# Executive Summary– Build and Evaluate Classification Model and Participate in the Multi-Class Prediction of Obesity Risk Kaggle Competition (Late Submission).

By Samuel Mbah Nde – DDS 8555 – Assignment 5

## Project Overview.

The work I did this week focused on performing multiple classification tasks. The work is divided into two sections. In section 1, I focus on answering the conceptual questions from Chapter 4 of ISLP (James, Witten, Hastie, Tibshirani, & Taylor, 2023). In section 2, I implement 5 multiclass classification model models and submit the most performant model in the Multi-Class Prediction of Obesity Risk Kaggle (Reade & Chow, Multi-Class Prediction of Obesity Risk, 2024). The code used for this analysis can be found in my GitHub repository https://github.com/ndesamuelmbah/DDS-8555/tree/main/AssignmentFiles/Week5 (Nde, 2024).

Section 1: Conceptual Questions.

Question 1: This is question 1 extracted from page 193 of ISLP (James, Witten, Hastie, Tibshirani, & Taylor, 2023) and is displayed in the screenshot below.

A close-up of black text

AI-generated content may be incorrect.  
**Solution**

Recall

If we multiply both sides of equation 4.2 by , we get

By subtracting from both sides, we get

Thus, as shown above, both equations are equivalent (⬄).

Question 2: This is question 13 extracted from page 196 of ISLP (James, Witten, Hastie, Tibshirani, & Taylor, 2023) and is displayed in the screenshot below.

A screenshot of a cell phone

AI-generated content may be incorrect.

Solution

I began my analysis by exploring the data set in Microsoft Excel and then computed some summary statistics in the jupyter notebook attached to this file also hosted on my Github repository https://github.com/ndesamuelmbah/DDS-8555/tree/main/AssignmentFiles/Week5 (Nde, 2024). Figure 1 below shows a pair plot of the variables in that dataset colored by direction.  
**Figure 1**  
Pairplot of variables in the dataset colored by market direction.

A graph of a diagram

AI-generated content may be incorrect.  
Note. *Notice that the most noticeable predictor of the Direction of the market is the current Today, the current direction of the market. The volume by year also shows an exponential growth which speaks to the fact that more and more companies are listed on the market daily. Also note the presence of* ***outlier*** *spread around each scatterplot.*

A summary of the first model predicting the direction from the five lag variables and the volume with a logistic regression model is shown in Figure 2 below.  
**Figure 2**  
*Model summary for logistic regression using the lag variables and volume to predict market Direction and its confusion matrix.*

A table of numbers and letters

AI-generated content may be incorrect. A chart of a different color

AI-generated content may be incorrect.  
Note. *Left: Model summary showing indicating that only one predictor (Lag2) is reported as statistically significant in the model. Right: Confusion matrix showing a high number of false positives (Predicted market up 1 when in reallity, the market was down).*

To improve the model, I looped through all the possible set of variables from the initial 5 variables, creating models and tracking the features that improved the precision score. I used precision as the score for this model because it is important to minimize false positives in a situation like this where financial stake are involved. The table in figure 3 below shows the features that improved the scores sorted in by the score in descending order. It is important to note that even though the variable Today proved to have a clear division between the market Directions, I excluded it from my predictors because is typically not available for prediction until after the fact.

Table 1  
Feature sets that improved logistic regression model with the precision scores  
A screenshot of a computer

AI-generated content may be incorrect.  
**Note.** *These are not the only models that were scored. These are the models that improved the precision score. There is a relatively small gain in precision in using the best (4 variable) model versus using the Lag2 (single variable) model. This small gain reflects the difficulty in predicting the market. It may also mean the logistic regression model is not the best for the dataset.*

I used a similar approach for k nearest neighbors (KNN) classifier, changing the number of neighbors (k) from 1 to 12 and looping through the unique feature sets, saving the best feature set for each k. The results printed in Figure 4 below shows that the best feature set for each k remained the same. This makes sense because the KNN classifier works by measuring distances between points and the points are static in space (Jain, Kumar, & Roy).

**Figure 4**  
*Code snippet looping through features looking for the most predictive feature set for each k.*  
A screenshot of a computer program

AI-generated content may be incorrect.  
Note: *Each k from 1 to 12 produced the same optimal features with different precision score. The best score occurs at k=2 but this value of k is too small and the model with k=2 may be unstable as it would be very sensitive to outliers. k=4 provides an alternative option that may be less sensitive to outliers and therefore more stable.*

Section 2: Participating in Multi-Class Prediction of Obesity Risk Kaggle Competition (Reade & Chow, Multi-Class Prediction of Obesity Risk, 2024). I fitted 5 classification models on the dataset using the code hosted in my GitHub repository https://github.com/ndesamuelmbah/DDS-8555/tree/main/AssignmentFiles/Week5 (Nde, 2024). Figure 5 below shows the score of each of the models on a test dataset.   
F**igure 5**  
Code snippet showing model results for multiclass classification of Risk of Obesity.  
A screenshot of a computer code

AI-generated content may be incorrect.  
**Note.** *I choose recall to be the primary metric for evaluating these models because the priority in a medical context like this is to identify all positive cases so that they could be treated.   
Assumptions and Interpretation of Results.*

Since recall is the primary metric of interest, SVM (0.8507) emerges as the best-performing model. This means SVM correctly identifies more cases of obesity risk than any other model, making it ideal if missing positive cases is costly.

1. SVM (Overall Best, Recall = 0.8507)
   * SVM performs well when classes are separable, even in high-dimensional spaces.
   * The linear kernel assumption might work well, but non-linear kernels (RBF, polynomial) could be worth testing.
   * Assumes data is somewhat balanced though SVM can handle imbalanced classes with proper weighting.
2. Logistic Regression (Recall = 0.8423)
   * A close second! Logistic regression assumes linear relationships between predictors and log-odds.
   * Works best when there’s no multicollinearity and predictors are independent.
   * Less flexible than SVM but more interpretable coefficients provide insight into feature importance.
3. QDA (0.8112) > LDA (0.7935)
   * QDA outperforms LDA, suggesting quadratic decision boundaries work better than linear ones.
   * LDA assumes equal covariance among classes, while QDA relaxes this assumption hence, QDA's superior recall.
   * Works well for normally distributed features, which seems like the case from the graph of the density plots of the dataset (diagonal of pair plots).
4. Naïve Bayes (Lowest Recall - 0.6506)
   * Performs worst, likely due to its strong independence assumption between predictors (James, Witten, Hastie, Tibshirani, & Taylor, 2023).
   * This assumption often doesn’t hold in real-world datasets, leading to high bias and poor recall in this case.
   * While it’s fast and scalable, its oversimplification fails to capture complex relationships in obesity risk.

Final Takeaway

* If recall is critical as I have assumed (missing an obesity risk case is costly), go with SVM.
* If you want an interpretable alternative, Logistic Regression isn’t far behind.
* QDA may work better than LDA if feature interactions matter.
* Avoid Naïve Bayes unless you’re looking for speed over accuracy.

Next Steps?

* Fine-tune SVM hyperparameters (kernel choice, C parameter).
* Check for transformation of variable to into normal distribution to improve LDA model.
* Consider ensemble methods to boost recall further!
* Add more features to the model – especially the many categorical features that were not used such as smoke status which is a risk factor of type 2 diabetes (D, et al., 2019).

# Bibliography

D, C., A, A., A, D. P., C, R., Calogero AE, P. F., & R., P. (2019). Smoking and diabetes: dangerous liaisons and confusing relationships. *National Library of Medicine*. doi:10.1186/s13098-019-0482-2

Jain, S., Kumar, S., & Roy, N. (n.d.). *Futuristic Trends in Network and Communication Technologies.* doi:10.1007/978-981-16-1480-4\_18

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning: with Applications in Python.* Springer Nature.

Nde, S. M. (2024, 02). *DDS-8555*. Retrieved from Github: https://github.com/ndesamuelmbah/DDS-8555.git

Reade, W., & Chow, A. (2024). *Multi-Class Prediction of Obesity Risk*. Retrieved from Kaggle: https://kaggle.com/competitions/playground-series-s4e2

Reade, W., & Chow, A. (2024). *Regression with an Abalone Dataset*. Retrieved from Kaggle: https://kaggle.com/competitions/playground-series-s4e4