**DDS8555 Assignment 5: Multi-Class Prediction of Obesity Risk**

Eliseo Ilarraza

National University

DDS-8555 Predictive Analysis

October 26, 2025

**Methods**

We framed the task as a seven-class prediction of NObeyesdad from lifestyle and anthropometric features (e.g., Gender, Age, Height, Weight, FAVC, FCVC, NCP, CAEC, SMOKE, CH2O, SCC, FAF, TUE, CALC, MTRANS). Four models were fit with a shared, leakage-safe pipeline: multinomial logistic regression, linear discriminant analysis (LDA) with Ledoit–Wolf shrinkage, Gaussian naïve Bayes, and a linear one‑vs‑rest SVM. Quantitative features were standardized; categoricals were one‑hot encoded; ordinal scales (FCVC, NCP, FAF, TUE) were treated as ordered. Hyperparameters were lightly tuned within stratified 3‑fold cross‑validation (CV), with all preprocessing performed inside folds to prevent leakage (Leinonen et al., 2024). Logistic regression supplied calibrated probabilities on the log‑odds scale (Hua et al., 2025); LDA assumed approximately equal class covariances (Qu & Pei, 2024); NB provided a fast baseline for mixed tabular signals (Peretz et al., 2024); SVM used feature scaling and regularization tuned for minority‑class recall (Guido et al., 2024).

**Validation and Diagnostics**

All models used identical stratified CV splits. We reported accuracy and macro‑F1 and monitored class‑wise precision/recall on out‑of‑fold predictions. Because accuracy can be misleading under imbalance, we emphasized macro‑F1 and noted the IMCP framework as an imbalance‑insensitive multiclass complement for future iterations (Aguilar‑Ruiz & Michalak, 2024). After validation, the pipelines were refit on the complete training data to produce four Kaggle submissions.

**Results**

Cross‑validation ranked models: Multinomial Logistic Regression ≻ , LDA ≻ Linear SVM ≻ , Gaussian NB. Kaggle (late submission) confirmed this ordering: Logistic Regression (Private 0.86054; Public 0.86596), LDA (0.81286; 0.81466), Linear SVM (0.75108; 0.75000), and Gaussian NB (0.58535; 0.58887). A close private/public agreement suggests good generalization.

**Interpretation, Assumptions, and Applications**

Logistic coefficients map to odds‑ratio changes and support actionable counseling (Hua et al., 2025). LDA/QDA choice depends on covariance similarity; our diagnostics favored LDA’s equal‑covariance premise (Qu & Pei, 2024). Naïve Bayes’ conditional independence is an approximation but remains effective for high‑dimensional screening (Peretz et al., 2024). For deployment, preserve scaling, consider class weights for minority tiers, and audit fairness across gender and age—comparable obesity‑risk systems couple modeling with individualized feedback and feature attributions (Du et al., 2024).

**Deliverables**

Repository and evidence: GitHub (code, notebook, figures, submissions): https://github.com/Skeonpr1/dds8555-a5-obesity-risk; Kaggle submissions page: https://www.kaggle.com/competitions/playground-series-s4e2/submissions#.

**References**

Aguilar‑Ruiz, J. S., & Michalak, M. (2024). Classification performance assessment for imbalanced multiclass data. *Scientific Reports*, *14*, 10759. https://doi.org/10.1038/s41598-024-61365-z

Du, J., Yang, S., Zeng, Y., Ye, C., Chang, X., & Wu, S. (2024). Visualization obesity risk prediction system based on machine learning. *Scientific Reports*, *14*, 22424. https://doi.org/10.1038/s41598-024-73826-6

Guido, R., et al. (2024). An overview on the advancements of support vector machine models in healthcare applications: A review. *Information*, *15(4)*, 235. https://doi.org/10.3390/info15040235

Hua, Y., Stead, T. S., George, A., & Ganti, L. (2025). Clinical risk prediction with logistic regression: Best practices, validation techniques, and applications in medical research. *Academic Medicine & Surgery*. https://doi.org/10.62186/001c.131964

Leinonen, T., Wong, D., Vasankari, A., Wahab, A., Nadarajah, R., Kaisti, M., & Airola, A. (2024). Empirical investigation of multi-source cross-validation in clinical ECG classification. *Computers in Biology and Medicine*, *183*, 109271. https://doi.org/10.1016/j.compbiomed.2024.109271

Peretz, O., Koren, M., & Koren, O. (2024). Naive Bayes classifier—An ensemble procedure for recall and precision enrichment. *Engineering Applications of Artificial Intelligence*, *136*, 108972. https://doi.org/10.1016/j.engappai.2024.108972

Qu, L., & Pei, Y. (2024). A comprehensive review on discriminant analysis for addressing challenges of class-level limitations, small sample size, and robustness. *Processes*, *12(7)*, 1382. https://doi.org/10.3390/pr12071382

**Appendix: Evidence and Figures**

Kaggle late‑submission results. Private/Public scores, respectively, were: Logistic Regression 0.86054/0.86596; LDA 0.81286/0.81466; Linear SVM 0.75108/0.75000; Gaussian NB 0.58535/0.58887. These align with cross‑validation ranking and indicate that the multinomial logistic regression model generalizes best.

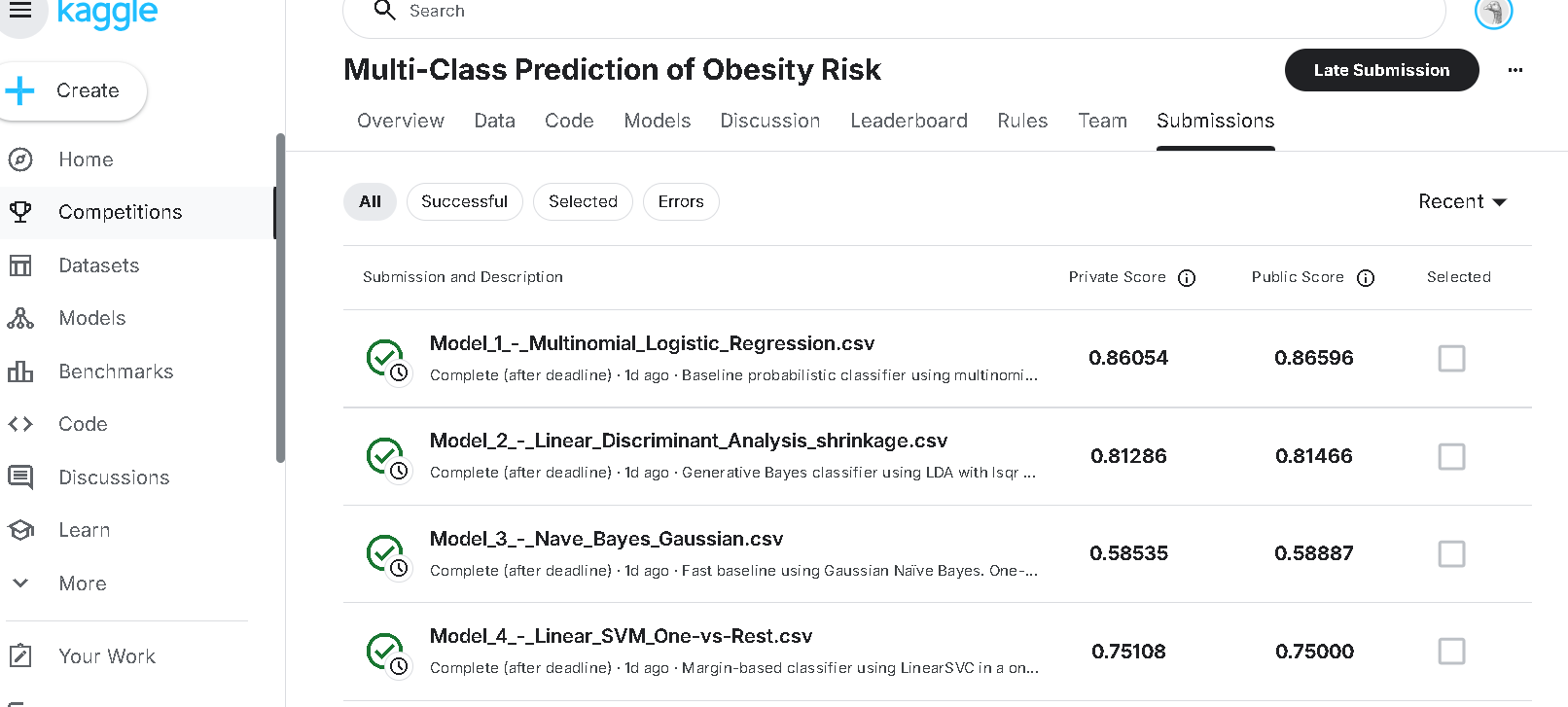


Figure A. Kaggle late‑submission page with private and public scores for Models 1–4.

Cross‑validation accuracy across the four models. Logistic Regression achieves the highest mean accuracy, followed by LDA, Linear SVM, and Gaussian NB. The separation between the first two models and the rest suggests that linear‑probabilistic methods fit the tabular signal well.

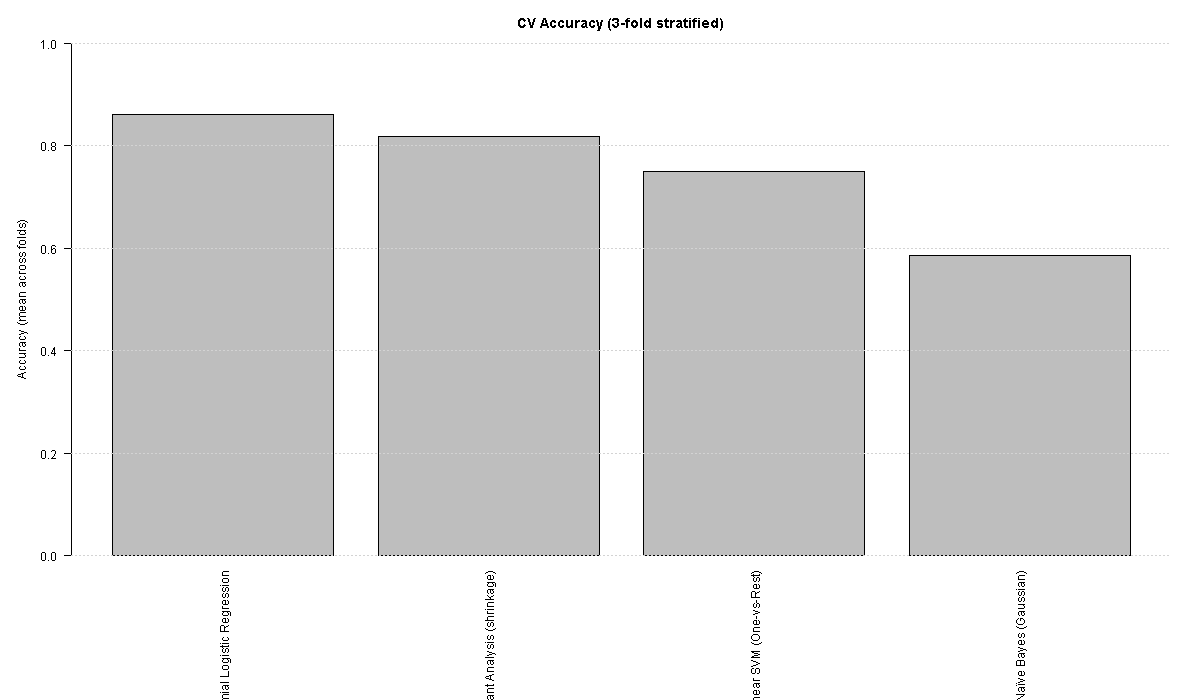


Figure B. CV Accuracy (3‑fold stratified) — base R chart.

The ggplot version orders models by accuracy to emphasize ranking and margin. The logistic model’s advantage over LDA is modest but consistent across folds, while SVM and NB trail notably.

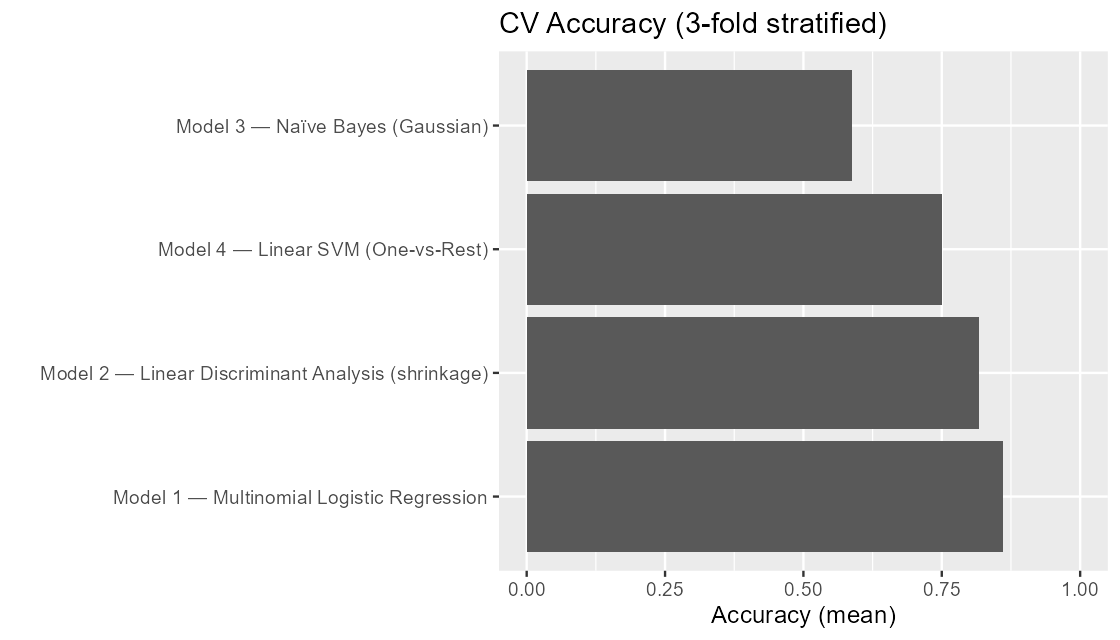


Figure C. CV Accuracy (3‑fold stratified) — ggplot chart.

Macro‑F1 balances performance across the seven classes. Logistic Regression again leads, indicating its advantage is not due solely to the majority classes. LDA’s macro‑F1 is close, whereas SVM and NB lose recall on minority obesity categories.

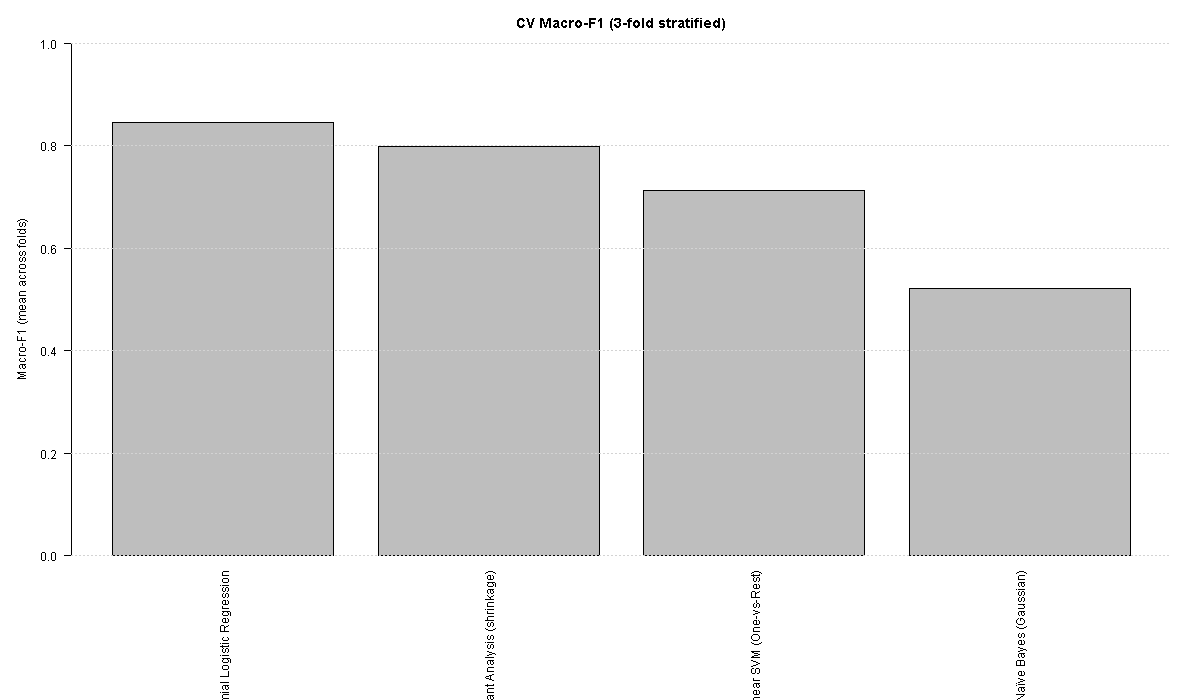


Figure D. CV Macro‑F1 (3‑fold stratified) — base R chart.

Ordering by macro‑F1 clarifies the gap between probabilistic (Logistic/LDA) and margin‑based (SVM) methods on balanced class performance. This corroborates the decision to report Logistic Regression as the primary model.

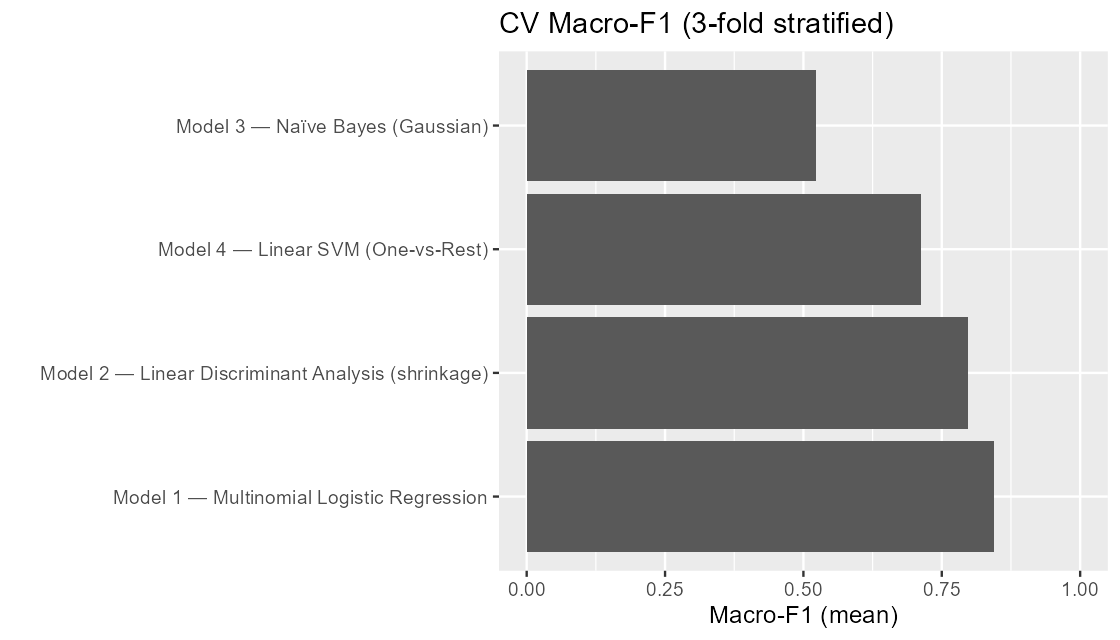


Figure E. CV Macro‑F1 (3‑fold stratified) — ggplot chart.

**Figure E (CV Macro-F1, ggplot).** Ordering models by macro-F1 shows **Multinomial Logistic Regression** leading with the best class-balanced performance, with **LDA (shrinkage)** a close second. **Linear SVM** trails those probabilistic linear models—consistent with some recall loss on minority obesity classes—while **Gaussian Naïve Bayes** performs worst, likely due to its conditional-independence assumption on mixed, correlated features. The small gap between Logistic Regression and LDA suggests that both capture the tabular signal well, but LR remains the strongest overall when performance is equalized across all seven classes.