Introduction to LaTeX

Lulu Allen
CS 800 - Research Methods

Current draft: 2/16/26 at 7:08am EDT

About Me: I am a US Navy veteran and a Ph.D student specializing in Computer Engineering with a focus in Cybersecurity at Old Dominion University. My career aspirations include becoming a Cybersecurity Architect or a Software Engineer.

1 URIs

Below are clickable links to my pages:

`https://www.cs.odu.edu/~lalle003/`

`https://github.com/lalle003`

2 Images

All figures must have a caption and must be referenced in the text. See the example below.

Figure 1 shows an original PNG with no scaling or cropping. The original dimensions are 400 x 532 (or, 2in x 2.66in). Figure 2 shows an example of cropping the image using the `trim, clip` options to `include graphics`.

3 Quotation Marks

Within my recent research, I focused on hard drive failures to incorporate predictive maintenance by addressing challenges such as temporal drift and class imbalance and adaptive models like Random Forest and XGBoost. The XGBoost modeling results demonstrated that machine learning methods can learn useful patterns from these complex signals, despite noise and missing values.

4 Tables

Table 1 shows a simple example table. Table 2 shows an example of my education and professional experience. This employs rows that span multiple columns (multicol) and columns that span multiple rows (multirow).

Table 1: Academic Career

Education	**Institution**	**Achievements**
Ph.D Computer Engineering	ODU	Expected (12/2026)
Masters' in Cybersecurity	ODU	Graduated
DevOps Bootcamp	Intellectual P.T.	Sec+, Splunk Core

Figure 1: Original PNG

Table 2: Experience in Years

		Training	
		Licenses	Education
Positions	Aviation Mechanic	5	1
	Technical Logistics	3	6

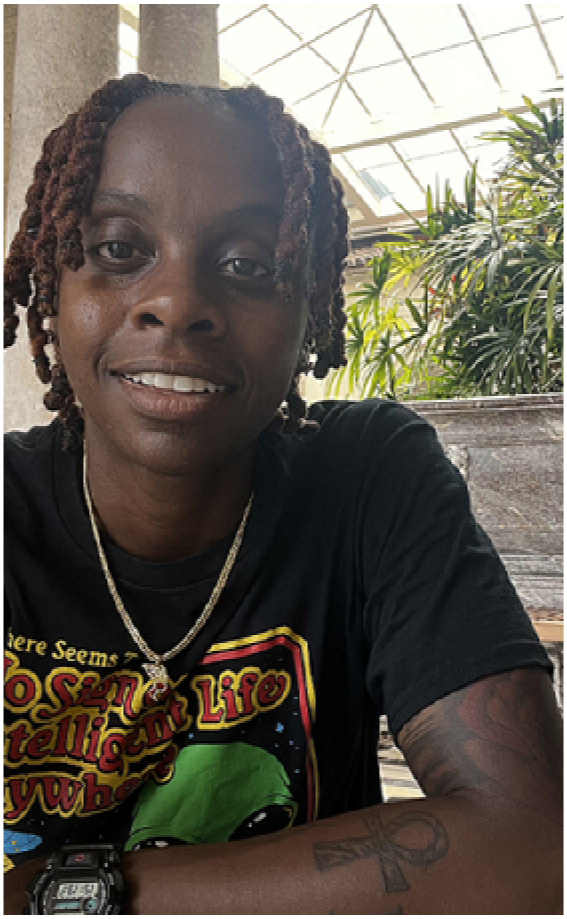
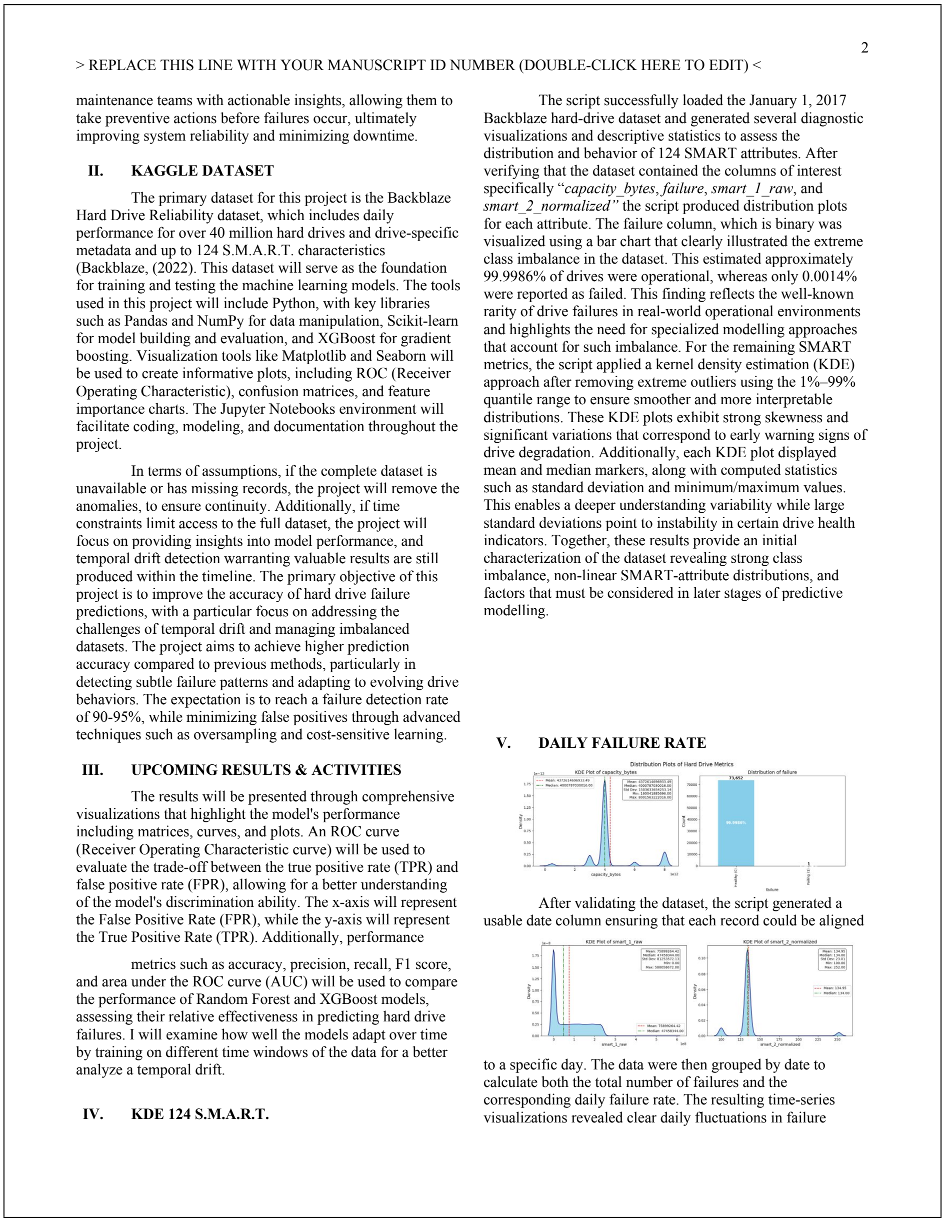


Figure 2: Cropped PNG - 0.25in from left, 0.5in from bottom, 1in from right, 0.3in from top



maintenance teams with actionable insights, allowing them to take preventive actions before failures occur, ultimately improving system reliability and minimizing downtime.

II. KAGGLE DATASET

The primary dataset for this project is the Backblaze Hard Drive Reliability dataset, which includes daily performance for over 40 million hard drives and drive-specific metadata and up to 124 S.M.A.R.T. characteristics (Backblaze, (2022). This dataset will serve as the foundation for training and testing the machine learning models. The tools used in this project will include Python, with key libraries such as Pandas and NumPy for data manipulation, Scikit-learn for model building and evaluation, and XGBoost for gradient boosting. Visualization tools like Matplotlib and Seaborn will be used to create informative plots, including ROC (Receiver Operating Characteristic), confusion matrices, and feature importance charts. The Jupyter Notebooks environment will facilitate coding, modeling, and documentation throughout the project.

In terms of assumptions, if the complete dataset is unavailable or has missing records, the project will remove the anomalies, to ensure continuity. Additionally, if time constraints limit access to the full dataset, the project will focus on providing insights into model performance, and temporal drift detection warranting valuable results are still produced within the timeline. The primary objective of this project is to improve the accuracy of hard drive failure predictions, with a particular focus on addressing the challenges of temporal drift and managing imbalanced datasets. The project aims to achieve higher prediction accuracy compared to previous methods, particularly in detecting subtle failure patterns and adapting to evolving drive behaviors. The expectation is to reach a failure detection rate of 90-95%, while minimizing false positives through advanced techniques such as oversampling and cost-sensitive learning.

III. UPCOMING RESULTS & ACTIVITIES

The results will be presented through comprehensive visualizations that highlight the model's performance including matrices, curves, and plots. An ROC curve (Receiver Operating Characteristic curve) will be used to evaluate the trade-off between the true positive rate (TPR) and false positive rate (FPR), allowing for a better understanding of the model's discrimination ability. The x-axis will represent the False Positive Rate (FPR), while the y-axis will represent the True Positive Rate (TPR). Additionally, performance

metrics such as accuracy, precision, recall, F1 score, and area under the ROC curve (AUC) will be used to compare the performance of Random Forest and XGBoost models, assessing their relative effectiveness in predicting hard drive failures. I will examine how well the models adapt over time by training on different time windows of the data for a better analyze a temporal drift.

IV. KDE 124 S.M.A.R.T.

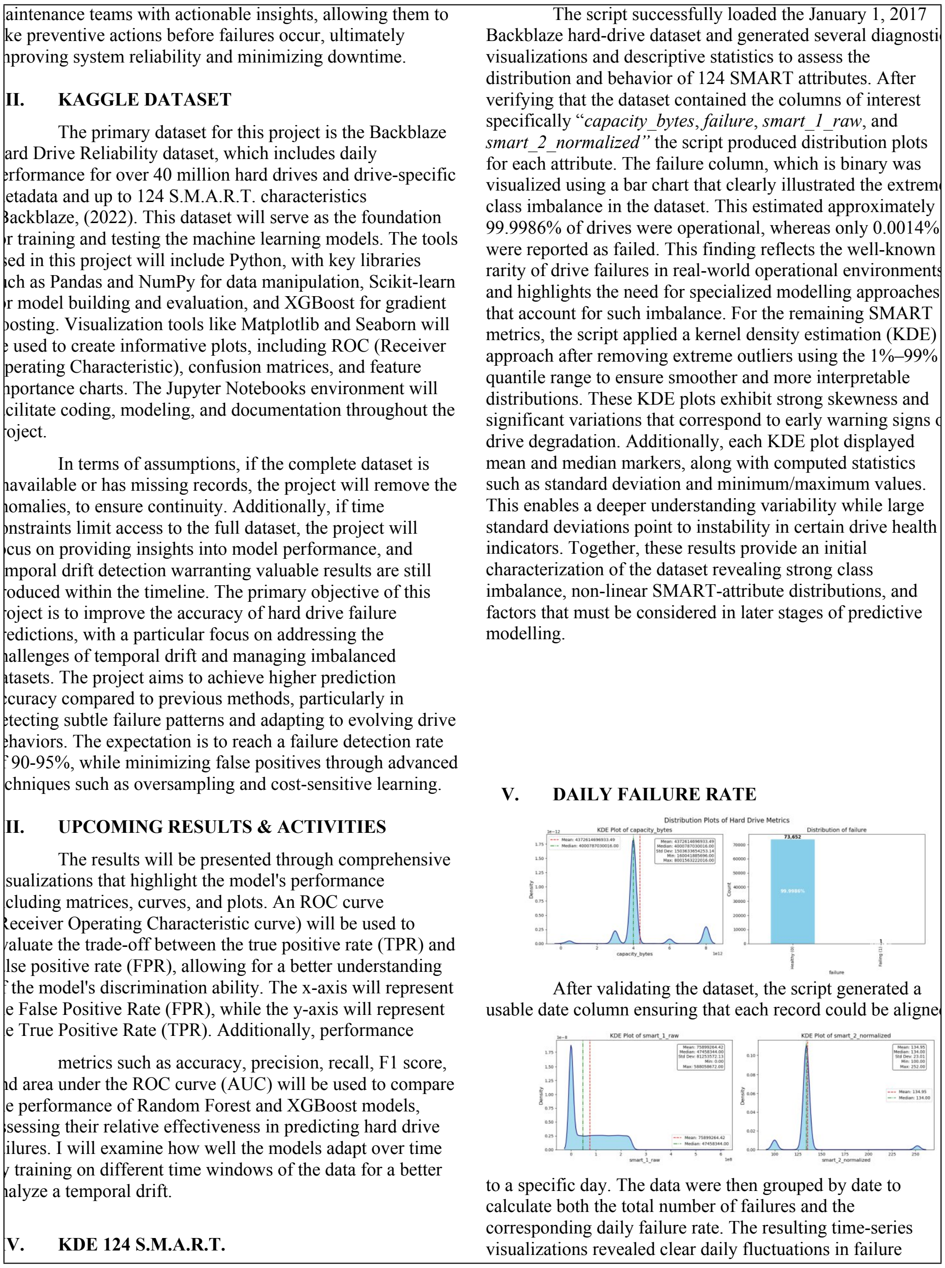
The script successfully loaded the January 1, 2017 Backblaze hard-drive dataset and generated several diagnostic visualizations and descriptive statistics to assess the distribution and behavior of 124 SMART attributes. After verifying that the dataset contained the columns of interest specifically "*capacity_bytes*, *failure*, *smart_1_raw*, and *smart_2_normalized*" the script produced distribution plots for each attribute. The failure column, which is binary was visualized using a bar chart that clearly illustrated the extreme class imbalance in the dataset. This estimated approximately 99.9986% of drives were operational, whereas only 0.0014% were reported as failed. This finding reflects the well-known rarity of drive failures in real-world operational environments and highlights the need for specialized modelling approaches that account for such imbalance. For the remaining SMART metrics, the script applied a kernel density estimation (KDE) approach after removing extreme outliers using the 1%–99% quantile range to ensure smoother and more interpretable distributions. These KDE plots exhibit strong skewness and significant variations that correspond to early warning signs of drive degradation. Additionally, each KDE plot displayed mean and median markers, along with computed statistics such as standard deviation and minimum/maximum values. This enables a deeper understanding variability while large standard deviations point to instability in certain drive health indicators. Together, these results provide an initial characterization of the dataset revealing strong class imbalance, non-linear SMART-attribute distributions, and factors that must be considered in later stages of predictive modelling.

V. DAILY FAILURE RATE

After validating the dataset, the script generated a usable date column ensuring that each record could be aligned

to a specific day. The data were then grouped by date to calculate both the total number of failures and the corresponding daily failure rate. The resulting time-series visualizations revealed clear daily fluctuations in failure

Figure 3: Inserted PDF



aintenance teams with actionable insights, allowing them to ke preventive actions before failures occur, ultimately nproving system reliability and minimizing downtime.

II. KAGGLE DATASET

The primary dataset for this project is the Backblaze ard Drive Reliability dataset, which includes daily erformance for over 40 million hard drives and drive-specific etadata and up to 124 S.M.A.R.T. characteristics 3ackblaze, (2022). This dataset will serve as the foundation r training and testing the machine learning models. The tools sed in this project will include Python, with key libraries ich as Pandas and NumPy for data manipulation, Scikit-learn r model building and evaluation, and XGBoost for gradient posting. Visualization tools like Matplotlib and Seaborn will e used to create informative plots, including ROC (Receiver perating Characteristic), confusion matrices, and feature nportance charts. The Jupyter Notebooks environment will cilitate coding, modeling, and documentation throughout the oject.

In terms of assumptions, if the complete dataset is navailable or has missing records, the project will remove the nomalies, to ensure continuity. Additionally, if time onstraints limit access to the full dataset, the project will ocus on providing insights into model performance, and mporal drift detection warranting valuable results are still oduced within the timeline. The primary objective of this oject is to improve the accuracy of hard drive failure edictions, with a particular focus on addressing the hallenges of temporal drift and managing imbalanced atasets. The project aims to achieve higher prediction ccuracy compared to previous methods, particularly in etecting subtle failure patterns and adapting to evolving drive ehaviors. The expectation is to reach a failure detection rate f 90-95%, while minimizing false positives through advanced chniques such as oversampling and cost-sensitive learning.

II. UPCOMING RESULTS & ACTIVITIES

The results will be presented through comprehensive sualizations that highlight the model's performance cluding matrices, curves, and plots. An ROC curve Receiver Operating Characteristic curve) will be used to aluate the trade-off between the true positive rate (TPR) and lse positive rate (FPR), allowing for a better understanding f the model's discrimination ability. The x-axis will represent e False Positive Rate (FPR), while the y-axis will represent e True Positive Rate (TPR). Additionally, performance

metrics such as accuracy, precision, recall, F1 score, nd area under the ROC curve (AUC) will be used to compare e performance of Random Forest and XGBoost models, ssessing their relative effectiveness in predicting hard drive ilures. I will examine how well the models adapt over time training on different time windows of the data for a better nalyze a temporal drift.

V. KDE 124 S.M.A.R.T.

The script successfully loaded the January 1, 2017 Backblaze hard-drive dataset and generated several diagnosti visualizations and descriptive statistics to assess the distribution and behavior of 124 SMART attributes. After verifying that the dataset contained the columns of interest specifically "*capacity_bytes*, *failure*, *smart_1_raw*, and *smart_2_normalized"* the script produced distribution plots for each attribute. The failure column, which is binary was visualized using a bar chart that clearly illustrated the extrem class imbalance in the dataset. This estimated approximately 99.9986% of drives were operational, whereas only 0.0014% were reported as failed. This finding reflects the well-known rarity of drive failures in real-world operational environment and highlights the need for specialized modelling approaches that account for such imbalance. For the remaining SMART metrics, the script applied a kernel density estimation (KDE) approach after removing extreme outliers using the 1%–99% quantile range to ensure smoother and more interpretable distributions. These KDE plots exhibit strong skewness and significant variations that correspond to early warning signs c drive degradation. Additionally, each KDE plot displayed mean and median markers, along with computed statistics such as standard deviation and minimum/maximum values. This enables a deeper understanding variability while large standard deviations point to instability in certain drive health indicators. Together, these results provide an initial characterization of the dataset revealing strong class imbalance, non-linear SMART-attribute distributions, and factors that must be considered in later stages of predictive modelling.

V. DAILY FAILURE RATE

After validating the dataset, the script generated a usable date column ensuring that each record could be aligne to a specific day. The data were then grouped by date to calculate both the total number of failures and the corresponding daily failure rate. The resulting time-series visualizations revealed clear daily fluctuations in failure

Figure 4: Trimmed PDF