**Machine Breakdown Predictions**

**Harrison Howze, Sam Schmall, Omar Abdelsalam, Daniel Byerly**

1. **Problem Statement and Background (15%)**

Problem: Minimize costs and downtime of a business by predicting machine failures and the factors that contribute to those failures.

Where: We found the initial dataset from Kaggle. The data includes rotational speed, air and process temp, torque of the machines, tool wear in minutes. There are 2 target variables, whether the machine fails and what specific failure type it is if it does fail.

Informal success measures: accuracy

Who cares: Business owners, machine operators and maintenance.

What impact: Financial impact, reduced downtime.

Implications: More data tailored to specific machines.

Related work: Predictive software analysis.

*[Give a clear and complete statement of the problem. (Do NOT describe data, methods or tools yet – see below.) Where does the data come from, what are its characteristics? Include informal success measures (e.g. accuracy on cross-validated data, without specifying ROC or precision/recall, etc.) that you plan to use. Include background material as appropriate: who cares about this problem, what impact it has, what implications better solutions might have. Included a brief summary of any related work you know about.]*

1. **Data and Exploratory Analysis (15%)**

We have potentially unnecessary columns, the UID and product ID which allow the algorithm to learn which items in the training set always fail. We also are not going to use failure type, as our focus is predicting whether the machine will fail or not regardless of type. Our focus at the start is binary classification, failure or not. There is no missing data, tool wear has a large amount of variance but is standardized.

R-code: TODO

*[Describe the data set you will be using. Discuss anything you had to do clean the data and why. Describe what tools and R code you used to extract, clean, and generate the data for your experiments. Some potential questions of the data might be: any anomalies or outliers? – no. Did you need to impute any of the data in order to get it to work for any proposed algorithms? – no.]*

1. **Methods (10%)**

*[Describe the methods you explored (usually algorithms, or data cleaning or data wrangling approaches). Justify your methods in terms of the problem statement. What did you consider but \*not\* use? In particular, be sure to include every method you tried, even if it didn't "work". When describing methods that didn't work, make clear how they failed and any evaluation metrics you used to decide so.]*

1. **Tools (10%)**

*[Describe the tools that you used and the reasons for their choice. Justify them in terms of the problem itself and the methods you want to use. Tools will probably include machine learning, and possibly data wrangling and visualization. Please discuss all of them. How did you employ them? What features worked well and what didn't? What could be improved? Describe any tools that you tried and ended up not using. What was the problem?]*

1. **Results (35%)**

*[Give a detailed summary of the results of your work. Here is where you specify the exact performance measures you used. Usually there will be some kind of accuracy or quality measure. There may also be a performance (runtime or throughput) measure. Please use visualizations whenever possible. Include links to interactive visualizations if you built them. You should attempt to evaluate a primary model and in addition a "baseline" model. The baseline is typically the simplest model that's applicable to that data problem, e.g. Naive Bayes for classification, or K-means on raw feature data for clustering. If there isn't a plausible automatic baseline model, you can e.g. compare with human performance by having someone hand-solve your problem on a small subset of data. You won’t expect to achieve this level of performance, but it establishes a scale by which to measure your project's performance. Compare the performance of your baseline model and primary model and explain the differences.]*

1. **Summary and Conclusions (10%)**

*[In this section give a high-level summary of your results. If the reader only reads one section of the report, this one should be it, and it should be self-contained. You can refer back to the "Results" section for elaborations. This section should be less than a page. In particular, emphasize any results that were surprising.]*

1. **Appendix (5%)**

Include the link to your github/gitlab repository (that I can access) containing your R programs/scripts, and link to the data.