component in classifying applicants. The underwriting process is used to make decisions on applications and price policies. Due to the rising number of applicants for life insurance, many insurance companies seek a faster, automated process to classify applicants and make decisions. Through the use of supervised learning algorithms, automating such a process has been proven to be both possible and effective. This is the main purpose of Noorhannah Boodhun and Manoj Jayabalan's "Risk Prediction in Life Insurance Industry Using Supervised Learning Algorithms. To determine whether the results of this initial analysis is valid, this project aims to recreate the main feature tools being modeled, the Principal Components Analysis (PCA) and the COrrealtion-Based Feature Selection (CFS). However, the utilization of such algorithms and the relevant data has various concerns, particularly ethical considerations. This stems from a sense of privacy in the data and whether the process itself is fair. As such, the breakdown for how these algorithms are made and utilized must be studied to determine whether they are morally sound to implement.

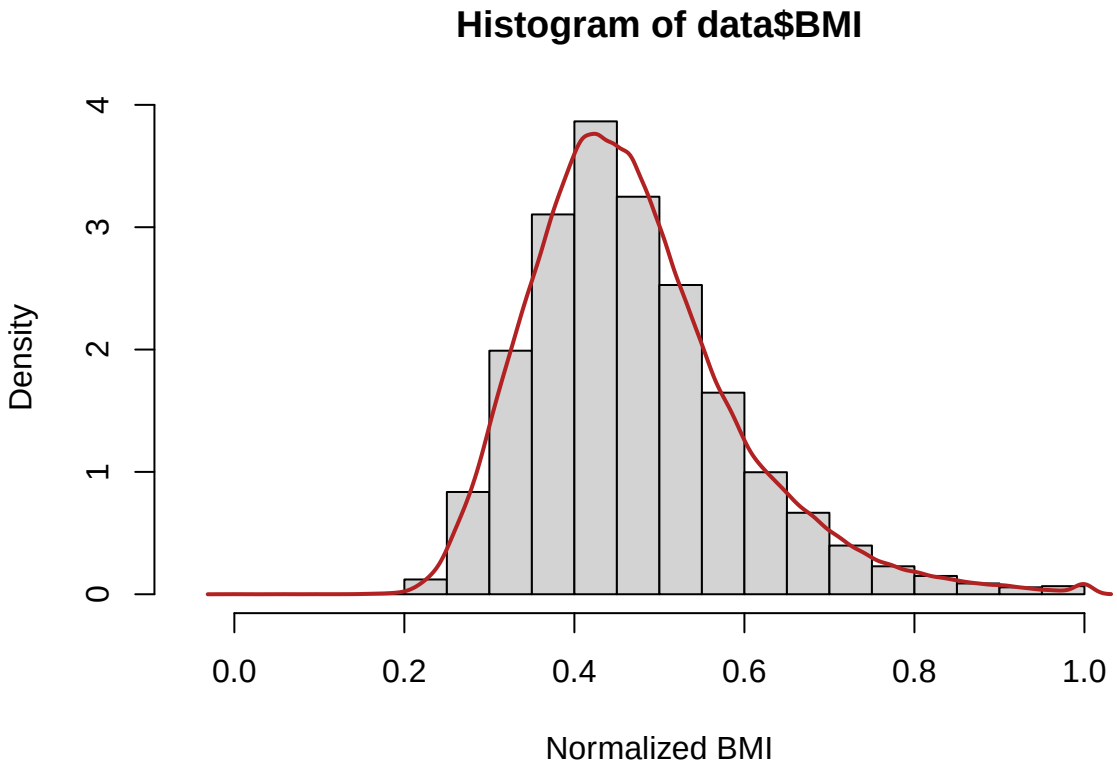
Analysis of Methods

In order to test the validity of one of the methods used in Boodhun and Jayabalan's assessment, the results needed to be verified. Too do this, the same data used in the original analysis was obtained publicaly through the Prudential Life Insurance Assessment Data on Kaggle. Before a method to verify results were simulated, initial exploratory data analysis was done to get an understanding of the data. For this, an initial summary was done, with general information being provided below:



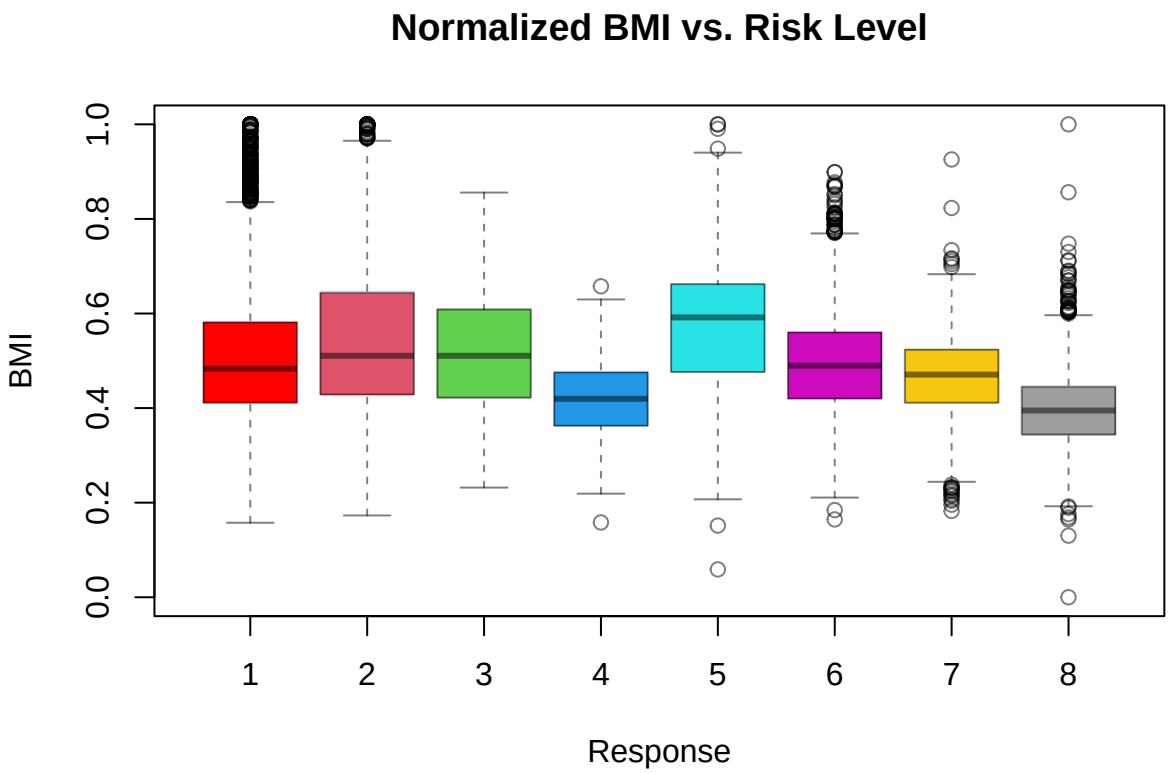
To gauge a better understanding of the data, an initial response variable was created to relate to specific vatiables for analysis. In this case, the sample BMI variable was normalized before undergoing various levels of analysis. First, a univariate analysis of the variable was done through the construction of a histogram:


Response
Frequency
0
5000
10000
15000
20000
1
2
3
4
5
6
7
8



Based on the histogram, the majority volume of life insurance applicants fall in the range between 0.2 and 0.8 for their BMI value (normalized).

Then, a bivariate analysis was done on BMI through a boxplot that looks at its interaction with the risk level response variable, following training of response data:



The boxplot data provides far more insight into the interaction between normalized BMI and the the risk level of applicants.

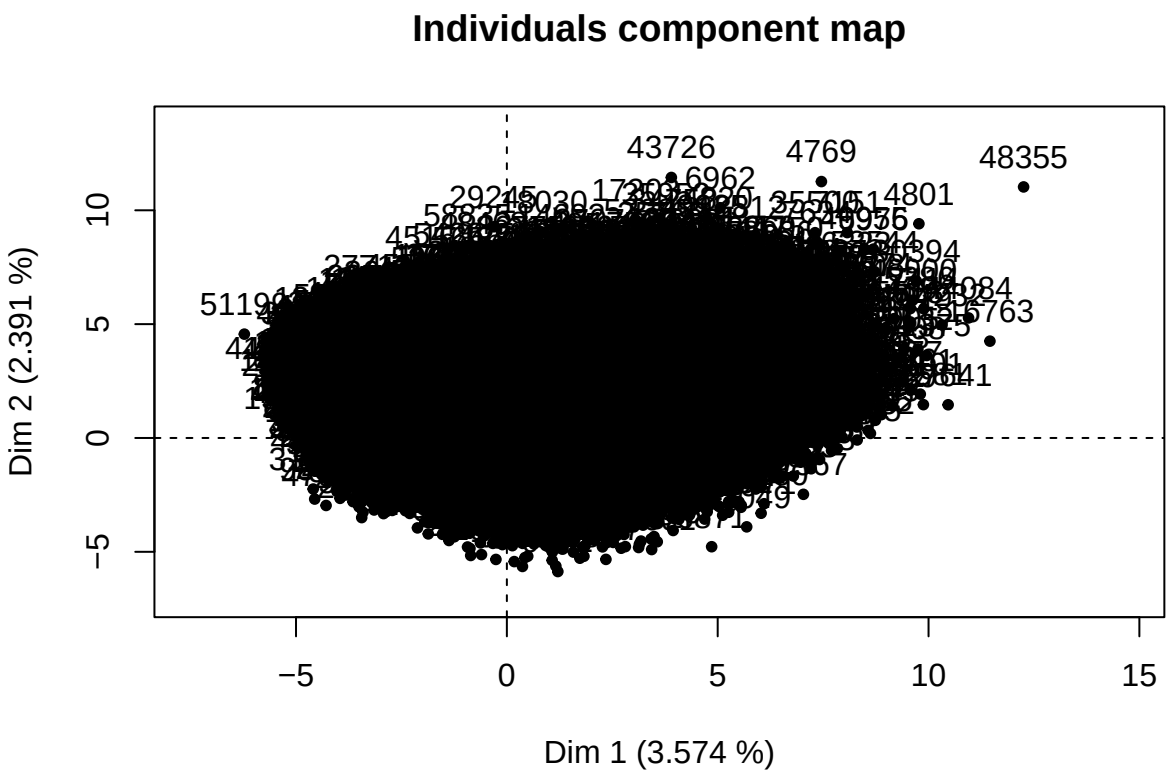
Following the EDA, testing the reproducibility of the PCA and CFS began. Prior to this, data was tested for Missing Completely At Random through Little's Test, with the results shown below:

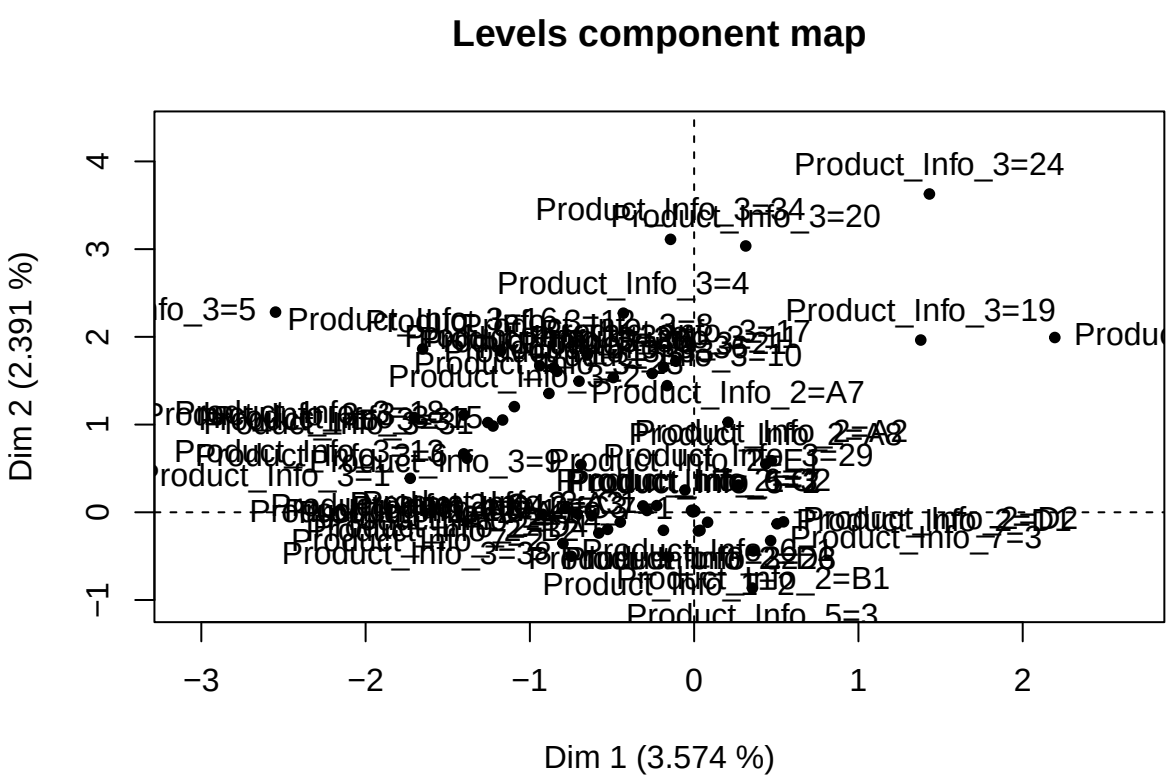
```
## Warning: package 'remotes' was built under R version 4.3.3
## Warning in mysort(data): NAs introduced by coercion to integer range
## Warning in mysort(data): NAs introduced by coercion to integer range
## this could take a while
## [1] 0
## Warning in mysort(data): NAs introduced by coercion to integer range
## Warning in mysort(data): NAs introduced by coercion to integer range
## this could take a while
## [1] 1
## this could take a while
## [1] 0
```

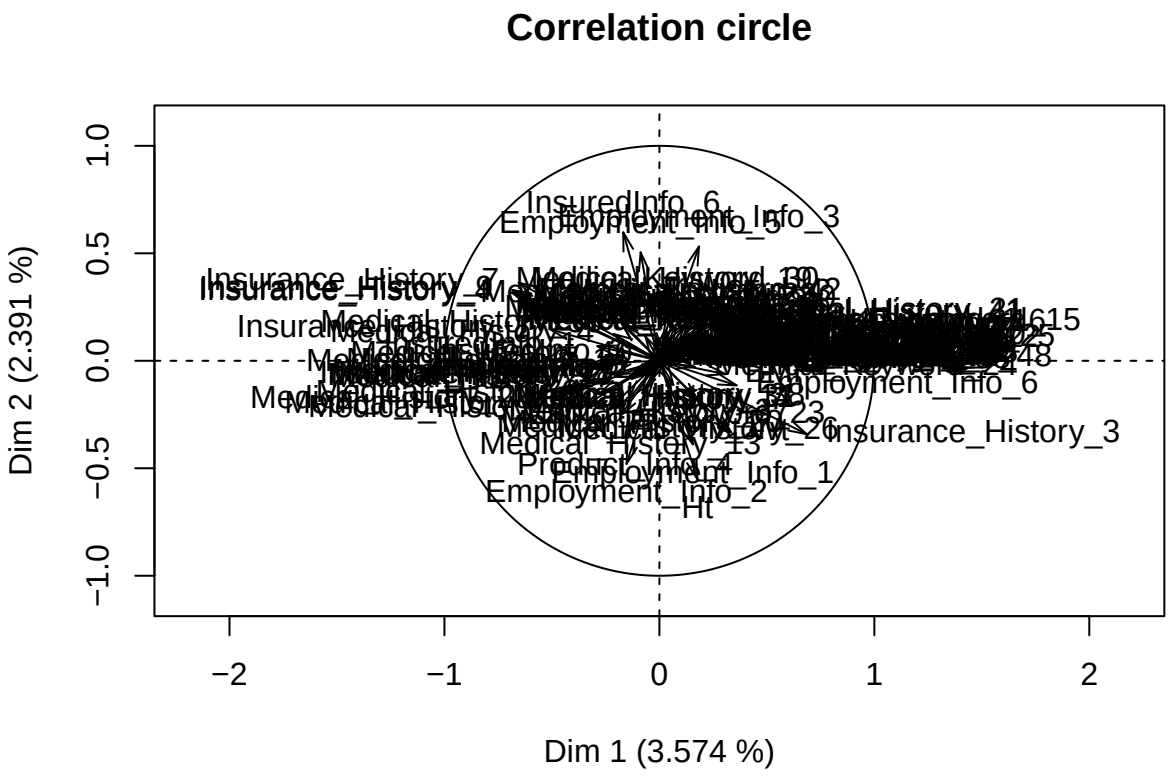
From this, a significance value of 0.000 is show, determining that missing data was not completely at random, as done in the original article. Knowing this, I moved onto cleaning the data with multiple imputations to replace missing values. This was done using the MICE (Multivariate Imputation via Chained Equations), as done in the original, keeping in mind that I was assuming missing data to be Missing at Random.

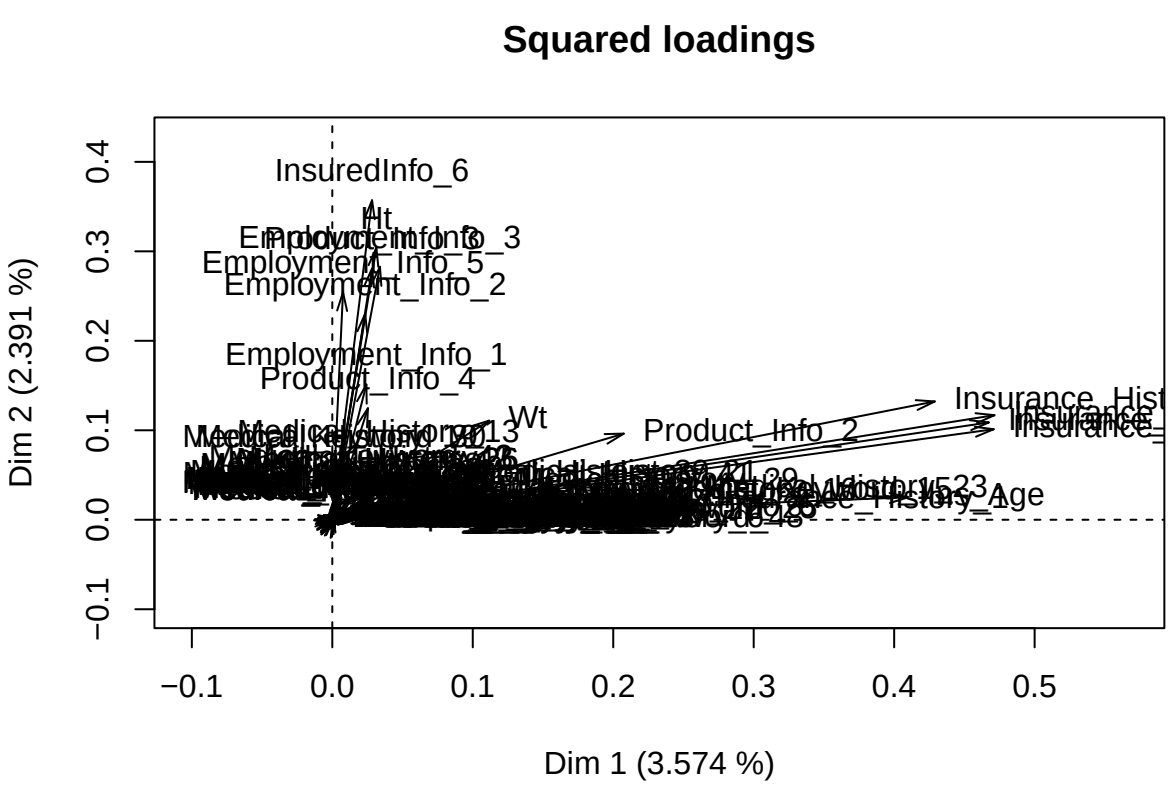
Following imputation, the PCA and CFS models were recreated:

```
## Warning: package 'PCAmixdata' was built under R version 4.3.3
## [1] 59381   118
```




Levels component map
Product_Info_3=24
Product_Info_3=20
Product_Info_3=4
Product_Info_3=19
Product_Info_2=A7
Product_Info_2=A2
Product_Info_3=29
Product_Info_7=3
Product_Info_2=B1
Product_Info_5=3
Dim 2 (2.391 %)
Dim 1 (3.574 %)
4
3
2
1
0
−1
−3
−2
−1
0
1
2


Correlation circle
InsuredInfo_6
Employment_Info_3
Employment_Info_5
Insurance_History_4
Insurance_History_3
Employment_Info_6
Medical_History_13
Product_Info_4
Employment_Info_1
Employment_Info_2
Ht
Dim 2 (2.391 %)
Dim 1 (3.574 %)
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2




```
##       dim1              dim2              dim3              dim4
##  Min.   :-6.22549   Min.   :-5.8665   Min.   :-5.4229   Min.   :-7.2561
##  1st Qu.:-2.01000   1st Qu.:-1.4409   1st Qu.:-1.3315   1st Qu.:-1.0961
##  Median :-0.09788   Median :-0.2578   Median :-0.1315   Median :-0.3984
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 1.63113   3rd Qu.: 1.2121   3rd Qu.: 1.2174   3rd Qu.: 0.5550
##  Max.   :12.25464   Max.   :11.4418   Max.   :10.2678   Max.   : 6.9564
##       dim5              dim6              dim7              dim8
##  Min.   :-5.9594   Min.   :-9.9695   Min.   :-9.09270   Min.   :-6.87074
##  1st Qu.:-1.0565   1st Qu.:-0.8189   1st Qu.:-0.69250   1st Qu.:-0.58624
##  Median :-0.1826   Median : 0.1192   Median : 0.07848   Median :-0.07608
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.9113   3rd Qu.: 0.9036   3rd Qu.: 0.82801   3rd Qu.: 0.40662
##  Max.   : 9.5186   Max.   : 9.7468   Max.   : 7.59285   Max.   : 9.64524
##       dim9              dim10             dim11             dim12
##  Min.   :-8.72862   Min.   :-6.58996   Min.   :-6.39327   Min.   :-7.1396
##  1st Qu.:-0.54120   1st Qu.:-0.79593   1st Qu.:-0.75429   1st Qu.:-0.5553
##  Median :-0.06365   Median :-0.03836   Median :-0.02222   Median :-0.1391
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.38442   3rd Qu.: 0.75556   3rd Qu.: 0.66924   3rd Qu.: 0.3250
##  Max.   : 7.65258   Max.   : 7.63806   Max.   : 9.01384   Max.   :11.3179
##      dim13             dim14             dim15             dim16
##  Min.   :-7.4739   Min.   :-7.85104   Min.   :-5.7014   Min.   :-8.53042
##  1st Qu.:-0.7162   1st Qu.:-0.60094   1st Qu.:-0.8313   1st Qu.:-0.45238
##  Median :-0.0330   Median : 0.06373   Median :-0.1366   Median : 0.01604
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000
```

```
##  3rd Qu.: 0.7193   3rd Qu.: 0.70734   3rd Qu.: 0.6982   3rd Qu.: 0.48757
##  Max.   : 6.7733   Max.   : 6.74961   Max.   :10.2797   Max.   :10.91519
##      dim17             dim18              dim19              dim20
##  Min.   :-8.7929   Min.   :-10.64627   Min.   :-8.92497   Min.   :-5.68425
##  1st Qu.:-0.6171   1st Qu.: -0.54681   1st Qu.:-0.51859   1st Qu.:-0.60981
##  Median :-0.1031   Median : -0.06243   Median :-0.03814   Median :-0.02634
##  Mean   : 0.0000   Mean   :  0.00000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.4857   3rd Qu.:  0.47893   3rd Qu.: 0.50520   3rd Qu.: 0.51027
##  Max.   :12.9185   Max.   : 10.03553   Max.   :10.20213   Max.   :10.42919
```

```
## Warning: package 'rcompanion' was built under R version 4.3.3
```

```
## Cramer V
##   0.1445
```

```
##  [1]  11  10  53  84  72  36  67   8   4 117  38  46  44  68  23  26  14  59  92
## [20]  61  94  22  33  24 111   9 107  57  48  70  56  18   2
```

Based on the results of the PCA and CFS, both methods are shown to be reproducible with the same results as the analysis by Boodhun and Jayabalan. This can be further used on other data in order to prepare for supervised learning algorithms, such as Multiple Regressions, REPTrees, and Artificial Neural Networks. These feature selections are a key component in being able to produce accurate outcomes for the methods. Without such selection methods, the validation of further methods must be called into question. In the case of this model, the analysis of Boodhun and Jayabalan is held valid.

Analysis of Normative Consideration

While such algorithms are efficient for companies to evaluate the risk levels of insurance patients, there are a number of ethical concerns with this process. The primary issue at hand stems from the ambiguity of data privacy and consent. The process details that a step includes ensuring anonymity of its applicants. However, this information is still accessible prior to data cleaning and manipulation. This is best highlighted in the case of this analysis, where the initial data was easily discovered and used to validate the results of the initial analysis. Even in the case that such information is not provided, the primary variables that remain, employment history and medical history, can still be traced back to the applicant. Due to this concern, the purpose of consent is considered. While applicants may comply to the standards of consent provided by the company, the consideration of informed consent v tacit consent arises. Are applicants truly provided with all information that is digestible to them and are they providing their consent knowing everything they are complying to? In the case of tacit consent, potential misuse of data can occur, and the true benefits of providing consent must be explored.

Apart from data privacy and consent, an argument of fairness arises in the process to determine the risk prediction for each applicant. Justice as equality versus merit serves as a major point of consideration. In this field, merit would be used as the main factor, exemplified by the use of employment history and medical history as the primary variables to consider. On the other side of this argument, are these variables a fair measure of providing that value to the applicants? Without such considerations, the reliance on these variables could potentially result in a form of algorithmic bias that favors certain types of applicants over others. Overall, a consideration of the true degree of consent and the extent to which this process is fair must be done.

Conclusion

Undoubtedly, the utilization of supervised learning algorithms provides a faster, more automated process for risk assessment. These algorithms are also determined to be valid in the results they produce, and are capable of reproducing initial output, proving how efficient and accurate such learning algorithms are. Yet, these types of advanced learning methods also cross a line of acceptable ethical considerations, particularly

in regards to the field in which such algorithms are being implemented. In this case of risk assessment and life insurance, the data being handled by such algorithms is that of extreme privacy. Determining who has access to such data is a serious consideration, followed by how much of that data should even be provided to others. In all cases, a question rises: is the applicant aware that their information is readily accessible? As determined by the case of this paper, the main company in charge of this study, Prudential, has provided such information in a public space, albeit in an anonymous manner. Even in this case, so much information about each applicant is revealed; if there are others who have access to more advanced forms to look at the data, what other information can be found? May it even be possible to trace the information back to the applicant, even if they are supposedly anonymous? While the use of such algorithms are definitely a step to be taken in the future for the field of risk assessment, the concerns of privacy and consent regarding applicant data must first be addressed for this practice to truly be safe.