Examining the Risk Factors Associated with Hypertension in the Australian population

Executive summary

Hypertension is a condition characterised by consistent high blood pressure readings, effecting ~4 million Australians. Shockingly, 50% of sufferers of severe hypertension die within 10 years from cardiovascular implication including heart disease, stroke or uraemia. However, in 90% of cases, the causes are elusive and since the disease is asymptomatic, is often left untreated. It is therefore important to identify the risk factors that are associated with this disease. To study the risk factors of hypertension the data from the 2011-2013 Australian Health Survey will be utilised.
This paper will investigate the prevalence of known risk factors of hypertension and attempt to classify people as hypertensive using their age, diet and physical measurements.

Key questions

The key questions addressed are:

- What variables (from a list of known factors compiled from papers) can we identify as being related to the occurrence hypertension?
- Is there a relationship between salt intake and blood pressure in those who haven't been told they have hypertension?
- What relationship does alcohol intake have on the incidence of hypertension/blood pressure?
- Can we use a person's characteristics (BMI, waist circumference, age…) to classify them as hypertensive or not?

Main findings

The main findings of this report are:

- 9 of the 13 variables showed significant difference in hypertension groups
- Salt was found to have no correlation with blood pressure
- Alcohol in small amounts may decrease blood pressure but in general contributes to an increase
- Logistic regression and SVM (linear kernel) perform quite well in classification with 76 % success rate

Shortcomings

The major shortcomings of this analysis stemmed mainly from the data used. Firstly, about a quarter (~3000) of the observations were lost in the cleaning process due to missing values in certain variables, there is possible bias in that some characteristics of a person could be common in the non-responders. The response variable had most people classed as 'never told has hypertensive disease', this is not a clear response to whether the person doesn't have the disease as they could but haven't been diagnosed, the method used to remedy this was not guaranteed to work either. Another variable could have had issues that could not be addressed, a large portion of alcohol intake responses were zero, this did not seem like a true reflection of the populations habits so it is possible that non respondence was simply coded as zero. Lastly the classification problem was likely effected by the fact that people with hypertension have changed the habits that lead to the onset of the disease, as since being diagnosed they would be receiving treatment and dietary/exercise plans. This effect was possibly why exercise and most dietary factors were not significant in classification models.

Problem

What continuous variables (from a list of known factors compiled from papers) can we identify as being related to the occurrence hypertension?

- The data used had 132 variables, with continuous and categorical data types, including the persons state of hypertension diagnosis (do they have it now, never been diagnosed etc..) and systolic and diastolic blood pressure measurements. After narrowing down the variables to a set of those mentioned in previous studies and are continuous, we ask the question: which variables are different in those people that have and do not have hypertension? This can be answered though observing the distributions of measurements through boxplots for those with and without hypertension and performing a statistical test if there is a significant difference at the 0.05 significance level.

Is there a relationship between salt intake and blood pressure in those who haven't been told they have hypertension?

- Salt is one of the most commonly known factors for high blood pressure, it would be a fair statement to say that when asked what one should do to reduce their risk of hypertension or to cure it, they would reply "cut down on salt in your diet". This question will aim to determine if salt intake really does have a significant impact on blood pressure. The boxplot of salt intake between those with and without hypertension will reveal some information into the problem, then the analysis will continue with determining if salt intake is correlated with blood pressure in individuals who have not been medically diagnosed with hypertension.

What relationship does alcohol intake have on the incidence of hypertension/blood pressure?

- Another factor said to be associated with high blood pressure is alcohol intake, but the relationship might not be so simple as a linear correlation. Some reports suggest that certain small amounts of alcohol intake on a regular basis can lead to lower blood pressure that those who abstain, and it is generally accepted that large intakes lead to high blood pressure. By analysing the incidence of hypertension in different classes of drinkers (non-drinkers,lower half and upper half of drinkers) and looking at how alcohol intake is correlated with amount of alcohol consumption.

Can we use a person's characteristics (BMI,waist circ., age…) to classify them as hypertensive or not?

- Hypertension is not something a person can know if they have without going to a doctor or performing a blood pressure measurement, for this reason it is important to know what factors can make a person at risk of the disease. This motivates the question if a classifier can be made that uses a person's physical measurements, diet and other factors to determine if that person is at risk of hypertension. This problem becomes one of classification: can we classify someone as hypertensive using these factors to some degree of accuracy.

Data

The data used to answer the questions is from the 2011-2013 Australian Health Survey. This was a large scale survey in which around 12,000 people from all ages (>2yr) and places responded to a wide variety of questions. The questions include personal traits, dietary habits, physical measurements, incidence of hypertension and many more. To reduce the number of variables used in our analysis, research was conducted into all known factors that may be related to hypertension. From these, all that had continuous measurements were kept for analysis.

These factors were (with there code name used):

- Percentage of daily energy intake from alcohol (%) 'alcohol.energy'
- Body mass index (kg m-2) 'bmi'
- Daily salt intake (mg) 'salt'

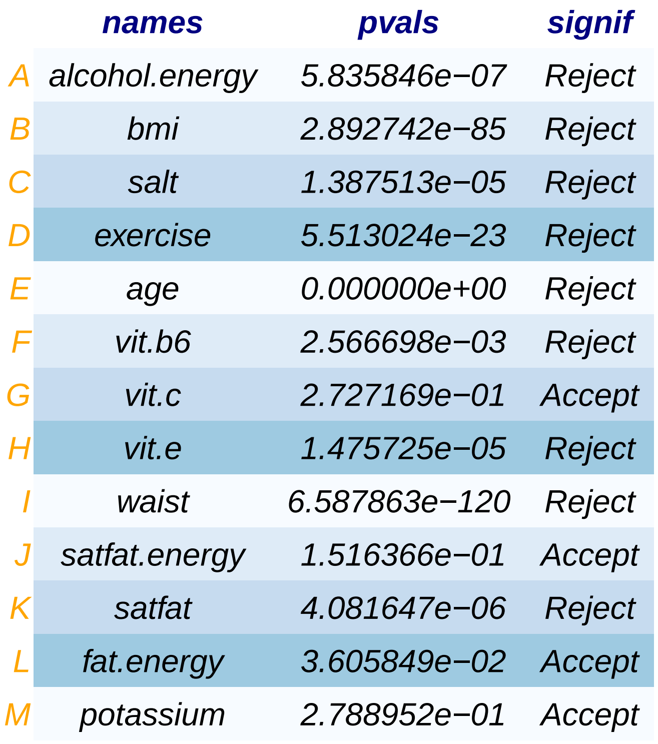
- Weekly exercise total (min) ‘exercise’
- Age (yrs) ‘age’
- Daily intake of vitamin B6 (mg) ’vit.b6
- Daily intake of vitamin C (mg) ‘vit.c’
- Daily intake of vitamin E (mg) ‘vit.e’
- Waist circumference (cm) ‘waist’
- Percentage of daily energy intake from saturated fat (%) ‘satfat.energy’
- Daily intake of saturated fat (g) ‘satfat’
- Percentage of daily energy from all fat (%) ‘fat.energy’
- Daily intake of potassium (ug) ‘potassium’

The hypertension was categorical with 4 categories: 1) Ever told has hypertensive disease, still current and long term 2) Ever told has hypertensive disease, still current but not long term 3) Ever told has hypertensive disease, not current 4) Never told has hypertensive disease. To simplify analysis, the hypertension variable was transformed into a factor with only 2 values ‘yes’ if they currently have hypertension and ‘no’ if they don’t. The last category ‘never told has hypertensive disease’ was very interesting. It creates a problem in that we don’t know if some of these people have hypertension or not. We wish to classify people as having hypertension or not, yet we don’t have the true state of hypertension for these people. For this reason, the data used in classification was transformed so that those in category 4 with systolic blood pressure >140 mmHg and/or diastolic blood pressure >90 are not transformed as ‘no’ but instead ‘yes’. This is a bold attempt to reduce error in the classification problem. The cleaning of the data of the selected variables was a simple process of removing data points with missing responses or unrealistic values.

Analysis

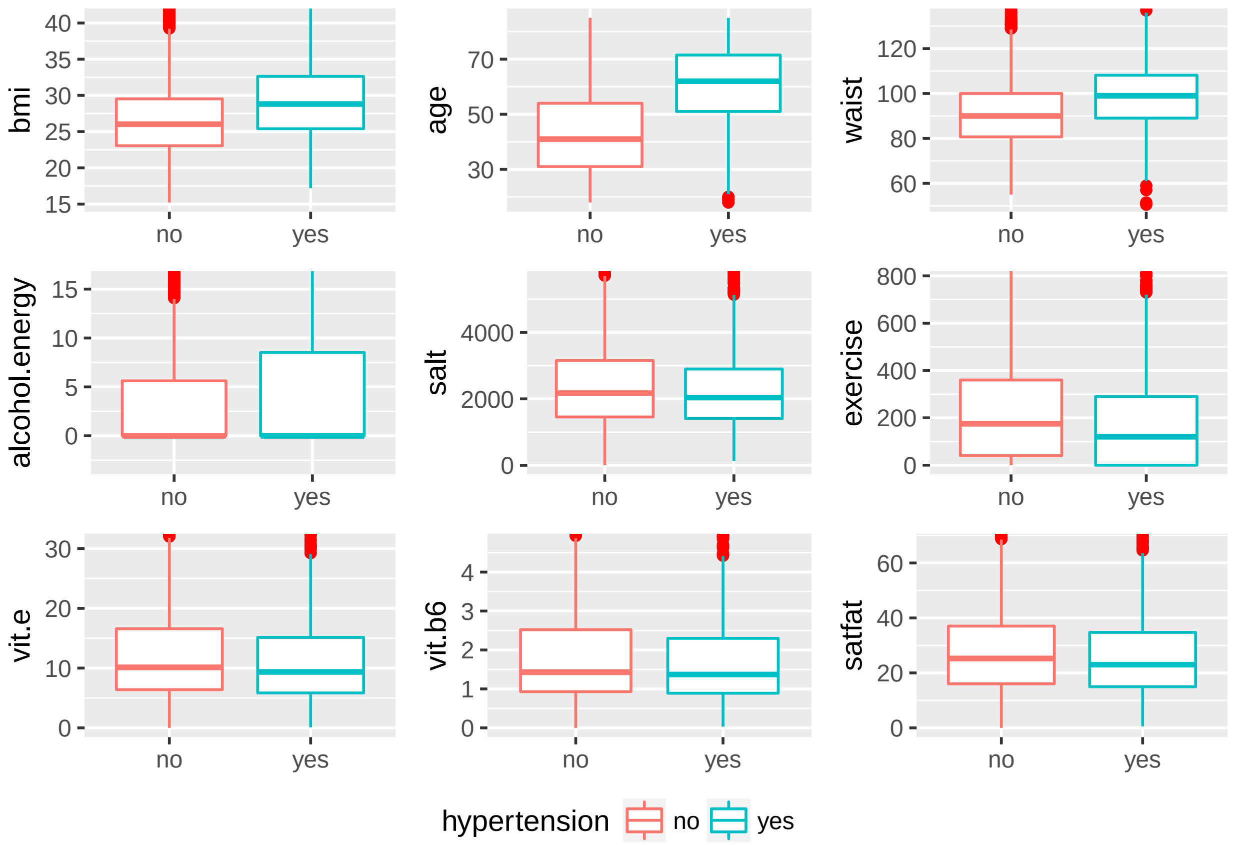
Variables related to incidence of hypertension

All the variables used have been known to have some relation to hypertension. To test if there is a difference in the typical values for these variables between the hypertensive people and normotensive (normal blood pressure) a Mann-Whitney U test was performed for each variable. The null hypothesis is that there is no significant difference in the distribution of the values for the variables between the hypertensive and normotensive groups, the alternate hypothesis is that the distributions are different. A significance level of 0.01 was used and a Bonferroni correction made, this was done to ensure the conclusions were accurate.



	names	pvals	signif
A	alcohol.energy	5.835846e−07	Reject
B	bmi	2.892742e−85	Reject
C	salt	1.387513e−05	Reject
D	exercise	5.513024e−23	Reject
E	age	0.000000e+00	Reject
F	vit.b6	2.566698e−03	Reject
G	vit.c	2.727169e−01	Accept
H	vit.e	1.475725e−05	Reject
I	waist	6.587863e−120	Reject
J	satfat.energy	1.516366e−01	Accept
K	satfat	4.081647e−06	Reject
L	fat.energy	3.605849e−02	Accept
M	potassium	2.788952e−01	Accept

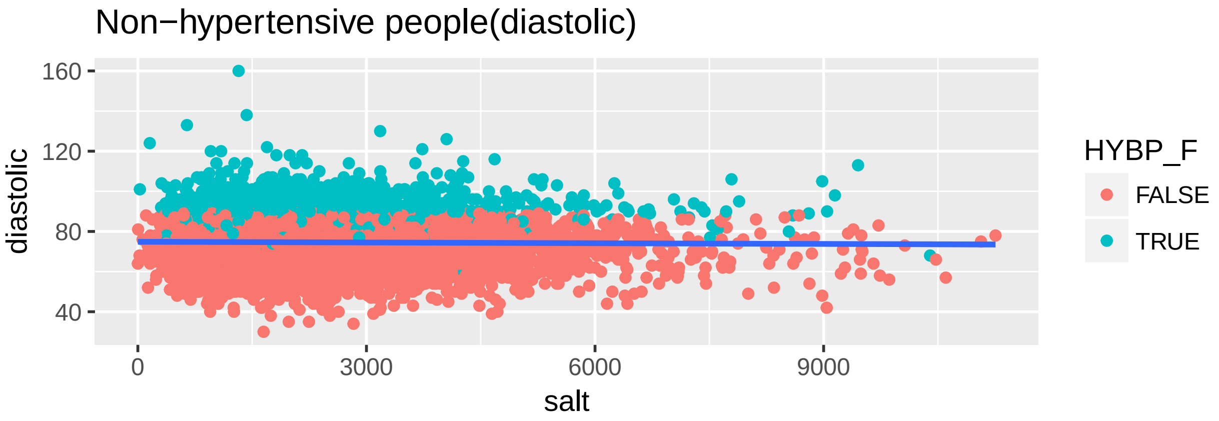
The test has identified 9 of the 13 variables that accept the alternative hypothesis that the distribution for the values of those variables is different between hypertensive and normotensive groups. To gain better insight to the differences the boxplots of these variables for each group are plotted below.

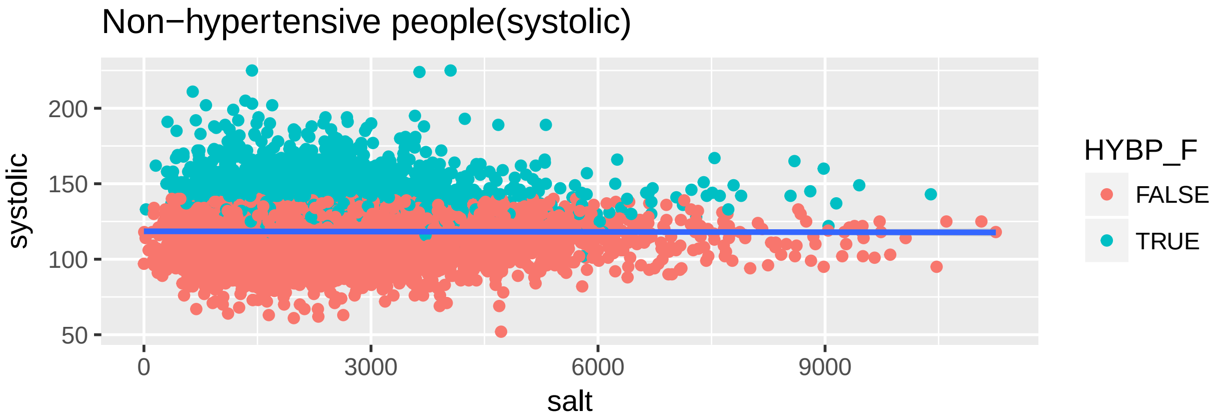


The boxplots give information of how the distributions of the variables differ between the hypertensive and normotensive groups. The first row shows that those with hypertension have generally a larger BMI, are older and have a bigger waistline. It is not surprising that these variables are related in this way, it is interesting to note how different the age distributions are in median at almost 20 years. The second row shows that those with hypertension drink more and exercise less which is expected, but surprisingly hypertensive people appear to eat less salt. This is likely since those who are told they have hypertension are also told to reduce their salt intake, but one would expect this to also be true for alcohol as well if this were the case. The last rows have small differences as they are all nutritional factors, hypertensive people generally have less vitamin E and B6, surprisingly they also eat less saturated fat. It appears that some dietary factors have counterintuitive relations with hypertension and that there is possibly interference from treatment of the condition. This could mean that classification using these variables could be rendered useless as the trends they have are not reflecting how a risk factor should influence the incidence of the disease. To understand more about how some of these variables are related to hypertension, more analysis is needed.

Salt intake and blood pressure

Due to the counterintuitive result from the boxplots it is important to explore further the relationship between salt and hypertension. To do this the relationship between salt intake and the measured values for systolic and diastolic blood pressure is analysed. It would be inappropriate to include those people that have been diagnosed with hypertension in this analysis, as they likely have had treatment to reduce salt intake which would impact the goal of the analysis. Below is the scatterplots of diastolic and systolic blood pressure, those who with systolic blood pressure >140 mmHg or diastolic blood pressure >90 mmHg are coloured in blue to indicate those with a high blood pressure measurement.

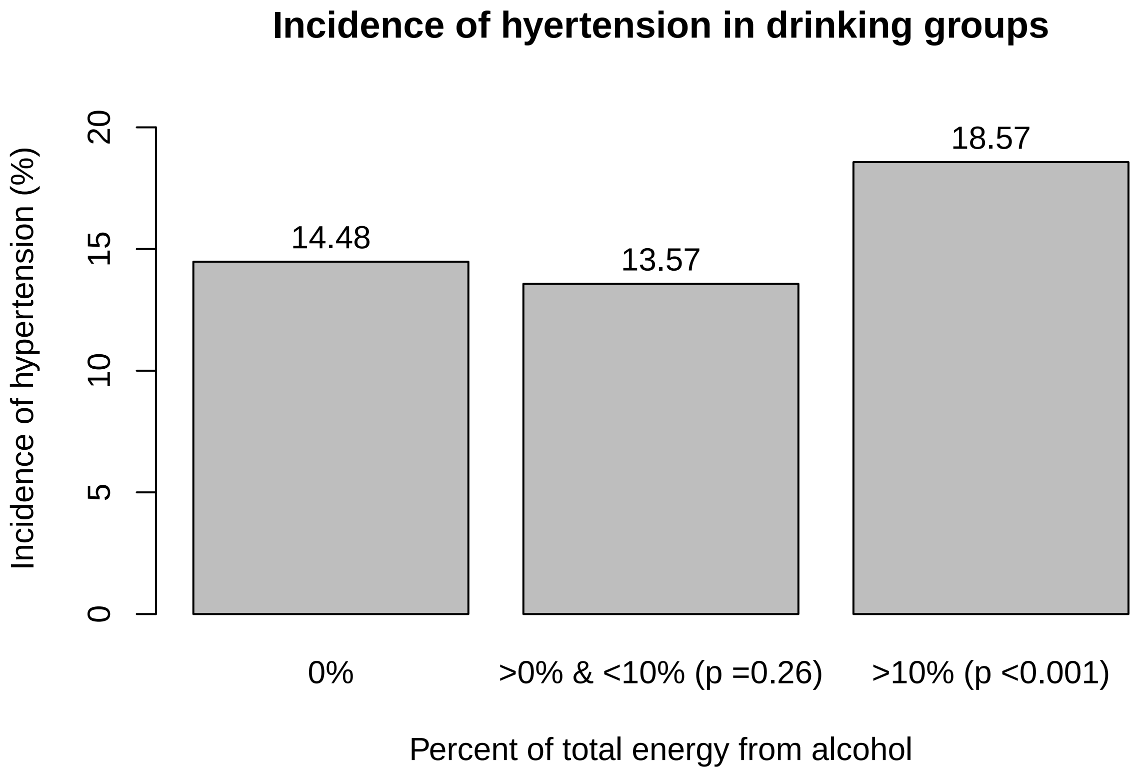




The p value for the slope is greater than 0.1 for both cases, indicating that there is no significant correlation between salt and blood pressure in people who have not been diagnosed with hypertension. This result simply indicates that there is no evidence to suggest that changing your salt intake will affect your blood pressure in any way. This suggests that maybe the only reason that salt was seen to have a significant relation with hypertension is the fact that those with the condition have cut down their salt intake intentionally as part of a way to treat the condition. However, this is purely a simple linear regression so we are not considering other confounding factors.

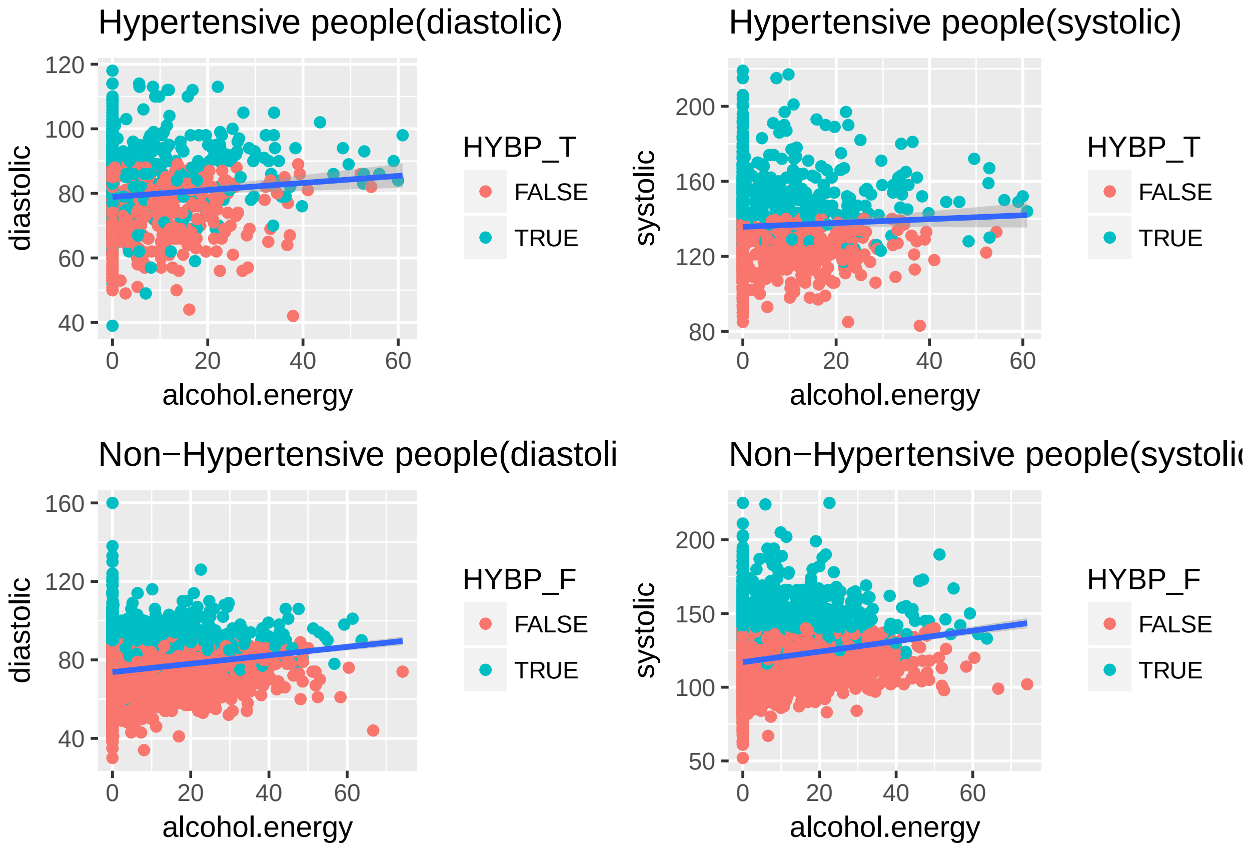
Alcohol and hypertension/blood pressure

It has been documented but not fully studied that light drinking can lead to lower blood pressure levels than abstaining from drinking all together. However, it is widely accepted that alcohol is positively correlated with blood pressure increase. The variable used here will be the percentage of energy from alcohol an individual consumes. To observe the how the incidence of hypertension changes with drinking patterns, a bar plot is constructed to show the proportion of hypertension in non-drinkers, lighter drinkers (bottom 50% of drinkers) and heavier drinkers (top 50%). P values were determined using a binomial test.



Just like the report referenced above, it is found that light drinkers have a lower incidence of hypertension that abstainers. However, like the report, the difference between light drinking and abstaining is not statistically significant with a p value of 0.26 and the heavy drinkers had significantly more hypertension with a p value less than 0.001. The phenomenon in the light drinkers would need to be more thoroughly studied to prove a significant decrease in hypertension incidence, it may be caused by some physiological protective effect small intakes of alcohol has.

To investigate the relationship alcohol has with hypertension plots were made to see how alcohol is correlated with systolic and diastolic blood pressure. The plots will be split up by data for hypertensive people and normotensive people. Again, those who with systolic blood pressure >140 mmHg or diastolic blood pressure >90 mmHg are coloured in blue to indicate those with a high blood pressure measurement.



All but the top right has significant (at 0.05 level) positive slopes indicating a clear rise in blood pressure with increasing alcohol. The gradients in both non-hypertensive plots are much larger than in the hypertensive plots. Diastolic blood pressure rises 0.2mmHg every percent more of your energy intake being alcohol and systolic blood pressure rising 0.35mmHg every percent, in non-hypertensive people. In both hypertensive models, the gradient is 0.1mmHg/percent energy intake from alcohol, with the systolic having an insignificant slope.

The reason the plots have been divided by hypertension groups is that it was assumed there would be a difference in the hypertensive people as they are likely undergoing treatment and altering their alcohol intake due to the presence of the condition. This assumption appears warranted that the plots offer different relationships with alcohol. Another interesting observation to note in bottom left plot is there seems to be people in the hypertensive group with high systolic BP (blood pressure) and quite low diastolic BP. Possibly certain hypertensive conditions have this kind of behaviour or it is a treatment which is causing this.

Classification Analysis

After the cleaning of original data, we were left with 7512 samples and they were split up (randomly) as follows:

- Training set (60 %) - 4507
- Testing set (20 %) - 1502
- Cross-Validation (CV) set (20 %) - 1503

The below three classifiers were used on the cleaned data:

```
- KNN: K-Nearest Neighbour
- Logistic Regression

- SVM:  Support Vector Machine (Linear and radial kernel)
```

Reasons for choosing the above three classifiers:

None of the three models make any strict assumptions about the decision boundaries for the data. In this instance we only looked at linear and radial kernels for SVM but (time permitting) we could have looked at polynomial kernels as well. The only assumption logistic regression makes is that the logit is a linear function of the predictors.

Procedure:

The whole report was run with set.seed(1)

KNN

- Data was scaled using scale()
- 10 fold Cross-validation was used to pick the parameter "k", this was run 30 times and training data was used in this step. We did not think it would be appropriate to pick a parameter based on CV-set (as it was only 20% of the data).
- We used k = 20 based on the value of k that gave the minimum CV error in step 1.
- Using this value of k, we tested the KNN classifier on the testing data set.
- knn() package from the MASS library was used to perform classification.

Logistic Regression

- Full model (using all predictors) was fit using the training data set.
- glm() function was used from the base R.
- Then model selection was done using the step() function using both AIC and BIC criterion. This was done starting with both "full" and "null" model.
- AIC and BIC picked different models irrespective of starting with a full or null model. Both had "alcohol.energy", "bmi", "age" but AIC also picked "waist" (bmi and waist are highly correlated but bmi had a bigger coefficient).
- Test set was used on both models to check their performance and the BIC model performed better (21.7 % vs 22.02 %) But when using 10-fold CV, both models performed similarly (around 23 % error rate)
- Also, an arbitrary cut-off of 0.5 was chosen to assign the class labels to the prediction with >0.5 being marked as having hypertension. (ROC curve of this model suggests it is still a good model with another arbitrary cut-off)
- We picked the BIC model as it was simpler.

SVM

Linear kernel:

- tune() function from library e1071 was used for finding (the best model with appropriate cost parameter within the CV loop).
- Data is scaled internally by the tune() and svm() functions.
- For the cost parameter we tried the values (0.001, 0.01, 0.1, 1, 5, 10, 100). Again the training set was used instead of the CV set.
- 0.1 was picked by the tune function as the best parameter and also output the best model by fitting this cost parameter, which resulted in 2404 support vectors.

- We used the remaining data to test the performance of this model. Using just the test set (~1500 samples) gave an error rate of 21.69% Using all the remaining samples (~3000) gave an error rate of 22.77%

Radial kernel:

- Used the exact same process as above but this time there were two parameters to pick, "c", "gamma"
- We used cost=c(0.1, 1, 10, 100, 1000), gamma=c(0.5,1,2,3,4) as inputs to the tune() function and it output c = 1 and gamma = 0.5 as the best parameters.
- The best radial model had 3425 support vectors.
- We tested this model with the CV+test sample combined and it gave far worse error rate. Using all the remaining samples (~3000) gave an error rate of 24.82%

SVM

The best model from CV-tuning for the SVM linear kernel:

```
##
## Call:
## best.tune(method = svm, train.x = Ytrain ~ ., data = dat, ranges = list(cost = c(0.01,
##     0.1, 1, 5)), kernel = "linear")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  0.1
##       gamma:  0.07692308
##
## Number of Support Vectors:  2404
##
##  ( 1199 1205 )
##
##
## Number of Classes:  2
##
## Levels:
##  no yes
```

Confusion matrices - SVM:

Linear-CV+Test, Linear-test, Radial-CV+Test

```
##      no yes   no yes   no yes
## no  968 219 1941 463 2002 586
## yes 107 209  221 380  160 257
```

Performance of SVM models (%):

```
##         Linear-CV+Test Linear-test Radial-CV+Test
## Success      0.7723794   0.7831005      0.7517471
## Error        0.2276206   0.2168995      0.2482529
```

Linear kernel model outperforms the radial model by a significant margin.

Logistic Regression

Summary of results for Logistic regression

	Full	Full-BIC	Full-AIC
(Intercept)	**-7.012** ***	**-6.796** ***	**-7.168** ***
alcohol.energy	**0.013** **	**0.016** ***	**0.015** ***
bmi	**0.059** ***	**0.084** ***	**0.058** ***
salt	0.000		
exercise	-0.000		
age	**0.065** ***	**0.065** ***	**0.064** ***
vit.b6	0.002		
vit.c	-0.000		
vit.e	-0.001		
waist	**0.011** *		**0.012** **
satfat.energy	0.003		
satfat	0.000		
fat.energy	-0.007		
potassium	-0.000		
N	4507	4507	4507
logLik	-2108.271	-2115.253	-2111.880
AIC	4244.542	4238.507	4233.759
BIC	4334.330	4264.160	4265.826
null.deviance	5325.958	5325.958	5325.958
Deviance	4216.542	4230.507	4223.759

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Picking the FULL-BIC model based on prediction performance:

$$\mathbf{ln(\frac{p(X)}{1-p(X)}) = -6.796 + 0.016\ alcohol.energy + 0.084\ bmi + 0.065\ age}$$

Where: $p(X)$: *probability of developing hypertension*

And it is given by: $p(X) = \frac{1}{1+e^{-(-6.796+0.016\ alcohol.energy+0.084\ bmi+0.065\ age)}}$

Cross Validation error for the BIC model:

```
## [1] 0.2376831
```

This is higher than test-set data!

Model performance:

Full, AIC, BIC confusion matrices

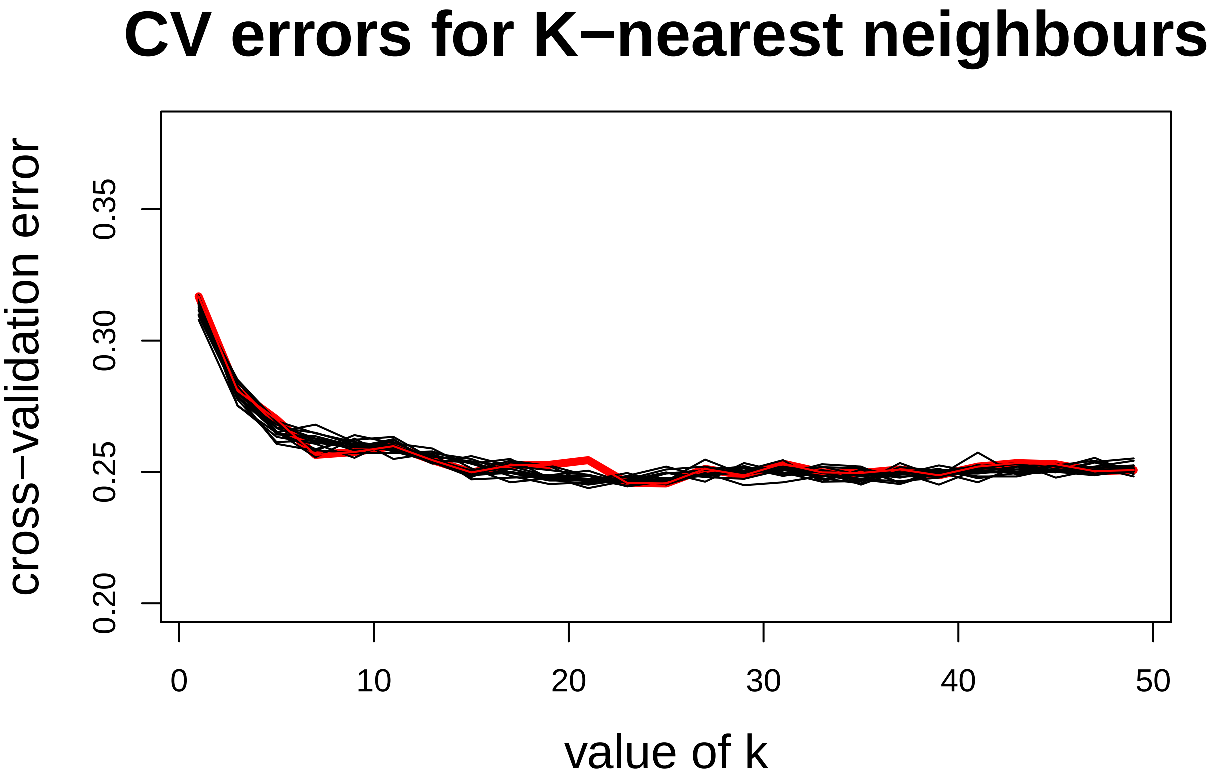
```
##      no yes  no yes  no yes
## no  966 223 969 225 974 226
## yes 109 205 106 203 101 202
```

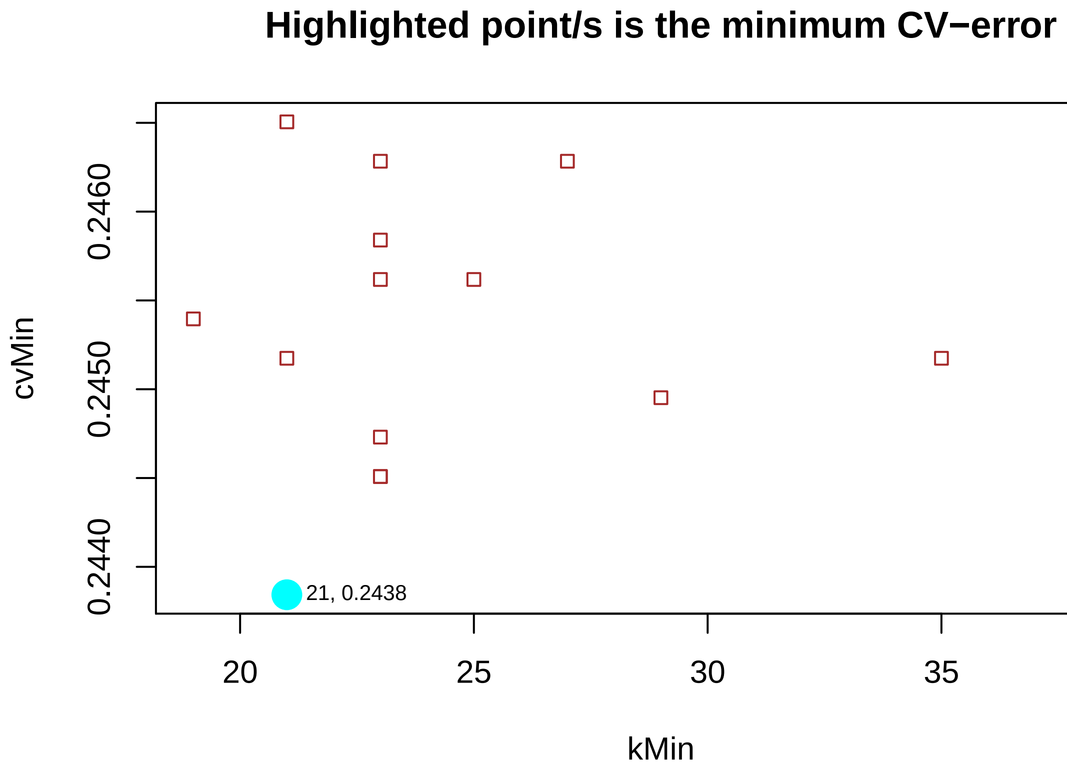
Performance of Logistic regression models (%):

```
##         Training-Full Test-Full Test-AIC Test-BIC   CV-BIC
## Success       76.9026  77.91084 77.97738 78.24351 76.23169
## Error         23.0974  22.08916 22.02262 21.75649 23.76831
```

Both AIC And BOC model are pretty close but their CV-error is markedly higher. It is more informative to look at their performance using the ROC curves.

KNN


CV errors for K−nearest neighbours
cross−validation error
0.20
0.25
0.30
0.35
0
10
20
30
40
50
value of k



Performance of KNN classifier:

KNN-Test, KNN-CV+Test-set confusion matrices:

```
##      no yes   no yes
## no  986 261 1978 550
## yes  89 167  184 293
```

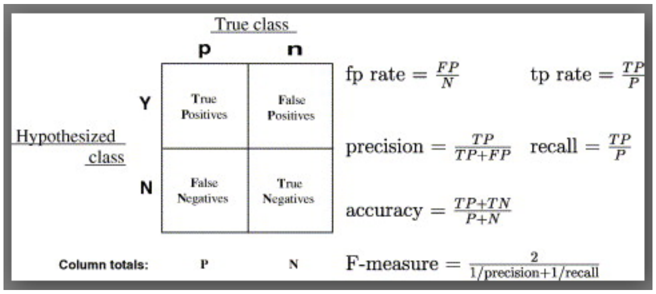
Performance of KNN classifier (%)

```
##         Test-set CV+Test-set
## Success 76.71324    75.57404
## Error   23.28676    24.42596
```

KNN model performs quite poorly compared to the other two classifiers.
Using the test-set it pares well but the above plot of CV errors shows that it does not dip below 24 % when cross validation is used.

ROC and confusion matrices

The below illustration is taken from “An introduction to ROC analysis” by Tom Fawcett.



		True class	
		p	n
Hypothesized class	Y	True Positives	False Positives
	N	False Negatives	True Negatives
Column totals:		P	N

fp rate $= \frac{FP}{N}$ tp rate $= \frac{TP}{P}$

precision $= \frac{TP}{TP+FP}$ recall $= \frac{TP}{P}$

accuracy $= \frac{TP+TN}{P+N}$

F-measure $= \frac{2}{1/\text{precision}+1/\text{recall}}$

Most of the performance measurements used in our analysis can be calculated from the confusion matrix as show in the above illustration.

The two main metrics we used are:

$Success\ rate = \frac{True\ Positives+True\ Negatives}{Total\ samples}$

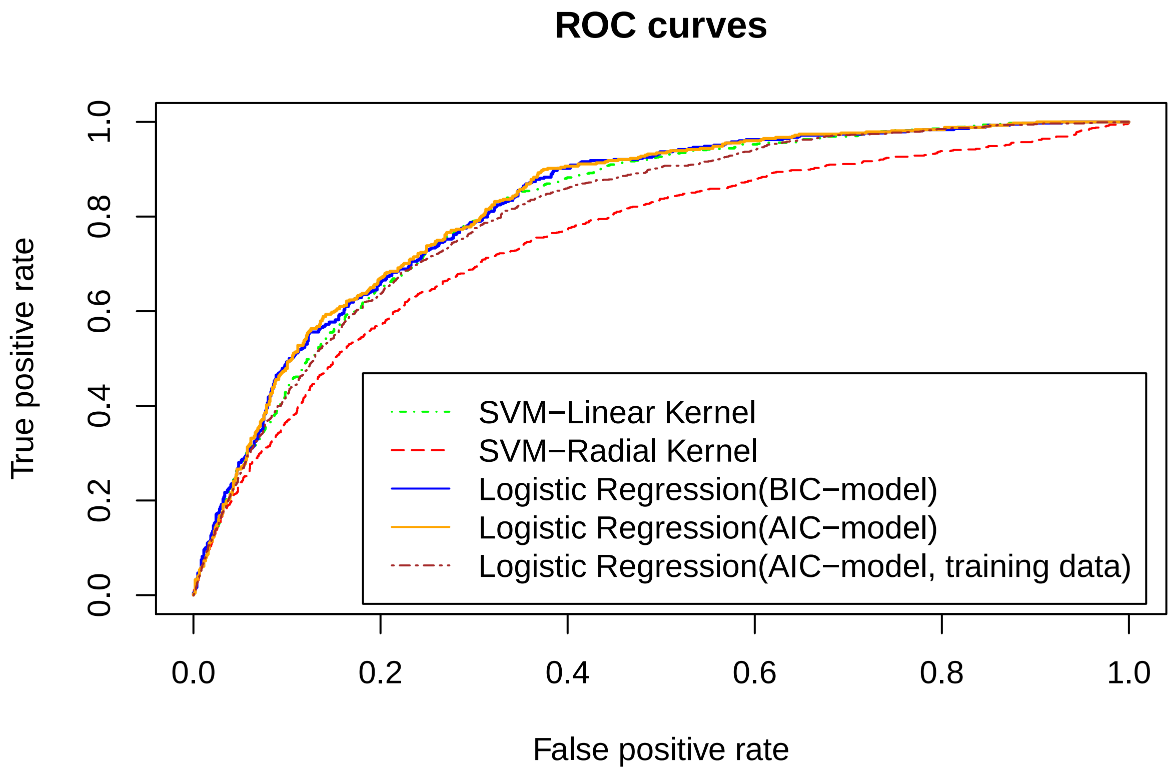
$Error\ rate = \frac{False\ Positives+False\ Negatives}{Total\ samples}$

I.e., to get the success rate add the diagonal entries and divide by the total sample size and for error rate add the other diagonal entries and divide by the total sample size.

How to read the ROC plot:

Y axis shows the “True positive (TP) rate” and X axis shows the “False positive (FP) rate”. Best performing models are the ones with the curve towards the northwest.

The below ROC curves calculate the FP and TP rate from several confusion matrices that use different cut-offs to classify the scores. For example, we used 0.5 as the cut-off for Logistic regression classification once we got the prediction scores. We could have easily picked some other cutoff. Relying solely on a single confusion matrix analysis will be misleading instead ROC curves paint a picture of how the model performs over a range of cutoffs.



Conclusion

This paper has been successful in examining the known factors for hypertension for the Australian population. It was shown that of 13 known variables only 9 had significantly different distribution between hypertensive and normotensive people at the 0.05 level. The variables that were determined not to be significant may be, they may confound with other factors that create the onset of hypertension.When the relation of salt to blood pressure was further analysised it was found that there was no linear correlation between either blood pressure measure and salt in people not diagnosed with hypertension. This could be that people with well known characteristics that indicate hypertension, and more likely to have it, have changed their salt intake as a precaution and hence the lack of relation.

It was gathered from other sources that alcohol intake might not have a simple relation to alcohol, with some data suggesting that small amounts may infact lower blood pressure. Similar analysis was conducted and it was found that there was a decrease in blood pressure in the lower half of drinkers, but it was not significant enough to say for sure. It was also confirmed that large alcohol intakes contribute significantly to higher blood pressure.

The above ROC plot shows that Logistic Regression model and SVM (linear kernel) model outperform the other models in classifying. They both have error rates between 22 - 24 % but looking at the above plot closely, it is clear that Logistic regression outperforms SVM (linear kernel) model at certain cutoffs but towards the northwest they all are packed closer except for the SVM (radial kernel) model. In conclusion we could use either model to classify with a success rate of atleast 76%.

References

1. *Appel LJ, Dietary approaches to prevent and treat hypertension: A scientific statement from the American Heart Association. Hypertension 2006; 47(2): 296–308.*
2. *Dakshinamurti K, Dakshinamurti S: Blood pressure regulation and micronutrients.Nutr Res Rev. 2001 Jun;14(1):3-44. doi: 10.1079/NRR200116.*
3. _Fujita T. Mechanism of Salt-Sensitive Hypertension: Focus on Adrenal and Sympathetic Nervous Systems. Journal of the American Society of Nephrology: JASN. 2014;25(6):1148-1155. doi:10.1681/ASN.2013121258._
4. *Whelton, Paul & He, Jiang & Appel, Lawrence & A Cutler, Jeffrey & Havas, Stephen & A Kotchen, Theodore & Roccella, Edward & Stout, Ron & Vallbona, Carlos & C Winston, Mary & Karimbakas, Joanne. (2002). Primary prevention of hypertension: clinical and public health advisory from The National High Blood Pressure Education Program. JAMA : the journal of the American Medical Association. 288. 1882-8. 10.1001/jama.288.15.1882.*
5. *Franz H Messerli, Bryan Williams, Eberhard Ritz, Essential hypertension, In The Lancet, Volume 370, Issue 9587, 2007, Pages 591-603, ISSN 0140-6736, https://doi.org/10.1016/S0140-6736(07)61299-9.(http://www.sciencedirect.com/science/article/pii/S0140673607612999)*
6. *Gleiberman L, Harburg E: Alcohol usage and blood pressure: A review. Hum Biol 1986; 58:1-31.*
7. *Russell M, Cooper ML, Frone MR, Welte JW. Alcohol drinking patterns and blood pressure. American Journal of Public Health. 1991;81(4):452-457.*
8. *Shankarishan P, Borah PK, Mohapatra PK, Ahmed G, Mahanta J. Population attributable risk estimates for risk factors associated with hypertension in an Indian population. Eur J Prev Cardiol. 2012;20:963–71.*
9. *Kaldmäe M, Viigimaa M, Zemtsovskaja G, Kaart T, Abina J, Annuk M. Prevalence and determinants of hypertension in Estonian adults. Scand J Public Health. 2014;42(6):504–510.*
10. *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani "An Introduction to Statistical Learning" http://www-bcf.usc.edu/~gareth/ISL/*
11. *TomFawcett, "An introduction to ROC analysis", Pattern Recognition Letters Volume 27, Issue 8, June 2006, Pages 861-874*

Statement of contributions

Cade:

- Wrote executive summary, problem,data, and analysis for variable relations,salt and alcohol.
- Created plots (in base R) and p values for sections written

Rahul:

- Wrote and coded all of classification section
- Utilised ggplot to improve quality of plots

Both Rahul and I worked well together on this project giving equal effort