Failure Forecasting in Multi-Brand HDDs Using Feature-Driven Supervised Learning

Rochit Madamanchi[1], Ritu Solanki[2], Ganavi S P[3],Dr. Peeta Basa Pati [4]
Department of Computer Science and Engineering, Amrita School of Computing,
Amrita Vishwa Vidyapeetham, Bengaluru, India
[1]bl.en.u4cse23148@bl.students.amrita.edu,
[2]bl.en.u4cse232165@bl.students.amrita.edu,
[3]bl.en.u4cse23124@bl.students.amrita.edu,
[4]bp_peeta@blr.amrita.edu

***Abstract*—Hard disk drive (HDD) failure within large-scale data centers can lead to catastrophic data loss, service interruptions, and maintenance costs. Anticipating such failures ahead of time is an essential process towards adopting proactive maintenance and system reliability. HDDs today are equipped with Self-Monitoring, Analysis, and Reporting Technology (SMART), which produces internal health indicators that can be used to predict failure. This project proposes a data-driven methodology for predicting HDD failure based on the publicly available Backblaze dataset, which contains daily SMART logs from thousands of drives from many manufacturers.**

***Index Terms*—Hard disk drive (HDD), SMART attributes, failure prediction, machine learning, predictive maintenance, data imbalance, feature selection**

I. Introduction

As the data-intensive applications and cloud storage services experienced exponential growth, the dependability of hard disk drives (HDDs) in data centers has increasingly become a matter of concern. Inadvertent drive failure can cause service downtime, data loss, and heavy economic impact. Conventional drive monitoring techniques tend to fail to anticipate failing drives proactively, particularly in large-scale environments.

The contemporary HDDs come with Self-Monitoring, Analysis, and Reporting Technology (SMART), which keeps track of several internal health parameters continuously. The SMART features provide useful information about the operational condition of a drive, including metrics such as reallocated sector count, uncorrectable error count, power-on hours, and temperature. If these parameters are analyzed judiciously, they can act as precursors to drive failure.

This project aims to construct a predictive machine learning model with the Backblaze publicly available dataset, which consists of daily SMART logs of thousands of hard drives from several manufacturers. The goal is to find and identify the most useful SMART attributes and use different supervised learning algorithms to classify the drives as healthy or in danger of failure. Considering the imbalanced state of the dataset, where failure events are rare, the research also ventures into data resampling methods and performance measures other than accuracy to guarantee efficient detection of rare yet devastating failures.

Through precise early prediction of HDD failures, this work is designed to make a positive impact on proactive maintenance practices, reduce system downtime, and increase the reliability of data centers. The suggested model can also provide a basis for more extensive applications to predictive maintenance in industrial and IT networks.

II. Literature Survey

Several studies have explored the use of machine learning techniques for thyroid disease classification and clustering. Prior work includes the use of k-means, DBSCAN, and hierarchical clustering methods. These methods aim to identify subgroups among patients that share similar thyroid function profiles, enhancing diagnosis and treatment strategies.

III. Methodology

The following methodology describes the complete pipeline for clustering thyroid patient data using biochemical features from the `thyroid.csv` dataset.

A. Thyroid.csv Input

The process starts by loading a raw dataset (`thyroid.csv`) of patient-level thyroid function data. The data in this dataset comprise biochemical measurements like TSH, T3, TT4, T4U, and FTI, among other clinical attributes.

B. Data Preprocessing Block

This module deals with cleaning and preparing data for analysis. All missing values (represented by "?") are replaced by `NaN`, followed by conversion of applicable columns into suitable numeric data types. Missing values are subsequently imputed using median imputation to maintain consistency of data while reducing the effect of outliers. This guarantees data integrity for further processing.

C. Feature Standardization (StandardScaler)

The chosen features are scaled with the `StandardScaler` function from Scikit-learn, which projects the features onto a space with zero mean and unit variance. Standardization is required to ensure that the clustering algorithm handles all input features equally, avoiding bias caused by various numerical scales.

D. KMeans Clustering

This module applies the KMeans clustering algorithm on standardized data. Patients are segmented into clusters by similarity of biochemical profiles, enabling unsupervised segmentation. The value of `k` can be specified using techniques such as the Elbow Method or Silhouette Score. A cluster label is assigned to each patient representing a subgroup of similar thyroid profiles.

E. PCA Dimensionality Reduction to 2D

Principal Component Analysis (PCA) is employed to project the high-dimensional feature space onto two principal components. This mapping facilitates easy visualization while preserving most of the dataset's variance. PCA assists in interpreting the distribution and separation of clusters in the reduced feature space.

F. Visualization Module

The final module plots the clusters on a 2D scatter plot. Each data point represents a patient in PCA space, and the points are color-coded based on their cluster assignment. The visualization helps assess the quality of the segmentation and highlights distinct groups of patients based on measurements of thyroid function.

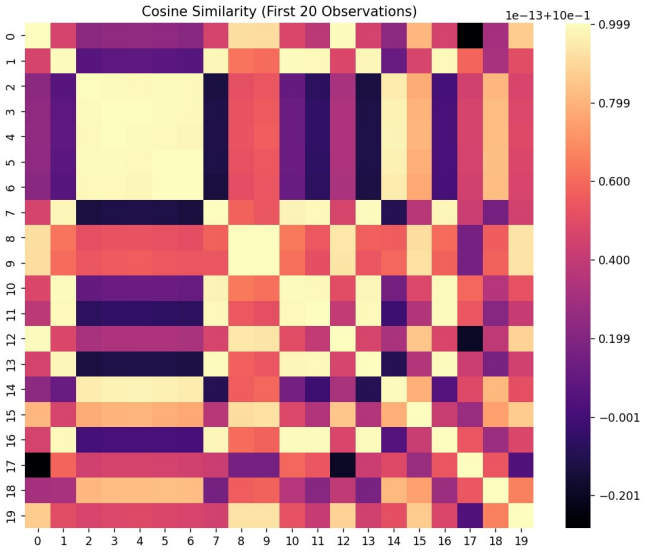

Fig. 1. Cosine Similarity

IV. Result and Analysis

The rank of an observation matrix determines its linear independence. A full-rank matrix avoids redundancy and enhances classification performance by ensuring distinct feature representations. Techniques like PCA help maintain meaningful feature spaces. Regression predicts continuous values (e.g., stock prices) using models like Linear Regression and Neural Networks, while classification assigns discrete labels (e.g., spam detection) using algorithms like Logistic Regression and SVMs. Regression uses MSE as a loss function, while classification relies on Cross-Entropy Loss. Stock price prediction relies on feature engineering, utilizing moving averages and volume trends. Historical data can be leveraged through models like ARIMA and LSTM. Market sentiment analysis, incorporating news sentiment and economic indicators, enhances predictions. Hybrid models that blend statistical and machine learning methods improve accuracy. Additionally, risk assessment using volatility measures like Bollinger Bands helps estimate uncertainty. Combining these techniques ensures more reliable and precise forecasts.

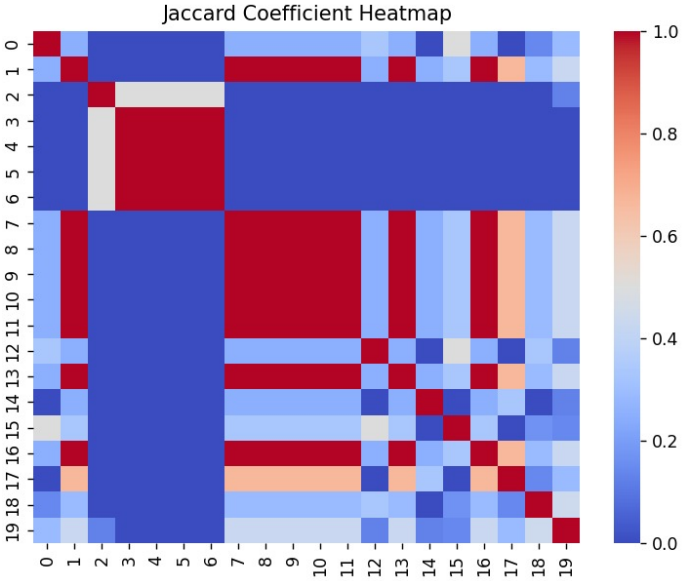

Fig. 2. Jaccard Coefficient Heatmap

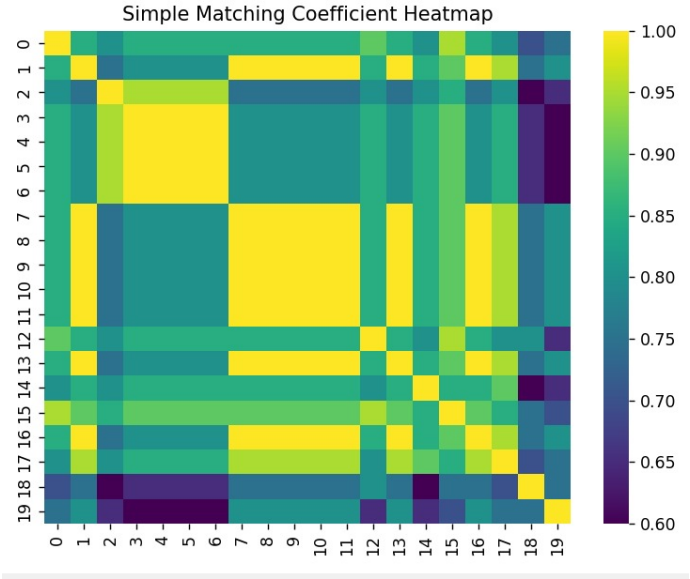

Fig. 3. Simple Matching Coefficient Heatmap

References