Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7747464

DESIGN AND IMPLEMENTATION OF AN ENSEMBLE MODEL FOR PREDICTING HARD DRIVE FAILURE

Lokendra Gour[1], Akhilesh A. Waoo[2]*
Department of Computer Science and IT, Faculty of Computer Applications, Information Technology and Sciences, AKS University, Satna (M.P.) India
* Corresponding Author - Dr. Akhilesh A. Waoo (akhileshwaoo@gmail.com)

Abstract- Disk failure predictions are essential in contemporary storage solutions for ensuring data integrity. Machine learning was shown to be an effective strategy for addressing the issue of disc failure forecasting. To better anticipate when a disc may fail, researchers have recently begun to include the idea of time windows in the field. By analyzing the correlation between the value of a disc and its age, these studies have been able to make some good predictions. Storage devices are among the most regularly replaced hardware devices and continue to provide difficulties in predicting failure. There have been several promising suggestions for developing a hard drive forecasting model based on SMART features, including the use of statistical and machine learning techniques. Although these models are used in production data centers, they were not evaluated under production conditions. In addition, hard drives deteriorate with time, although this slow decrease is not fully explained by current theories. To address the shortcomings of the currently available technology, this study introduces a novel hard drive failure prediction model that is based on Ensemble Learning. This model outperforms its predecessor, which used a Back propagation artificial neural network, in terms of prediction accuracy, model stability, and explanatory lucidity. With an overall forecast of 15 days, our proposed ensemble learning-based model is 96% accurate in predicting disc failures.
Keywords: Ensemble Learning, Back blaze predictions, Hard-drive failure, Machine learning, Data science.

INTRODUCTION

The whole system, including all of the disks in all of our data centers across the world, is captured in a snapshot once each day. To keep track of everything that happens each day, we keep a journal on individual hard drives. Around 191,000 daily reporting drives will have been logged by the end of September 2021. From April 2013 onwards, our database has approximately 266 million entries, each of which describes a specific hard drive by model, serial number, and any SMART attributes reported by the drive. Only 20–40 of the 255 possible pairs of SMART attributes may be reported by a given drive type. The date and the "failure" status are the only two fields we've added by hand; the rest of the information is provided directly by the drives themselves (with a "failure" status of 0 for operational and a "success" status of 1 for removed from service).[1] When it comes to large data centers, the consequences of a hard disk failure might be disastrous. All of the most crucial and secret information stored there is at risk. Businesses invest significant effort and resources towards predicting disk failures to take preventative measures. Organizations might save money, avoid data loss, and maintain consumer confidence if they knew when their hard drives would fail. Drives may malfunction in extreme or unexpected conditions (server power loss, natural calamity, etc.). Natural calamity, power loss at the data center, etc. Hard drives often exhibit indicators of diminishing performance in the hours and days leading up to their ultimate failure, while certain devices might die quickly. The end objective of our study is to employ machine learning to learn these patterns, identify potential reasons for hard drive failure, and forecast accidents before they occur. Our research into hard disk failure analysis and prediction employ both supervised and unsupervised learning techniques. As a means of comparing the two strategies, we'll employ both supervised and unsupervised algorithms.[2]

There has been a lot of study on fault intolerance and failure prediction because of the issue of hard disk failure. In contrast to reactive fault tolerance, proactive failure prediction may be able to foresee hard drive failure. This helps technicians prepare for the inevitable failure of a hard disk. Proactive failure prediction has been improved by algorithms including SMART (Self-Monitoring, Analysis, and Reporting Technology) components created by statisticians and machine learning experts. These methods have demonstrated some success in a range of settings, but they are not without their caveats. For instance, present prediction algorithms may only provide a "good" or "poor" rating on a hard drive, without any indication of how far or close to failure the drive is. Furthermore, these techniques only use a single instance of the SMART attributes as input for prediction, disregarding the dependency between the numerous states of a hard disk throughout the time horizon. To circumvent these limitations, we employ SMART features to accurately characterize sequential information and assess different states of hard disk health.[3] Several machine-learning-based prediction strategies have been proposed to improve HDD failure prediction accuracy. Methods like this include recurrent neural networks, convolutional neural networks, and Bayesian algorithms. Regarding FDR, prediction models based on RNNs are superior to those based on RFs (FDR). Researchers couldn't fully take advantage of RNN models' capacity to extract the temporal dependency of drive health data to acquire the characteristics of drive degeneration until significantly more data became available. However, traditional RNN models can only remember recent events due to the gradient fading or exploding. As a solution, some researchers have used a segmentation technique to standardize drive deterioration. Drives deteriorate more intricately in mobile edge computing because of the worse environmental conditions compared to the cloud. This method, therefore, is useless. The success of a prediction model is also heavily dependent on the accuracy with which the HDD health state is classified. Despite their widespread use, binary approaches and degradation degrees only look at the time sequence, rather than the HDD deterioration process, which is directly related to the current health situation. Another major challenge to HDD failure prediction is sample imbalance, which occurs when there are many more healthy drives than ill drives in the training dataset.[4] The data here was collected and analyzed from many computer systems and storage media located in a large data center. Over seventy days, we gathered information from 380,000 disks spread over 64 data centers, 10,000 racks of servers, and tens of thousands of computers. In all, this equates to about 2,600,000 hours of use time on mobile devices.

During this period, the data center housed more than two million hard drives, but we only included those having full SMART, performance, and location data in this analysis. To avoid

performance hits and privacy issues, not all drives have their SMART and performance data recorded and saved. We next propose a working definition of disk failure based on our analysis of the many disk events that have been documented in data centers. After that, we discuss the three sources of data—disk SMART properties, prior research, and our observations—that contributed to the development of our project (the standard for predicting disk failure in previous research) the dimensions and relative positioning of each disk are specified.[5]
To solve these problems, we provide a supervised machine-learning approach to detecting disks on the edge of failure and an inference approach to gathering data about failing disks. Within this study, we provide a technique for automatically labeling datasets that may be implemented on any kind of magnetic storage media. Case examples demonstrating the method's effectiveness were supplied by the European Organization for Nuclear Research (CERN). In light of CERN's focus on High Energy Physics (HEP) issues, storage services constitute an essential component of the current cloud options. For a corporation like CERN, which employs many models, accurate drive failure projections would enhance service quality while decreasing costs. To begin, we have a look at where hard disk failure prediction stands at the moment (Section 2). The continuation appears below. In Section 3, we provide a high-level summary of the machine learning strategy we've settled on, and in Section 4, we demonstrate how we put that approach into action to solve our classification problem (Section 4). This part is the last one. Using our method on the CERN disk population, we present a summary of the results and explore potential future steps in Section 5.[6]

LITERATURE REVIEW

(C. Green, 2022) [7] Back Blaze Hard Drive Data, in particular, is one of the most researched hard drive failure data sets in the literature, As a result of the vastly disproportionate number of functional drives in the active data centers to the small number of failing drives, Conventional classification or regression methods have previously made use of disc drives' SMART (self-monitoring, analyzing, and reporting technology) features. Some machine learning (ML) models have attempted to anticipate failures; these include tree-based approaches and ensemble learning algorithms.
(Ding et al., 2022) [8] As real-time monitoring systems and the Internet of Things have become more widespread, so needs scalable and parallel algorithms that can forecast when mechanical breakdowns will occur and how long a manufacturing system or its components will continue to be useable (IIoT) (IIoT. Machine learning algorithms can generate accurate predictive models, but only with access to enormous amounts of training data. Furthermore, a parallel random forest approach is developed utilizing In particular, the parallel random forest method can yield substantial speedups.
(Su & Li, 2022) [9] There have been instances where the failure rate of storage devices reached 14%. Unfortunately, there aren't always any signs of impending storage device failure. There must be a real-time predictive maintenance system that automatically predicts when maintenance should be done to drastically cut down on unscheduled downtime. Detecting a failing storage device is a time series prediction challenge, as opposed to the more usual regression predictive modeling. Time series forecasting forecasts are possible with LSTM because of its gated architecture, which automatically takes in the relevant background information. Excellent for monitoring the condition of storage devices over time.

(Chhetri et al., 2022) [10] The hard drive is crucial to a computer because it stores data and can be accessed quickly and easily. As a result, unwavering dependability from a hard disk is crucial. As a result of the gravity of the issue, several research initiatives have been launched to improve the accuracy of hard disc failure predictions. However, the numerous advantages of machine learning, such as improved pattern detection and prediction ability, are lost on research that relies solely on semantic technology. Our research presents a knowledge graph-based approach to hard disk failure prediction that draws on the most promising features of machine learning and semantic technologies. The experimental results favor our proposed method over the current best practices.

(Wen et al., 2022) [11] Prognostics used in conjunction with preventative maintenance can improve the robustness and security of engineered systems. Data-driven prognostic techniques have recently undergone a rebirth in favorability because of the widespread use of state-of-the-art sensor technology and big data analytics technologies. The purpose of the research is to provide a comprehensive review of recent and forthcoming advancements in data-driven machine prognostics, with an eye on their potential usefulness in real-world settings.

(Zafra, 2022) [12] The primary goal of the study of predictive maintenance, which is a subfield of maintenance science, is to determine the best time and method for doing maintenance on a wide range of industrial systems. The goal is to reduce the amount of time spent on system upkeep while increasing the amount of time that the monitored system is available to users. Improvements in monitoring industrial systems as part of the Industry 4.0 paradigm are driving a revolution in predictive maintenance. This has made the study of computers at the interface of industry and academia one of the most active fields of study, making predictive maintenance research one of the most active areas of study. The goal of this study is to perform a comprehensive literature analysis to examine the existing status of predictive maintenance studies that integrate data mining.

(Tomar et al., 2021) [13] Data loss due to disc failures or other system challenges can harm a data center's efficiency and accessibility. Replacing a hard disk drive (HDD) in a data center is a time-consuming and expensive process. Consequently, to lessen the frequency of this problem, several hard drive manufacturers have started prioritizing efficiency over profit. You can utilize SMART data from your hard drive to predict its failure, and many HDD manufacturers collect this kind of data from regular use. Now, with the help of SMART characteristics and some simple threshold values, you can tell when a disc drive is about to fail. The purpose of this research is to employ Machine Learning techniques for precise and preventative prediction of hard disc drive failures since these techniques are the most successful at handling learning and rare event-based problem. We compared the Random Forest algorithm to the Decision tree technique and the Naive Bayes algorithm to discover the one that created the best dependable prediction model.

(Li et al., 2021) [14] A major focus for cloud storage companies is ensuring consumer confidence by predicting when users' hard drives will fail. By looking at data collected by the hard drive itself, we found that the long-term temporal change point dependency (LTCD) of hard drive failure poses new difficulties for failure prediction. This study uses three key identifying factors, change point features, and amplifying change point dependency to present the failure prediction for temporal dependence (FPTD), which is more vulnerable to the failure of hard drives with LTCD. The experimental results show that the five FPTD evaluation measures, with Accuracy having the highest value at 99.0% and Recall having the lowest value at 97.6%, have an average score of over 94%. Due to its increased stability and superior prediction quality, the FPTD is a better fit for long-term and short-term hard disc failure prediction.

(Leukel et al., 2021) [15] Predicting whether or not a material system of interest will fail in the future is known as failure prediction. The results of this study have far-reaching consequences for predictive maintenance and other approaches to maintenance in the industry. Machine learning (ML)-based prediction models have been demonstrated to be accurate, hence this approach is being utilized more frequently to solve the issue. We looked through scholarly journals for accounts of studies done between 2012 and 2020. Only 34% of the 1,024 articles that were initially submitted for consideration were ultimately accepted. Specifically, we investigated feature engineering, feature processing, model training, and model evaluation. While our research indicated that state-of-the-art preprocessing and training methods are used, we also discovered that (3) performance evaluation is inadequate, which contributes to the overfitting problem, and (4) the reporting of experimental designs and results differs considerably. We point up areas for further research and offer ideas to improve the synthesis and comparison of data from various studies.

(Demidova & Fursov, 2021) [16] The difficulty of estimating the RUL of a hard drive is discussed in this article. Decision trees, random forests, Simple RNN, GRUs, and LSMs are only some of the machine learning and deep learning techniques that are considered alongside other possible solutions (LSTM). In this post, we'll take a look at the pros and cons of using feature creation to enhance time series forecast accuracy, comparing the two approaches side by side. Using data that is already out there, researchers examine the behaviour of tens of thousands of disc drives. All used algorithms and methodologies are laid out in a complete comparison matrix.

(Amram et al., 2021) [17] Modern machine learning algorithms are often opaque black boxes with extraordinary predictive power when it comes to data-driven predictive maintenance. It becomes harder for people to utilize these models to comprehend and gain insights into the failure processes at the core of them, which lowers confidence in the system's performance on new data. We investigate the issue of predicting hard disk failure in data centers using state-of-the-art, clear machine learning techniques. We demonstrate that these methods maintain a high level of forecasting accuracy while delivering crucial details about the drive's current and future health. We also show that these analyses can still be used in situations when data collection has barely begun, as they provide valuable insights even with minimal background information.

(Vollert et al., 2021) [18] Machine learning (ML) has been increasingly popular in the field of predictive maintenance in recent years (PdM). Despite significant performance improvements in recent years, the credibility of ML models may be damaged by a lack of explaining ability or interpretability. That is true for both models that already have a justification and those that don't. Ability, awareness of the potential for black box model explanation techniques to exhibit black box behavior, ambiguous language, a lack of time series data enquiry, insufficient explanation coverage, and the incorporation of domain knowledge are all necessary.

(Otaku, 2021) [19] The expectation in the industry is that machine learning technology will be used to construct a hard disc failure prediction model, which will allow for more accurate detection of hard disc failures in advance, a reduction in operation and maintenance costs, and an improvement in the quality of the customer experience. In this particular scenario, a hard disc failure prediction model will be trained with the use of a random forest technique.

(Zhai et al., 2021) [20] Predictive maintenance is one of the most cutting-edge technologies of recent years, garnering interest from both academics and industry (PdM). Advanced machine learning (ML) models for RUL prediction are being developed by researchers, but the great majority of these models have not been validated on industrial data and were not developed with practical use in mind. To unite academics and businesspeople and generate value, we present a unified strategy that seeks to combine PdM models with production scheduling. PdM integration into production scheduling requires a separate health prognostics model for each operation (PdM-IPS). We provide a generative deep learning model based on the conditional variational autoencoder for the goal of extracting a health indicator (HI) specific to a given operation from large volumes of industrial CM data (CVAE). To successfully use PdM in the business world, we use an unsupervised learning method because labeled failure data is extremely rare. The health prognostics model enables the PdM-IPS by providing a quantitative assessment of deterioration for a specified production run.

(Ren, 2021) [21] The advantages of predictive maintenance over more conventional maintenance tactics like corrective and preventative upkeep have led to its growing popularity. The past two decades have seen the rise of machine learning, which has helped remedy many of the drawbacks of earlier methods of maintenance prediction. Machine learning has shown unprecedented predictive capability in the realm of maintenance planning and optimization. Corrective, preventative, and predictive maintenance are compared, and the advantages and disadvantages of each are discussed. The advantages of machine learning for predictive maintenance are discussed, as are the reasons for its increasing popularity.

(Cerquitelli et al., 2021) [22] Predictive maintenance is a strategy for ensuring that assets always function at their best by proactively inspecting and fixing any problems that may arise. Fewer unplanned breakdowns and production halts, as well as reduced maintenance expenses, could lead to cheaper production costs. By building on previously stated standards and key data-driven techniques, this chapter seeks to define the essential operational blocks for predictive maintenance using modern advances in ICT and AI. Technical specifics about feasible data models for storing and distributing critical information, as well as a variety of deployment options for predictive analytics, are presented in the final section of the chapter.

THE PROPOSED METHOD

Data collection, pre-processing, data scaling, data normalization, model building, and model validation are all depicted in Figure I. These phases make up the present technique, which may be seen in its entirety below. The Python-based Ensemble model was used in this research to construct the classifier that was applied to the data. Both the outcomes and the precision of the Hard disk Failure prediction are greatly improved as a result of this research. We offer a collection of machine learning methods based on a soft vote classifier that can differentiate between states of Hard Disk Failure failure and non-failure. These algorithms are designed to detect if a hard disk is failing or not. The data were first cleaned and prepared, and then they were scaled so that they could be used in the model. Performing some exploratory data analysis was the first step we took before beginning the first processing and visualization of the dataset. A representation in schematic form of the suggested ensemble technique, which makes use of a voting classifier that is more forgiving. First, the data needed for this study were collected, then the data were analyzed with EDA, and finally, the data were integrated utilizing shared properties. After that, the data were standardised. Alterations were made to the parameters in a variety of different ways as well. When applied to training data, the Ensemble Classifier offers a method of categorization that is characterized by a high degree of accuracy. Both the test set and the training set are constructed using the complete dataset as their starting point. The model is initially trained with 80% of the data from the dataset, and then the remaining 20% is used for testing.

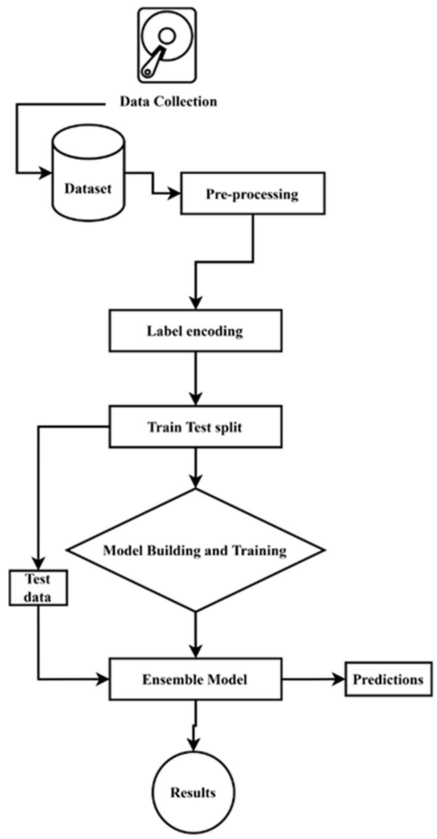

Figure1: Process Diagram of Proposed system

Dataset Description:

We use the public dataset that was provided by Backblaze, which captured a daily snapshot of each operational disk drive in their data centers. This allows us to evaluate our learning technique using this dataset. The snapshot stores not only the SMART values of a disk but also other basic information about it as well. Researchers have pointed out that different disk types need to be modeled in their unique way because of changes in the encoding and meaning of the SMART values that are provided by different manufacturers. Every every day, an image of all of the available hard disks is captured using the cloud storage facility provided by Backblaze. This snapshot includes fundamental drive information in addition to the S.M.A.R.T. statistics that were reported by the respective drive. One record or row of data constitutes the daily snapshot of a single drive. The snapshots taken by all of the active hard drives for a certain day are each saved in their row within a single file. This file can be opened in the Comma Separated Values format (often known as "CSV"). The date is added to the end of the filename, making it 2019-07-01 in our example. The name of the file for the first of July 2019 would be csv.2019-07-01.

Pre-Processing:

The data needs to be pre-processed before the machine learning algorithm can perform at its best. This is a prerequisite step. Following the completion of the data normalization stage, the subsequent pre-processing procedures can then get underway. One application that can profit from utilizing this strategy is the transformation of linear data. During the process that is known as "Standard scaling," the values of all of the attributes are standardized such that they fall within the range [0,1]. In the data's attributes, it was usual to see the value "0" as well as blanks. The gaps in the data were filled up with the median sales price of the relevant category of properties.

Model Building:

Several distinct machine learning methods, including Decision Tree, Random Forest, and Light GBM, are incorporated into our suggested strategy. Accuracy can be improved by using a voting classifier in conjunction with the aforementioned methods. The next section offers a brief overview of a few of these algorithms.

Decision Tree:

Every possible outcome for a given input can be represented in a decision tree, a graph that uses a branching method. Decision trees can be drawn by hand, although graphical editors and specialized software are also viable alternatives. To avoid clutter, we label drives as "failed" when we observe a collection of drives over time and discover that some of them fail at different times. All drives that are still working properly are good or healthy.

Random Forest:

Random forest is one of the most effective machine learning models for classification accuracy. A classifier known as a random forest employs many decision trees to arrive at a single classification. The inputs to such trees are selected using independent, uniformly distributed random variables, and the final categorization is reached by a majority-rule procedure.

Light GBM:

Microsoft's "Light Gradient Boosting Machine" (abbreviated as "LightGBM") is a free, open-source, distributed gradient boosting platform for machine learning. It's an approach to machine learning that uses decision trees to do tasks like ranking and categorization. LightGBM, which stands for "Light Gradient Boosting Machine" was developed by Microsoft and can be used for free if you're in the machine learning industry. This machine learning tool is based on decision tree algorithms and is employed in ranking and classification tasks.

Proposed Model:

As a concept, it might make use of consensus voting to integrate the best features of many machine learning algorithms into a unified, more reliable prediction. Both a hard and soft vote can be used by a voting classifier. The aggregator uses majority voting in hard voting to choose the most preferred class prediction from the underlying models before generating a final prediction. Even in the most basic realizations of soft voting models, the Predict probability technique is crucial. The vote classifier improves accuracy by mixing data from several base models. The suggested model was developed by extensive use of many categorization techniques. Several techniques, such as the Decision Tree, the Random Forest, and the Light GBM classifier, are used. A soft voting classifier can determine the likelihood of each target variable by looking at the prediction proba column of the attributes table. Next, we used Decision Trees, Random Forests, and Light GBM models to analyze the combined data and training set. The final prediction is reached by a majority vote utilizing a vote aggregate and a soft vote approach, with each model contributing its very own projection.

RESULT AND ANALYSIS

Three different machine learning models—the Decision Tree, the Random Forest, and the Light GBM—are used in the suggested approach. Data collected by Backblaze SMART was used in the "Hard Disk Failure" study. The data set is evenly divided between test and training data, accounting for 80% of the total. Common metrics used to judge an algorithm's efficacy and trustworthiness include accuracy, precision, recall, and F1 score. When the expected class value is 1 and the observed class value is also 1, this is known as a true positive (tp). For a true negative, it is sufficient that the observed value for a given class also is 0 (tn). Either a false negative (fn) or a false positive (fp) may be obtained if the predicted class and the actual class are different. Accuracy, defined as the ratio of correct predictions to total observations, is the most important metric. Here are several formulae for calculating precision, accuracy, recall, and F1 score.

$$Accuracy = \frac{tp + tn}{tn + tp + fp + fn} \quad (1)$$

$$Precision = \frac{tp}{tp + fp} \quad (2)$$

$$Recall = \frac{tp}{tp + fn} \quad (3)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

In this post, we discuss our findings from an investigation of the most used machine learning algorithms that are used to determine whether or not a hard drive has failed. This is done so that we may compare the effectiveness of several different strategies side by side with one another. Table 1 reveals that the ensemble-suggested voting classifier outperformed prior machine learning algorithms and research in terms of accuracy (96.35%), precision (96.26%), F1 score (96.36%), and recall (96.99%).

TABLE: 1 Accuracy table

Model	Accuracy	F1-Score	AUC Score	Precision	Recall
K-NN	0.7093	0.6727	0.7066	0.7412	0.6159
Logistic-regression	0.5693	0.4258	0.5624	0.6029	0.3292
Decision Tree	0.9343	0.9325	0.9343	0.9288	0.9362
Random Forest	0.9604	0.9596	0.9607	0.9505	0.969
Light-GBM	0.9604	0.9596	0.9607	0.9505	0.969
Ensemble-Model	**0.9635**	**0.9626**	**0.9636**	**0.9555**	**0.9699**

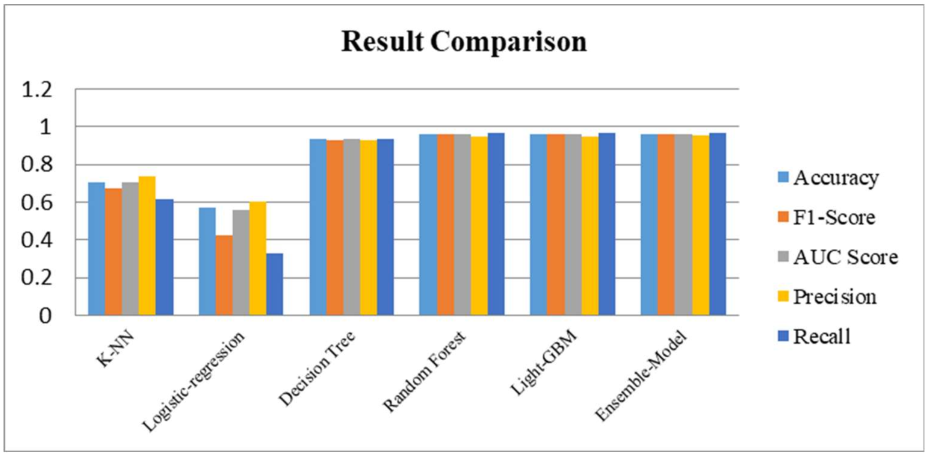

Figure: 2 Accuracy Comparison Graph

Confusion Matrix: Due to its performance, a classification algorithm may incorrectly label certain data. As such, the confusion matrix may be used to make inferences about the accuracy and frequency of mislabeling in a given dataset. According to the definition provided in the scholarly literature, the matrix's diameter represents the accuracy with which data has been categorized. If the matrix is all zeros except the diagonal, this approach is accurate to within a decimal point. To see if the data are spread uniformly amongst the various buckets, utilize the confusion matrix.

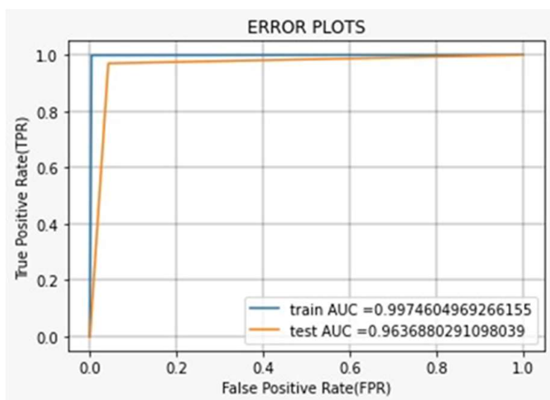

Figure: 3 AUC Plot

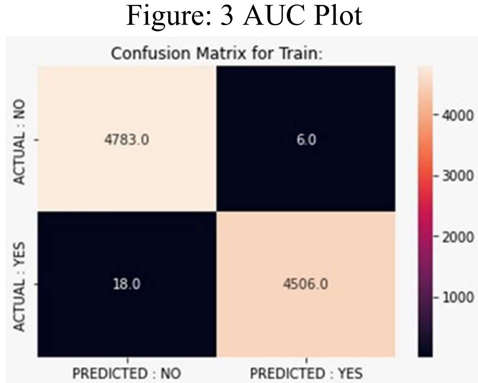

Figure: 4 Confusion MatricesFor Train

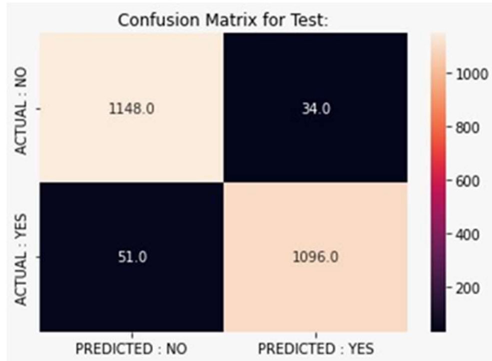

Figure: 5 Confusion Matrices For Test

CONCLUSIONS:

Our study presents a suite of tree-based algorithms for classifying and forecasting the likelihood of hard drive failure. Both in terms of how well predictions hold up over time and how easily they can be understood by others. Here, we put a voting classifier to the test in the context of prediction models. Compared to the option of not altering the models at all, the smoothness of our modified strategies maintains higher prediction accuracy. Therefore, we provide a voting classifier model that makes use of three machine learning techniques (Decision Trees, Random Forest, and light GBM). The proposed method has been applied to the experimentally collected Hard Disk Failure dataset. On the Testing set, the ensemble classifier achieved a 96.3% success rate.

REFERENCES:

[1] A. Klein, “Using Machine Learning to Predict Hard Drive Failures,” 2021.

[2] Akarshit Wal, “Prediction-of-Hard-Drive-Failure,” 2020.

[3] C. Xu, G. Wang, X. Liu, D. Guo, and T. Y. Liu, “Health Status Assessment and Failure Prediction for Hard Drives with Recurrent Neural Networks,” IEEE Trans. Comput., 2016, doi: 10.1109/TC.2016.2538237.

[4] J. Shen, Y. Ren, J. Wan, and Y. Lan, “Hard Disk Drive Failure Prediction for Mobile Edge Computing Based on an LSTM Recurrent Neural Network,” Mob. Inf. Syst., 2021, doi: 10.1155/2021/8878364.

[5] S. Lu, B. Luo, T. Patel, Y. Yao, D. Tiwari, and W. Shi, “Making disk failure predictions SMARTer!,” 2020.

[6] F. Gargiulo, D. Duellmann, P. Arpaia, and R. Schiano Lo Moriello, “Predicting hard disk failure by means of automatized labeling and machine learning approach,” Appl. Sci., 2021, doi: 10.3390/app11188293.

[7] J. and R. C. Green, “Predicting severely imbalanced data disk drive failures with machine learning models,” 2022.

[8] Y. Ding, Y. Zhai, Y. Zhai, and J. Zhao, “Explore deep auto-coder and big data learning to hard drive failure prediction: a two-level semi-supervised model,” Conn. Sci., 2022, doi: 10.1080/09540091.2021.2008320.

[9] C. J. Su and Y. Li, “Recurrent neural network based real-time failure detection of storage devices,” Microsyst. Technol., 2022, doi: 10.1007/s00542-019-04454-8.

[10] T. R. Chhetri, A. Kurteva, J. G. Adigun, and A. Fensel, “Knowledge Graph Based Hard Drive Failure Prediction,” Sensors, 2022, doi: 10.3390/s22030985.

[11] Y. Wen, M. Fashiar Rahman, H. Xu, and T. L. B. Tseng, “Recent advances and trends of predictive maintenance from data-driven machine prognostics perspective,” Meas. J. Int. Meas. Confed., 2022, doi: 10.1016/j.measurement.2021.110276.

[12] A. E. and A. Zafra, “Data mining in predictive maintenance systems: A taxonomy and systematic review,” 2022.

[13] V. Tomer, V. Sharma, S. Gupta, and D. P. Singh, “Hard disk drive failure prediction using SMART attribute,” 2021. doi: 10.1016/j.matpr.2021.03.229.

[14] X. Li, L. Zhu, C. Zhang, H. Yang, H. Wang, and J. Zhang, “Failure prediction for temporal dependency of hard drives,” 2021. doi: 10.18178/wcse.2021.06.054.

[15] J. Leukel, J. González, and M. Riekert, “Adoption of machine learning technology for failure prediction in industrial maintenance: A systematic review,” Journal of Manufacturing Systems. 2021. doi: 10.1016/j.jmsy.2021.08.012.

[16] L. A. Demidova and I. A. Fursov, “Aspects of Feature Engineering in the Problem of Predicting the Service Life of Hard Drives,” 2021. doi: 10.1109/SUMMA53307.2021.9632089.

[17] M. Amram, J. Dunn, J. J. Toledano, and Y. D. Zhuo, “Interpretable predictive maintenance for hard drives,” Mach. Learn. with Appl., 2021, doi: 10.1016/j.mlwa.2021.100042.

[18] S. Vollert, M. Atzmueller, and A. Theissler, “Interpretable Machine Learning: A brief survey from the predictive maintenance perspective,” 2021. doi: 10.1109/ETFA45728.2021.9613467.

[19] T. Otaku, “Hard disk failure prediction based on random forest algorithm,” 2021.

[20] S. Zhai, B. Gehring, and G. Reinhart, “Enabling predictive maintenance integrated production scheduling by operation-specific health prognostics with generative deep learning,” J. Manuf. Syst., 2021, doi: 10.1016/j.jmsy.2021.02.006.

[21] Y. Ren, “Optimizing Predictive Maintenance with Machine Learning for Reliability Improvement,” ASCE-ASME J. Risk Uncertain. Eng. Syst. Part B Mech. Eng., 2021, doi: 10.1115/1.4049525.

[22] T. Cerquitelli et al., “Data-Driven Predictive Maintenance: A Methodology Primer,” 2021. doi: 10.1007/978-981-16-2940-2_3.