# Group Project:
# *Evaluation of Statistical Model Accuracy*

Bryce Robinette Jacelyn Villalobos Kursten Reznik
David Koster Jacob Ruiz

## Introduction

In this paper, we compare the performance of four statistical models using randomly generated data with normal distribution $N(\mu, \sigma^2)$: The models we are investigating are: K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and the Logistic Regression model (GLM in R). That is, we analyze these models' accuracy with respect to sample size, number of input variables, the variance of the data, and normalization of the data.

We have been given four precise scenarios for which these models are to be tested. We outline them here:

**Scenario 1**
·Sample Size $N = 50$.
·Number of Inputs $\rightarrow 2$.
·No scaling or normalizing the data.

**Scenario 2**
·Sample Size $N = 500$.)
·Number of Inputs $\rightarrow 2$.
·Normalized Data.

**Scenario 3**
·Sample Size $N = 50$.
·Number of Inputs $\rightarrow 20$.
·No scaling or normalizing the data.

**Scenario 4**
·Sample Size $N = 500$.)
·Number of Inputs $\rightarrow 20$.
·Normalized Data.

In each of these scenarios, we will look at the accuracy of each model as it relates to the variance of the data. It should also be noted that this paper assumes that the reader is familiar with these statistical models and the programming language R.

## Materials and Methods

To begin, we perform each model with the parameters outlined in the four given scenarios. These four scenarios will give us our blueprint for analysis. Indeed, for each scenario, we run each model against a number of data sets of identical properties but with increasing variance in our randomly generated data. Each of these data sets are also sampled many times in order to return each model's average accuracy for that data's given variance. Moreover, we then evaluate the accuracy of each model as a function of the variance of the data sets in each scenario and plot the results. This methodology should yield the generalized accuracy of the model given the variance of the data.

Furthermore, we compare the accuracy of the models to gain insight into which model should likely be used given a set of data of normal distribution.

### Data Generation

We generate our data such that the response variable $y$ has two classes. That is, $y = 1$, or $y = 2$ with class means $\mu_1 = 1$ and $\mu_2 = 0$, respectively. Then we randomly generate values with distribution $N(\mu, \sigma^2)1$ for our predictors.

Should I include code here to show or just assume that our R-file is sufficient?

It should also be noted that when we have a smaller amount of observations $N$, there is a chance that when sampling the data into testing and training data that we could end up with an imbalanced sample of our response value. To mitigate this, we could sample an approximately equal amount of data with each value of $y$; however, since we are running a multitude of simulations on the data, we can rely on the central limit theorem.

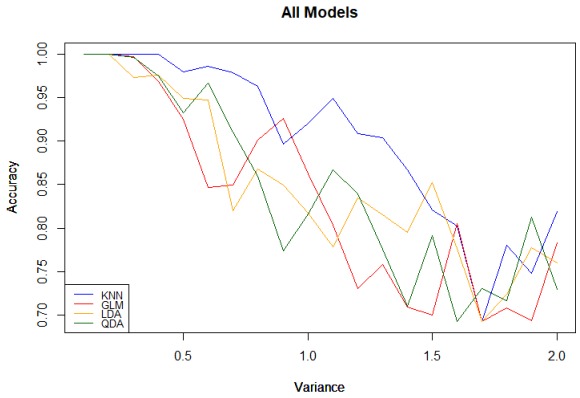
### Choosing our best $K$

In this section we present our best $k$. The very best $k$. In fact, it is the best $k$ in the history of $k$'s. So good it will blow your mind. Here is how we did it....

WORDS

We simulated our KNN model over a number of generated data sets with differing variances and chose the most common "best $k$" that was associated with these particular data sets. We took this common $k$ to be our generalized $k$-value for our following analysis.

## Scenario 1

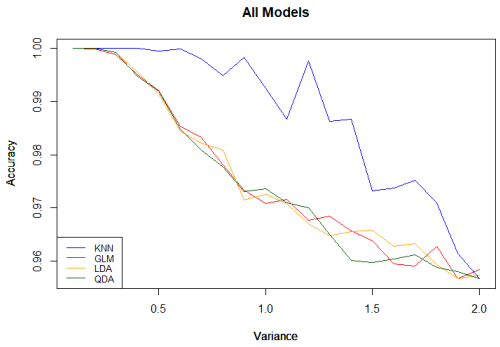



In this scenario, we compare out models with $N = 50$ observations, $p = 2$ parameters, and without scaling or normalizing the data.

From the resulting figure, we can conclude that all models diminish in accuracy as the variance increases. This is to be expected given that our data was generated randomly without any underlying function dictating the data generating process. Indeed, LDA and Logistic Regression assume a linear relationship between the response variable and the predictors while KNN is non-parametric. QDA is a decent compromise between KNN and any linear model as it assumes a nonlinear relationship between the predictors and the response.

In this scenario, the KNN model performs the best. This is to be expected due to the fact that we have chosen a relatively small $k$ for our analysis which makes our particular KNN model more flexible than it would be with a larger $k$-value. In this regard, we migh5 expect our KNN model to perform better under the assumed circumstances for our analysis

## Scenario 2




In this scenario, we take $N = 500$, parameters $p = 2$, and we scale the data.

We immediately see that the KNN model performs the best. Here is where we see the flexible model shine. Indeed, as we know with our brilliant minds, when we are presented with a large number of observations and the value of our predictors is small, we want to choose a more flexible model. Due to the large sample size, we are less likely to over-fit, even with a more flexible model.

The LDA, QDA, and Logistic Regression models all perform at approximately the same level of accuracy throughout the data variances. (*like the Backstreet boys → TELL ME WHY???*)

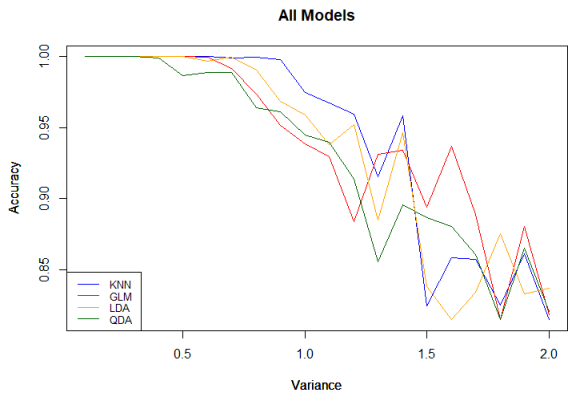
The other three models are less flexible the KNN. This incl;udes QDA. QDA allows for more flexibility but it requires more parameters to estimate the accuracy.

If you have many classes and not so many sample points, this can be a problem for the QDA model.

All of our models seem to be doing well; however the flexible nature of the KNN model makes it a superior predictor in this case.

IDENTICAL Y-AXES/VERTICAL SCALING

### Scenario 3




Now we take $N = 50$, and increase our predictors to $p = 20$. The data is not to be scaled in this scenario.

As we can see, this is some bullshit...

From the resulting plot, we immediately conclude that an increase in our predictors $p$ has yielded higher model accuracy over increasing variance of the data.

I suppose that one may conclude in cases where data variance is high, LDA and QDA may perform better with a larger number of predictors. However, we should be wearing of increasing the number of predictors too high that we start over-fitting and incorporating too much noise in our models.

Comparisons of KNN and GLM are to come as their information is made available.

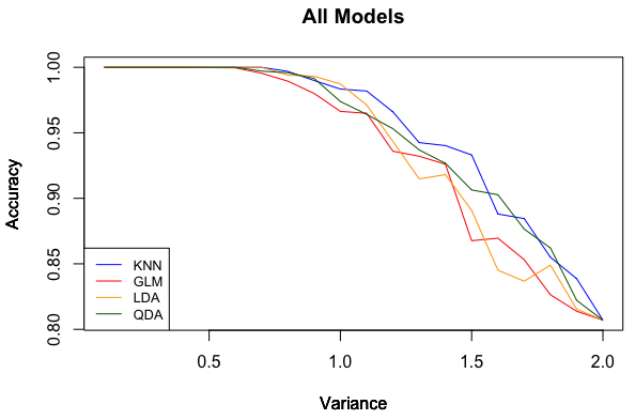
• Maybe write something about scenario 3 having a problem with QDA. When we have too small a sample sixe to compare with our predictors, then shit gets whack.

Since we have designated or predictors to be $p = 20$ it may result that the QDA performes in a simi.l,ar fation to LDA since we are not allowing the model th convenience of *best predictora*. That is, QDA (in general) needs a larger amount of predictors, but if the predictors have little significance, then id doeznt contribute the overall model accuracy andf may lead to an overfit and shit and stuff and yeah dawg.

NO... Dr, Chan. we must bhave more data. If you dont like it, its Davids fault

## Scenario 4

We now increase our number of observations to $N = 500$, keep or parameters $p = 20$, and scale our data.




knn will operform a bit worse wityh a larger number of predictors since we awill almosy inevbitably introduce noise into the model.

as the number of predictors increases we will need to use a more inflexible model. Having a high number of predictors can lead to over fitting of the data.

what we wsanted to show her e was when n is large the models will will in a general sense perform better sinse the number oif pdata pointds will help inform the ,odel (tell me why).... however, we should expect toi see the KNN and thre QDA to outperform the linear models due to the fact that with a larger number of predictors, nonlinear and non-parametric mm,odels should outperform linrear models that are less felxible that their nonlinear counterparts.

## Writing Notes

•Would we have suspected the QDA to perform better because of its ability to accommodate more flexible decision boundaries that could be present in randomly generated data?

•The number of parameters needed to estimate QDA increases faster than LDA. The number of parameters estimated in LDA increases linearly with $p$ while that of QDA increases quadratically with $p$. So, from the resulting analysis, we can see that when we have a larger data set and the number of predictors is large, the Quadratic Discriminant Analysis performed better than the linear models due to the fact that with a larger set if data(larger variance) in scenario 4 we are give a larger value of predictors, which in turn benefits the quadratic model to accept nonlinear relationships between the response variable and the the predictors.,

# Discussion

## Extensions?

Instead of normally distributed data, we could assess performance or accuracy of the models over other distribution sets, such as a uniform distribution.

# 1 Conclusion

# 2 Credits

McDonald's: Obviously.

Alcohol: for its continuing effort to keep me sane.

Sir David Attenborough: for the love of all that is good.

I dunno... maybe my cat?

# References

ME... FOOLS... I am the only reference you need.

JK seriously give yer references..

### GLM (to be rewritten later as part of the scenario analysis)

Logistic regression is going to have a different set of standards than our other models (except LDA). It is assumed that the reader recollects that Logistic Regression is parametric, while KNN is non-parametric.

Use your brain for a moment.... You then realize that LDA and GLM require linearity for any meaningful result. QDA is a decent compromise between KNN and any linear model, (such as GLM), since QDA also assumes non-linearity.

Use your eyes and read:
KNN will support non-linear solutions where GLM can only support linear solutions. Furthermore, KNN cannot yield confidence levels due to its selection nature where GLM can because it has the property of linearity.

Neighborhoods have meaningful weight....

In this regard, I have to say at this particular moment, that *Logistic Regression* is a largely unnecessary model when dealing with data generated with a known normal distribution without a further informative function dictating the outcome or response variable.

Like, for reel check it motha fuggas..... GLM and LDA are based on identical assumptions of linearity and statistical regression, making one no more informative than the other in our particular cases with randomly generated data derived from $N(\mu, \sigma^2)$.