CSC311: Introduction to Machine Learning

Project Report

Optional Creative Project Title

Submitted by

Student 1, Student 2, Student 3, Student 4

University of Toronto
Fall 2025

1 Executive Summary

This section should be at most 1/4 page (1 paragraph).

- Summarize the model families you explored (at least three).
- State your best-performing model and its projected accuracy.
- Provide a high-level justification of why this model outperformed the others.

2 Data Exploration

This section should be at most 1 page.

Why is data exploration important in a machine learning project?

- Understand the dataset by identifying feature types (numerical, categorical, text), distributions, and class balance.
- Detect issues by finding missing values, outliers, or inconsistent entries that could harm model training.
- Inform pre-processing by deciding which transformations are needed (e.g., normalization, encoding, text vectorization).
- Guide model choice by identifying patterns (e.g., linear vs. nonlinear relationships) that suggest suitable algorithms.
- Generate insights by developing intuition about which features may be most predictive or relevant.

What you should write in this section:

- Summarize the dataset by describing the main feature types, distributions, and class balance.
- Identify issues you found such as missing values, outliers, or inconsistencies.
- Describe pre-processing by explaining what transformations you applied and why.
- Explain data splitting by describing how you reserved a portion of the data as a test set. This test set should **not** be used during data exploration.
- Use figures when helpful by including plots or tables and explaining what they reveal.
- Connect findings to model choices by highlighting patterns that influenced your selection of models.
- Share key insights by presenting observations about especially relevant or predictive features.

Note: Each student contributes three data points (one per class). These three points are closely related, so if they are split across training, validation, and test sets, information from one could leak into another set. To avoid this data leakage, all three points from the same student must remain together in the same split.

3 Methodology

This section should be at most 3 pages.

What you should write in this section:

- Present the three model families you used and explain why they are appropriate for the dataset.
- Describe the optimization techniques by specifying the optimizer (e.g., SGD, Adam), learning rate schedule if applicable, and any regularization or early stopping strategies.
- Explain the validation method by describing how you split the data into training and validation sets, or how you applied cross-validation. Summarize your hyperparameter tuning process by stating which hyperparameters you varied, how you chose values to try, and the evaluation metrics you used to compare them. Provide evidence that your hyperparameter choices are reasonable, noting that an exhaustive search is not required.
- Describe the evaluation metrics by explaining which metrics you used to evaluate your models. Use at least one additional metric (precision, recall, F1, etc.) beyond accuracy, and justify your choices.
- Provide implementation details by briefly describing the libraries or frameworks you used, and any custom code you wrote.
- (Optional) Describe any additional techniques you tried, such as feature engineering, clustering, or PCA.

Common mistakes:

- Using the test set during training, feature engineering, or hyperparameter tuning.
- Tuning hyperparameters without a validation set that is separate from the training and test sets.
- Omitting important training details in the report, making results hard to reproduce.
- Jumping straight to deep networks when a simpler model might be sufficient and more interpretable.
- Choosing the best model based only on default hyperparameters, then tuning only that model. This is unfair because other models might improve significantly with tuning as well. **This is the most common mistake we observed while grading the reports in winter 2025.**
- Presenting tables or plots without labels, units, or explanations.
- Reporting only accuracy and ignoring other evaluation metrics such as precision, recall, or F1.

4 Results

This section should be at most 1 page.

What you should write in this section:

- Report the results of your models using the evaluation metrics you described in the Methodology section, and present them clearly in tables or plots.
- Analyze errors by describing common misclassifications or weaknesses, and include a confusion matrix or illustrative examples if helpful.
- Compare models by summarizing performance across model families, and clearly state your final model choice as implemented in your prediction script.
- Estimate test performance by **providing a single-number estimate of your best model's accuracy on the unseen test set**. Providing a range will earn **no** marks.
- Justify your performance estimate by supporting it with empirical evidence such as validation results, cross-validation stability, learning curves, or consistency across seeds/folds.

5 Contributions and Learning

This section should contain a list of 3–4 bullets, one for each student. Each student should write at most three sentences describing their main contributions and the most important lesson learned.

6 References

You should comment out this section for the final report.

Use references whenever appropriate (e.g., when you consult textbooks, research papers, online tutorials, or external code). Cite in the text using `\cite{}`. For example, let's cite the two references provided in the `.bib` file [1, 2].

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.