**Name of the Project: Assert Failure Prediction**

**PROJECT REPORT**

**Submitted by : NEWBIE TEAM**

**TEAM MEMBERS: B.MUNEENDRA**

**S.HARIPRIYA**

**P.GUNAPRIYA**

**K.INDUPRIYA**

**B.BABYSAINI**

**In partial fulfilment for the award of the Certificate**

**Of**

**SUMMER INTERNSHIP PROGRAM**

****

**Department of Computer Science and Engineering**

**Annamacharya Institute of Technology and Sciences**

**Venkatapuram village, Renigunta Mandal, Tirupati, Andhrapradesh 517520**

**July 2019.**

**BONAFIDE CERTIFICATE**

**This is to certify that the project entitled ”ASSERT FAILURE PREDICTION” submitted by : B.MUNEENDRA, S.HARIPRIYA, P.GUNAPRIYA, K.INDUPRIYA, B.BABYSAINI in partial fulfilment for the requirements for the award of internship certification in technologies of Machine learning and Deep learning is an authentic work carried out by them under my supervision and guidance.**

**To the best of my knowledge, the matter embodied in the project report has not been submitted to any other University/Institute for the award of any Degree or Diploma.**

**Signature of Supervisor                         Signature of Head of the Department**

**Akshay Kumar Kothuri, M.Tech       Mrs.B.Rupa Devi M.Tech,Ph.D.,**

**AI & IoT Developer Assistant Professor &amp; HOD,**

**SmartBridge Educational Services PVT LTD Department of CSE,**

**Hyderabad – 500096 AITS,Tirupati.**

**Project Report Template**

**1. Title of the project**

**1.1. Introduction**

**1.2. Objective of Research**

**1.3. Problem Statement**

**1.4. Industry Profile**

**2. Review of Literature**

**3. Data Collection**

**4. Methodology**

**4.1. Exploratory Data Analysis**

**4.1.1. Figures and table**

**4.2 Statistical techniques and visualization**

**4.3 Data Modeling and visualization**

**5. Findings and Suggestions**

**6. Conclusion**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1.ASSERT FAILURE PREDICTION  1.1 INTRODUCTION:             Python is a popular programming language.It was created by Guido Van Rossum,and released in 1991.  PYTHON IS USED FOR:  \* web development(server-side), \* Software development, \* Mathemtics, \* System scripting.                         Python can be used on a server to create web applications.It can be used to handle big data and perform complex mathematics.Python works on different platforms(Windows,Mac,Linux,RaspberryPi,etc). Libraries used in pthon:-           Some of the libraries in python which can be used by developers to implement machine learning in their existing applications.  \*TensorFlow \*Scikit-Learn \*Numpy \*Keras \*Pandas \*Seaborn MACHINE LEARNING:      Machine learing is an application of artificial intelligence(AI)  that provides systems the ability to automatically learn and improve from experiece without being explicitly programmed.Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. USES OF MACHINE LEARNING:  \*It is used to detect patterns in a dataset and adjust program actions accordingly  \*It focus on the development of computer programs that can teach themselves to grow and change when exposed to new data  \*It enables computers to find hidden insight using iterative algorithms INTRODUCTION ABOUT PROJECT:                 Many types of machines are currently being used in modern industrial processes all of them susceptible to many different failure modes.These failure modes can have consequences ranging from a mild inconvenience to life-threatening situations.Additionally,every fault occurence has an associated cost.This cost include the parts and labor for repairing or replacing the failed system component,as well as the cost incurred by the production line shutdown,and possibly the repair of the  collateral damage.Because of the associated risks and costs of failures,early detection,isolation,and prediction of faults is important  1.2.OBJECTIVE OF RESEARCH:- \* Undertand failure machanisms of assets that lead to failure of Machinary products. \* Find out the features associated with different failure modes. \* Investigate different sensor technologies for future detection. \* Investigate different techniques and develop algorithms for detecting changes in the condition that influence the quality of machinary products.  2.REVIEW OF LITERATURE        Business failure prediction is a topic of great importance for a lot of people(shareholders,banks,investors,suppliers,..).Thats why a lot of models were developed in order to predict it.Statistical procedure(multiple discriminant analysis,logit or probit) were among the most used methods in this kind of problem.However,parametric statistical methods require the data to have a specific distribution.In addition to the restriction on the distribution involved,multicollinearity,autocorrelation and heteroscedasticity could lead to problems with the estimated model with some statistical methods.                            Because of these drawbacks,others methods have been investigated:   \* Multicriteria methods(i.e UTA,Electre tri,..) or  \* Machine learning methods(i.e neural networks,genetic algorithm,decision tree,instance based learning,...)                         Our main target is to provide a review of the literature but also to have a larger view than usually by evoking causes,symptoms and remedies of bankruptcy.We also proposed new perspectives and topics of research The blogs we have studied to fetch the info regarding projects :  \* Fault detection and prediction with application to rotating machinery  \* Error Log Processing for Accurate Failure Prediction  \* Market Economy and the economic case for a mandatory disclosure system  \* Machine learning in bioinformatics  \* Studying Test Case Failure Prediction for Test Case Prioritization  \* Random-forest-based failure prediction for hard disk drives  3.DATA COLLECTION  A data set is a collection of related,discrete items of related data that may be accessed individually or in combination or managed as a whole entity.  The data set which we use in this project consists of some attributes:  \*Temperature  \* Humidity  \*Measure 1  \*Measure 2  \*Measure 3  \*Measure 4  \*Measure 5  \* Failure  Here,this is the sample data     C:\Users\HP\Pictures\Screenshots\Screenshot (64).png   |  |  | | --- | --- | |  |  | |  |  |

4.METHODOLOGY

 The major component of the system combines financial ratio analysis and the statistical technique known as multivariate discriminant analysis, to produce a predictive model made up of seven variables, measuring distinct aspects of company financial structure, all transformed into a single value called the Z score. Good distinction between the scores of solvent and failed companies was provided. This technique is widely employed in the commercial sector with much of the work concentrated on failed and healthy companies. A secondary method was developed to reinforce the financial approach, whereby managerial performance aspects are weighted, combined and a cut-off, known as the A score value, determined to separate the two groups. The concept behind the A score is based on the belief that if a company is in financial difficulty the reason generally relates to inadequate management ability and errors perpetrated earlier. The A score is designed to address this aspect of failure prediction. By operating these two principal methods in conjunction, it is possible to predict with confidence who could be next to fail.

\*As our data is in the form of categorical i.e yes/no and it clearly shows that it is a classification model.So,we have saggregated our project into classification model.

\*RandomForestClassifier algorithm is used in this project.Because the accuracy which we obtain is approximately 0.8 . And it is a considerable model when compared with other algorithms

4.1.STATISTICAL TECHNIQUES AND VISUALIZATION

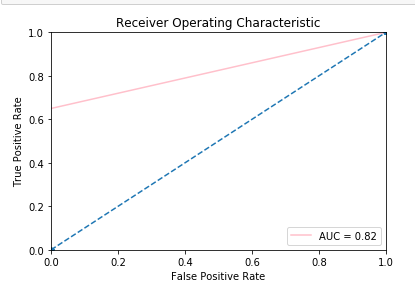
ROC: The true-positive rate is also known as sensitivity, recall or probability of detection in **machine learning**. ... The **ROC** is also known as a relative operating characteristic curve, because it is a comparison of two operating characteristics (TPR and FPR) as the criterion changes.

\*ROC curve can be generated by plotting the [cumulative distribution function](https://en.wikipedia.org/wiki/Cumulative_distribution_function) (area under the probability distribution from {\displaystyle -\infty } to the discrimination threshold) of the detection probability in the y-axis versus the cumulative distribution function of the false-alarm probability on the x-axis.

\*ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic [decision makin](https://en.wikipedia.org/wiki/Decision_making)

AUC: **AUC** is an abbrevation for area under the curve. It is used in classification analysis in order to determine which of the used models predicts the classes best. An example of its application are ROC curves. Here, the true positive rates are plotted against false positive rates.

\***AUC** = 1 means a perfect **classifier**, **AUC** = 0.5 is obtained for purely **random classifiers**. **AUC** < 0.5 means the **classifier** performs worse than a **random** one. The higher the value of GC, the better.



4.2.DATA MODELING AND VISUALIZATION

Failure prediction for hard disk drives is a typical and effective approach to improve the reliability of storage systems. In

a large-scale data center environment, the various brands and models of drives serve diverse applications with different

input/output workload patterns, and non-ignorable differences exist in each type of drive failures, which make this

mechanism much challenging. Although many efforts are devoted to this mechanism, the accuracy still needs to be

improved. In this article, we propose a failure prediction method for hard disk drives based on a part-voting random for-

est, which differentiates prediction of failures in a coarse-grained manner. We conduct groups of validation experiments

on two real-world datasets, which contain the SMART data of 64,193 drives. The experimental results show that our

proposed method can achieve a better prediction accuracy than state-of-the-art methods

Failure prediction for hard disk drives is a typical and effective approach to improve the reliability of storage systems. In

a large-scale data center environment, the various brands and models of drives serve diverse applications with different

input/output workload patterns, and non-ignorable differences exist in each type of drive failures, which make this

mechanism much challenging. Although many efforts are devoted to this mechanism, the accuracy still needs to be

improved. In this article, we propose a failure prediction method for hard disk drives based on a part-voting random for-

est, which differentiates prediction of failures in a coarse-grained manner. We conduct groups of validation experiments

on two real-world datasets, which contain the SMART data of 64,193 drives. The experimental results show that our

proposed method can achieve a better prediction accuracy than state-of-the-art methods

Failure prediction for hard disk drives is a typical and effective approach to improve the reliability of storage systems. In

a large-scale data center environment, the various brands and models of drives serve diverse applications with different

input/output workload patterns, and non-ignorable differences exist in each type of drive failures, which make this

mechanism much challenging. Although many efforts are devoted to this mechanism, the accuracy still needs to be

improved. In this article, we propose a failure prediction method for hard disk drives based on a part-voting random for-

est, which differentiates prediction of failures in a coarse-grained manner. We conduct groups of validation experiments

on two real-world datasets, which contain the SMART data of 64,193 drives. The experimental results show that our

proposed method can achieve a better prediction accuracy than state-of-the-art methods

Failure prediction for hard disk drives is a typical and effective approach to improve the reliability of storage systems. In

a large-scale data center environment, the various brands and models of drives serve diverse applications with different

input/output workload patterns, and non-ignorable differences exist in each type of drive failures, which make this

mechanism much challenging. Although many efforts are devoted to this mechanism, the accuracy still needs to be

improved. In this article, we propose a failure prediction method for hard disk drives based on a part-voting random for-

est, which differentiates prediction of failures in a coarse-grained manner. We conduct groups of validation experiments

on two real-world datasets, which contain the SMART data of 64,193 drives. The experimental results show that our

proposed method can achieve a better prediction accuracy than state-of-the-art methods.

Failure prediction for hard disk drives is a typical and effective approach to improve the reliability of storage systems. In

a large-scale data center environment, the various brands and models of drives serve diverse applications with different

input/output workload patterns, and non-ignorable differences exist in each type of drive failures, which make this

mechanism much challenging. Although many efforts are devoted to this mechanism, the accuracy still needs to be

improved. In this article, we propose a failure prediction method for hard disk drives based on a part-voting random for-

est, which differentiates prediction of failures in a coarse-grained manner. We conduct groups of validation experiments

on two real-world datasets, which contain the SMART data of 64,193 drives. The experimental results show that our

proposed method can achieve a better prediction accuracy than state-of-the-art methods.

Failure prediction for hard disk drives is a typical and effective approach to improve the reliability of storage systems. In

a large-scale data center environment, the various brands and models of drives serve diverse applications with different

input/output workload patterns, and non-ignorable differences exist in each type of drive failures, which make this

mechanism much challenging. Although many efforts are devoted to this mechanism, the accuracy still needs to be

improved. In this article, we propose a failure prediction method for hard disk drives based on a part-voting random for-

est, which differentiates prediction of failures in a coarse-grained manner. We conduct groups of validation experiments

on two real-world datasets, which contain the SMART data of 64,193 drives. The experimental results show that our

proposed method can achieve a better prediction accuracy than state-of-the-art methods.

Failure prediction for hard disk drives is a typical and effective approach to improve the reliability of storage systems. In

a large-scale data center environment, the various brands and models of drives serve diverse applications with different

input/output workload patterns, and non-ignorable differences exist in each type of drive failures, which make this

mechanism much challenging. Although many efforts are devoted to this mechanism, the accuracy still needs to be

improved. In this article, we propose a failure prediction method for hard disk drives based on a part-voting random for-

est, which differentiates prediction of failures in a coarse-grained manner. We conduct groups of validation experiments

on two real-world datasets, which contain the SMART data of 64,193 drives. The experimental results show that our

proposed method can achieve a better prediction accuracy than state-of-the-art methods

Failure prediction for hard disk drives is a typical and effective approach to improve the reliability of storage systems. In

a large-scale data center environment, the various brands and models of drives serve diverse applications with different

input/output workload patterns, and non-ignorable differences exist in each type of drive failures, which make this

mechanism much challenging. Although many efforts are devoted to this mechanism, the accuracy still needs to be

improved. In this article, we propose a failure prediction method for hard disk drives based on a part-voting random for-

est, which differentiates prediction of failures in a coarse-grained manner. We conduct groups of validation experiments

on two real-world datasets, which contain the SMART data of 64,193 drives. The experimental results show that our

proposed method can achieve a better prediction accuracy than state-of-the-art methods

Failure prediction for hard disk drives is a typical and effective approach to improve the reliability of storage systems. In

a large-scale data center environment, the various brands and models of drives serve diverse applications with different

input/output workload patterns, and non-ignorable differences exist in each type of drive failures, which make this

mechanism much challenging. Although many efforts are devoted to this mechanism, the accuracy still needs to be

improved. In this article, we propose a failure prediction method for hard disk drives based on a part-voting random for-

est, which differentiates prediction of failures in a coarse-grained manner. We conduct groups of validation experiments

on two real-world datasets, which contain the SMART data of 64,193 drives. The experimental results show that our

proposed method can achieve a better prediction accuracy than state-of-the-art methods

Failure prediction for hard disk drives is a typical and effective approach to improve the reliability of storage systems. In

a large-scale data center environment, the various brands and models of drives serve diverse applications with different

input/output workload patterns, and non-ignorable differences exist in each type of drive failures, which make this

mechanism much challenging. Although many efforts are devoted to this mechanism, the accuracy still needs to be

improved. In this article, we propose a failure prediction method for hard disk drives based on a part-voting random for-

est, which differentiates prediction of failures in a coarse-grained manner. We conduct groups of validation experiments

on two real-world datasets, which contain the SMART data of 64,193 drives. The experimental results show that our

proposed method can achieve a better prediction accuracy than state-of-the-art methods

Failure prediction for hard disk drives is a typical and effective approach to improve the reliability of storage systems. In

a large-scale data center environment, the various brands and models of drives serve diverse applications with different

input/output workload patterns, and non-ignorable differences exist in each type of drive failures, which make this

mechanism much challenging. Although many efforts are devoted to this mechanism, the accuracy still needs to be

improved. In this article, we propose a failure prediction method for hard disk drives based on a part-voting random for-

est, which differentiates prediction of failures in a coarse-grained manner. We conduct groups of validation experiments

on two real-world datasets, which contain the SMART data of 64,193 drives. The experimental results show that our

proposed method can achieve a better prediction accuracy than state-of-the-art methods

RandomForest Classifier:-It is ensemble algorithm. In next one or two posts we shall explore such algorithms. ***Ensembled algorithms*** are those which combines more than one algorithms of same or different kind for classifying objects. For example, running prediction over Naive Bayes, SVM and Decision Tree and then taking vote for final consideration of class for test object.

\*Basic parameters to Random Forest Classifier can be total number of trees to be generated and decision tree related parameters like minimum split, split criteria etc.

\* The code for using Random Forest Classifier is similar to previous classifiers.

1. Import library
2. Create model
3. Train
4. Predict

Supervised machine learning models boast remarkable predictive capabilities. But can you trust your model? Will it work in deployment? What else can it tell you about the world? We want models to be not only good, but interpretable. And yet the task of interpretation appears underspecified. Papers provide diverse and sometimes non-overlapping motivations for interpretability, and offer myriad notions of what attributes render models interpretable. Despite this ambiguity, many papers proclaim interpretability axiomatically, absent further expalnation .

**5.FINDINGS AND SUGGESTIONS**

The course will provide **basic** grounding in concepts such as training and tests sets, overfitting, and **error** rates. The course will also introduce a range of **model** based and algorithmic **machine learning** methods including regression, classification trees, Naive Bayes, and random forests.

A **prediction error is** the failure of some expected event to occur. ... **Prediction errors**, in that case, might be assigned a negative value and **predicted** outcomes a positive value, in which case the AI **would** be programmed to attempt to maximize its score.

A prediction error is the failure of some expected event to occur. When predictions fail, humans can use [metacognitive](https://whatis.techtarget.com/definition/metacognition) functions, examining prior predictions and failures and deciding, for example, whether there are [correlations](https://whatis.techtarget.com/definition/correlation) and trends, such as consistently being unable to foresee outcomes accurately in particular situations. Applying that type of knowledge can inform decisions and improve the quality of future predictions.

Predictive analytics software processes new and historical data to forecast activity, behavior and trends. The programs apply [statistical analysis](https://whatis.techtarget.com/definition/statistical-analysis)techniques, analytical queries and machine learning algorithms to data sets to create [predictive models](https://searchenterpriseai.techtarget.com/definition/predictive-modeling)that quantify the likelihood of a particular event happening.

Errors are an inescapable element of predictive analytics that should also be quantified and presented along with any model, often in the form of a confidence interval that indicates how accurate its predictions are expected to be. Analysis of prediction errors from similar or previous models can help determine confidence intervals.

**6.CONCLUSION**

Because of the associated risks and cost of failures,early detection and prediction of fault is very important . The performance of a data driven methodology is dependent on the suitability of 78 the data to the type of methodology to be applied (e.g. nonlinear data to be analyzed using linear techniques). In contrast, model based methodologies are more robust to system noise and disturbance as well as multiple operation modes than data based approaches. However, for a model based methodology to work satisfactorily, the model must adequately describe system dynamics. Such equations can be quite complex, computationally expensive, and difficult to obtain.

Future work utilizing in PCA in the context of FDP focus on determining how to evaluate the physical meaning if any of the principal components. Knowledge of the physical meaning of the principle components could provide insights as to which component(s) might be useful for the isolation and prediction of faults. The use of AIS in the future should extend not only into the detection and learning of faults, but also into areas of fault isolation and prognosis. For such an application, a bank of isolators would need to be accumulated. After isolator bank has been created and an OLAD been placed in the appropriate state variables, isolation should be possible through careful observation of each OLAD behavior and comparison to the behavior modeled by the faults in the isolation bank.