RQ: *Which machine learning method best predicts next-day stock price direction, and does any model outperform logistic regression?*

1. Abstract (½ page)

- Problem: predict next-day stock price direction
- Methods: logistic regression, LASSO, GAM, KNN, decision tree, random forest
- Data: one stock from 2015–2024
- Metrics: AUC, accuracy
- Main result: which model performs best
- Contribution: compare linear vs nonlinear vs tree models

2. Introduction (1 page)

- Why predicting stock direction is interesting
- Formulate the research question
- State that you compare multiple ML methods taught in the course
- Outline structure of the paper

3. Data (1 page)

- Source: Huge Stock Market Dataset (filtered to one stock)
- Variables used: Close, High, Low, Volume
- Engineered features:
    - Lagged returns
    - MA5, MA10
    - Volatility10
    - Momentum10
    - High–Low range
- Target variable: UpTomorrow = 1 if next day's return > 0
- Basic descriptive statistics

4. Methodology (2 pages)

4.1 Models (brief explanations)

- Logistic regression (baseline)
- LASSO (feature selection)
- GAM (nonlinear effects)
- KNN (memorization method)
- Decision tree
- Random forest (ensemble method)

4.2 Train/Test Split

- Train: 2015–2021
- Test: 2022–2024
- No shuffling (time order respected)

4.3 Evaluation Metrics

- Accuracy
- Confusion matrix
- ROC + AUC
- Baseline benchmark = majority class

5. Results (2 pages)

5.1 Performance Table

One table comparing AUC + accuracy for all models.

5.2 Key Visuals

- ROC curves (one combined plot)
- Random forest variable importance
- (Optional) 1–2 GAM spline plots

6. Discussion (1.5 pages)

- Which model performed best and why
- What features mattered most
- Linear vs nonlinear vs tree-based differences

- Limitations:
    - One stock
    - Daily data only
    - Low predictability of markets
- Practical meaning of weak/moderate predictability

7. Conclusion (½ page)

- Answer RQ clearly
- Summarize findings
- Note ML gives small but detectable improvements
- Suggest further work (more assets, more features, regime-based analysis)

Appendix

- All R code
- Extended tables and plots
- Additional diagnostics