AIR PRESSURE FAULT DETECTION SYSTEM

Jerusha Daflin Mahil T (Roll no: 20Z221)
Kirthi P (Roll no: 20Z227)
S Lakshitha (Roll no: 20Z241)
Sanjeeva Shekar (Roll no: 20Z245)
Shifa Mohamed Ibrahim (Roll no: 20Z249)

19Z610 – MACHINE LEARNING LABORATORY

BACHELOR OF ENGINEERING

Branch: COMPUTER SCIENCE AND ENGINEERING

FEBRUARY 2023

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
PSG COLLEGE OF TECHNOLOGY
(Autonomous Institution)
COIMBATORE – 641 004

Problem Statement:

There have been a lot of problems related to air pressure faults in trucks. This has caused many accidents due to brake failure and gear related problems . The project aims to predict air pressure fault in trucks in aspects such as braking and gear change. It uses a training set that consists of 60000 examples in total in which 59000 belong to the negative class and 1000 positive class for it to learn to distinguish between failure and proper functioning. The test set contains 16000 examples to verify if the ML model has been properly trained.

Dataset Description:

Link: https://archive.ics.uci.edu/ml/datasets/APS+Failure+at+Scania+Trucks#

The dataset consists of data collected from heavy Scania trucks in everyday usage. The system in focus is the Air Pressure system (APS) which generates pressurized air that is utilized in various functions in a truck, such as braking and gear changes. The datasets' positive class consists of component failures for a specific component of the APS system. The negative class consists of trucks with failures for components not related to the APS.

Number of Attributes: 171

Attribute Information:
The attribute names of the data have been anonymized for proprietary reasons. It consists of both single numerical counters and histograms consisting of bins with different conditions. Typically the histograms have open-ended conditions at each end. For example if we measuring the ambient temperature 'T' then the histogram could be defined with 4 bins where:

bin 1 collect values for temperature T < -20
bin 2 collect values for temperature T >= -20 and T < 0
bin 3 collect values for temperature T >= 0 and T < 20
bin 4 collect values for temperature T > 20

```
| b1 | b2 | b3 | b4 |
-----------------------------
-20 0 20
```

The attributes are as follows: class, then anonymized operational data. The operational data have an identifier and a bin id, like 'Identifier_Bin'. In total there are 171 attributes, of which 7 are histogram variables. Missing values are denoted by 'na'.

Tools To Be Used:

Pandas: Pandas is a Python library for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series.

Numpy: Numpy is a Python library for mathematical computation of large, multidimensional arrays and matrices. It offers a large collection of high-level mathematical functions to operate on these arrays.

Scikit-learn: Scikit-learn is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via an interface in Python. It can be used in bankruptcy prediction by providing necessary tools for feature scaling, model training and evaluation, data splitting and cross validation.

Robust scalar: Robust Scalar is a technique for transforming numerical features in a dataset to handle outliers. Unlike the normal scalar transformation which is sensitive to outliers, a robust scaler technique provides a more stable result by down-weighting the influence of outliers and preserving the majority of the distribution.

Simple imputed: Simple Imputation is a method used to handle missing values in a dataset. It replaces missing values with a substituted value such as the mean, median, or mode of the non-missing values. Simple imputation is a straightforward and commonly used method for dealing with missing data, but it has some limitations.

Mongodb: MongoDB is a cross-platform, document-oriented database system that is written in C++. It is classified as a NoSQL database, meaning that it does not use the traditional SQL relational model, but instead stores data in collections of BSON (Binary JSON) documents. This allows for more flexible and scalable data modeling, as well as faster and more efficient querying. MongoDB also provides features such as automatic sharding for horizontal scalability, aggregation for data analysis, and indexing for improved performance

Contributions:

Roll No	Name	Work Assigned
20Z221	Jerusha Daflin Mahil T	Training various models, ensemble models, applying hyperparameter tuning,cross validation.
20Z227	Kirthi P	Feature scaling.
20Z241	S Lakshitha	Evaluation metrics, Training various models
20Z2245	Sanjeeva Shekar	Encoding categorical data, filling missing numbers.
20Z249	Shifa Mohamed Ibrahim	Training various models, ensemble models, applying hyperparameter tuning,cross validation.

References:

1) https://typeset.io/papers/intelligent-fault-diagnosis-of-air-brake-system-in-heavy-gw4fr429fb

2) https://archive.ics.uci.edu/ml/datasets/APS+Failure+at+Scania+Trucks#

3) 3.3. Metrics and scoring: quantifying the quality of predictions — scikit-learn 1.2.1 documentation

4) https://machinelearningmastery.com/machine-learning-in-python-step-by-step/

5) Costa C.F., Nascimento M.A. (2016) IDA 2016 Industrial Challenge: Using Machine Learning for Predicting Failures. In: BostrÃ¶m H., Knobbe A., Soares C., Papapetrou P. (eds) Advances in Intelligent Data Analysis XV. IDA 2016. Lecture Notes in Computer Science, vol 9897. Springer, Cham

6) Gondek C., Hafner D., Sampson O.R. (2016) Prediction of Failures in the Air Pressure System of Scania Trucks Using a Random Forest and Feature Engineering. In: BostrÃ¶m

H., Knobbe A., Soares C., Papapetrou P. (eds) Advances in Intelligent Data Analysis XV. IDA 2016. Lecture Notes in Computer Science, vol 9897. Springer, Cham

7) Cerqueira V., Pinto F., SÃ¡ C., Soares C. (2016) Combining Boosted Trees with Metafeature Engineering for Predictive Maintenance. In: BostrÃ¶m H., Knobbe A., Soares C., Papapetrou P. (eds) Advances in Intelligent Data Analysis XV. IDA 2016. Lecture Notes in Computer Science, vol 9897. Springer, Cham

8) Ozan E.C., Riabchenko E., Kiranyaz S., Gabbouj M. (2016) An Optimized k-NN Approach for Classification on Imbalanced Datasets with Missing Data. In: BostrÃ¶m H., Knobbe A., Soares C., Papapetrou P. (eds) Advances in Intelligent Data Analysis XV. IDA 2016. Lecture Notes in Computer Science, vol 9897. Springer, Cham

9) https://www.simplilearn.com/tutorials/machine-learning-tutorial/machine-learning-steps

10) Choosing the right estimator — scikit-learn 1.2.1 documentation