**Assignment 5: Build and Evaluate Classification Models**

Andrew L.J. Rezendes

School of Technology and Engineering, National University

DDS-8550: Predictive Modeling

Dr. Alexander Watt

June 29, 2025

**Assignment 5: Build and Evaluate Classification Models**

Four multiclass classification models were developed and evaluated on their ability to predict obesity risk using the public dataset provided for the Kaggle competition "Multi-class Prediction of Obesity Risk." The classification models implemented were multinomial logistic regression, linear discriminant analysis (LDA), Gaussian naïve Bayes, and support vector machines (SVMs). These models were selected based on their foundational roles in supervised learning and their interpretability, generalizability, and relevance to health-related prediction tasks (James et al., 2021).

Data preprocessing included label encoding of the target variable representing obesity risk, nobeyesdad, one-hot encoding of categorical predictors, and standardization using z-score normalization. Feature scaling is a critical requirement for distance-based classifiers such as SVMs and for ensuring convergence in gradient-based optimization algorithms like those used in logistic regression (Kuhn & Johnson, 2013). This was applied across all numeric features. Every categorical feature was one-hot encoded, including variables such as gender, family history of obesity, transportation modes, and dietary habits. One-hot encoding was chosen over ordinal encoding to avoid introducing spurious ordinal relationships among nominal features (Chollet & Allaire, 2018).

The best-performing model was multinomial logistic regression, which achieved an accuracy of 86.8% and a macro-averaged F1 score of 0.85 on the validation set. This model demonstrated balanced predictive performance across all seven classes, particularly excelling in identifying severe obesity classes without sacrificing accuracy in borderline categories. Logistic regression's probabilistic nature and robust performance with standardized numeric inputs make it well-suited to structured health datasets (Hosmer et al., 2013).

The support vector machine classifier achieved comparable results, with an accuracy of 85.7% and a macro F1 of 0.84. SVMs are known for their effectiveness in high-dimensional spaces and their ability to model nonlinear decision boundaries via kernel functions (Cortes & Vapnik, 1995). However, the classifier’s performance degraded slightly in the overlapping overweight classes, which may be due to sensitivity to outlier distributions and limited interpretability compared to logistic regression.

The linear discriminant analysis model yielded an accuracy of 82.3%, showing strong separation in the more clearly defined obesity classes. However, its assumption of multivariate normality and equal class covariances likely contributed to performance losses in the more ambiguous categories, such as Normal\_Weight and Overweight\_Level\_I (James et al., 2021). LDA remains computationally efficient and interpretable, making it a reasonable baseline for multiclass prediction tasks.

The Gaussian naïve Bayes model underperformed relative to the others, achieving only 60.4% accuracy. Its underlying assumption of conditional independence among features was likely violated, which often results in miscalibration and poor generalization in health and behavioral datasets where predictor interactions are common (Zhang, 2004).

In conclusion, multinomial logistic regression was chosen as the best performing model due to its strong, balanced performance, interpretability, and alignment with the statistical structure of the dataset. SVM offered comparable accuracy but lacked transparency and was computationally more intensive. These results reinforce the importance of aligning model assumptions with data characteristics and choosing models that balance generalization with interpretability—especially in health-related classification tasks.

References

Chollet, F., & Allaire, J. J. (2018). *Deep learning with R*. Manning Publications.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. <https://doi.org/10.1007/BF00994018>

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118548387

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. <https://doi.org/10.1007/978-1-4614-6849-3>

Zhang, H. (2004). The optimality of naïve Bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference* (pp. 562–567). AAAI Press.

**Appendix**A screenshot of a web page

AI-generated content may be incorrect.

**Proof of Submission**

A screenshot of a computer

AI-generated content may be incorrect.