Quality of iron prediction in mining process using machine learning

Bharath Kumar Nakka
700744145
dept.Computer Science
University of Central Missouri
bxn41450@ucmo.edu

Mounika Prathapani
700745641
dept.Computer Science
University of Central Missouri
mxp56410@ucmo.edu

Srilaxmi Pavuluri
700745445
dept.Computer Science
University of Central Missouri
sxp54450@ucmo.edu

***Abstract*—In general, assessing and predicting iron quality requires analyzing the complex relationship between variables, which demands domain expertise as well as chemical analysis, which is a time-consuming process. Machine learning algorithms analyze the complex relationships and nonlinearity in the data with ease. Moreover, machine learning models adaptable to changes in the model parameters, and the scalability of the models are high; they can operate on large datasets as well. By leveraging the machine learning algorithms mining processing can be optimized in terms of manpower and operational costs. In this paper we are proposing principal component regression models means combining the principal components with regression models. The main objectives of the project are building multiple machine-learning regression models to predict the percentage of Silica Feed in Iron ore., Generating Principal Components, Combining PCA and Regression models and Evaluating the Models using various regression metrics r2 value, Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, and MAPE.**

[1] ***Index Terms*—MAPE(Mean Absolute Percentage Error), PCA(Principal Component Analysis), Percentage of Silica, R2 value and regression models,**

I. Introduction

Metallic iron can be extracted from iron ores. Ores that contain greater than 60 percent of hematite or magnetite are referred to as ”natural ore” or ”direct shipping ore.” They can be directly fed into iron-making blast furnaces. The usage of this metal is not popular till the 14th century though it has been around 3000 years. One of the most abundant rock-forming elements is iron (Fe). It accounts for around 5 percent of the Earth’s crust. It is the second most prevalent and extensively dispersed metal (the most common being aluminum).

In general, assessing and predicting iron quality requires analyzing the complex relationship between variables, which demands domain expertise as well as chemical analysis, which is a time-consuming process. Datasets like mining contain highly correlated features because it is a combination of diverse factors like pulp quality and airflow. To mitigate the multicollinearity between the variables, the dimensionality reduction method PCA is used along with the regression models.

To assess the model’s robustness various evaluation metrics are used. The main features of the project are: With the accurate prediction of Silica, we can control the quality of a lot of products where silica is a main component, for example, glass, steel, and ceramic. Another benefit of predicting the percentage of Silica is manpower planning and resource allocation in many

Datasets like mining contain highly correlated features because it is a combination of diverse factors like pulp quality and airflow. To mitigate the multicollinearity between the variables, the dimensionality reduction method is used along with the regression models. To assess the model’s robustness various evaluation metrics are used.

PCA is a technique for simplifying complex data. The PCA result has a simpler structure that may be processed using any algorithms, regression, or classification. We can project the high dimensional data onto a 2D plane. This study makes use of student academic performance data from Telkom

[1] https://github.com/BharathKumar-Nakka/Quality-prediction-in-mining-process

University's Informatics Study Programme and analyzes the student performance using time series analysis and classifies the performance using the Naive Bayes algorithm.

Student learning activity data exhibits complexity. These Patterns can be utilized to increase the quality of learning. However, due to the complexity and high dimensionality of the data, information mining frequently encounters several challenges. Recently, numerous approaches for extracting meaningful information from complex data have been proposed; one potential method incorporates data visualization and clustering, which provides some level of insight into the data. This work describes the effectiveness of PCA on complex data.

Existing methods predicted the percentage of silica in the iron ore. In this study, a novel approach is proposed to predict both targets at the same time using Multi-target regression methods. For the experimental analysis, RandomForest, AdaBoost, and KNN regressors are used and performance is evaluated using the R2 metric.

Existing methods depend on the laboratory results which are produced every 2 hours. In real-time with these discrete methods, linear relationships can not be approximated. In this study, a real-time monitoring system is developed using a deep neural network architecture. Deep Neural Networks allow us to predict continuous improvements and relationships

Iron ore comes in the form of rocks and pellets. These pellets'colorr depends on the iron oxide percentage of dark green, yellow, and grey. Green pellet quality prediction is used to plan the process t yield high-quality ores. In this paper whale optimization algorithm is used to optimize the GRU network. The results of this network are compared with the laboratory results and it concludes that the optimized GRU network outperformed the traditional deep learning algorithms

II. Motivation

The percentage of Silica in Iron ore decides so many factors in real life for example quality of the steel products. Silica is the most common impurity in iron ore extraction. The percentage of silica defines the purity of the iron. Knowing the accurate percentage of silica impurity helps to plan the manpower and operational costs. High amounts of silica form the slag and define the quality of steel products in steel making motivation of the project stems from the following advantages:

- In traditional laboratories the analysis takes a considerable amount of time. In machine learning the prediction and analyzing time is very less.
- In some situations it becomes a tedious task to analyze a lot of parameters to predict the percentage of Silica.
- Machine learning quantifies the prediction results using a wide range of parameters.

III. Objectives

The five main objectives of the project are:

- Building multiple machine learning regression models to predict the percentage of Silica Feed in Iron ore.
- Exploring the data variability and randomness
- To reduce the variability in the data PCA dimensionality reduction method is applied.
- Consolidating the results of the models and conducting an experimental analysis
- Projecting PCA transformed data onto the 2-dimensional plane

IV. Related work

This paper aims at detecting the intrusion with a low detection rate. Traditional methods like SVM(Support Vector Machine) are very slow. The proposed algorithm in this study is KPCA(Kernel Principal Component Analysis) which reduces the size of the feature matrix. To perform multiclass classification the ELM(Extreme Learning Machine) algorithm is implemented. The experimental results show that introducing KPCA to ELM improved the results [20].

Epileptic seizures (Epilepsy) are neurological illnesses that impact abnormal brain activity. Earlier epilepsy was classified using EEG data with machine learning techniques. Medical datasets are imbalanced, when classes are imbalanced, machine learning algorithms perform poorly. This paper addresses the limitations of traditional methods. In the first stage, our approach uses Principal Component Analysis (PCA) to extract both high- and low-variant Principal Components (PCs). In the second stage, different PCs are fed to machine learning models to classify the epilepsy categories [8].

The goal of this project is to solve the issue of the interdisciplinary fields of computer science and engineering. In this study, we proposed two methods for reducing the size of the families of d- and q-axis current curves that define the magnetostatic characteristics of a Synchronous Reluctance Machine (SynRM): Principal Component Analysis and Polynomial Interpolation [11].

Machine-Learning (ML) methods are most popular in diagnosing diseases and most importantly the progression of the disease. The goal of our project is to predict Coronary Artery Disease (CAD) using the Z-Alizadeh Sani dataset. To select important features t-test is used and PCA for dimensionality reduction. A 10-fold Cross-validation technique is used to evaluate the results. In total 6 machine learning models are deployed to predict the disease. Out of 6 machine learning models, the ANN model performed best and achieved an accuracy of 93 percent [3].

Data generated by human locomotion activities is a process that involves the analysis of hundreds or thousands of data in a short period. It is important to select the best features of the collected signals. With the help of principal component analysis (PCA) techniques we have selected the most relevant features. The use of machine learning techniques for reduced hardware devices in intelligent environments enables the creation of a solution for non-invasive activity supervision, supplementing the use of PCA with other classification algorithms suitable for the treatment of data with a large number of characteristics, such as support vector machines (SVM) [19].

With the growing demand for location-based services, floor localization, which can determine the floor level of a mobile user within a structure, has gained a lot of attention. This research proposes (RSSI)-an based method for floor localization using principal component analysis (PCA) and ensemble extreme learning machine (ELM) methodologies. PCA feature extraction pre-processing is used for dimension reduction of the training set. To improve the accuracy of the models instead of single ELM, ensemble ELM is used [14].

The project aims to predict cell type from single-cell RNA sequencing (scRNA-seq). In this study two best approaches are implemented for the prediction task one is integrating PCA into regression models and the other one implementing boosting methods. In the first method logistic regression (LR), k closest neighbor (kNN), and supporting vector machine (SVM is used along with PCA. Experimental results show that the integration of PCA shows improvement in accuracy [5].

In this paper, 34 basic elements of digital literacy of students in vocational education are reduced to important features using PCA. These reduced features are divided into 20 principal components. Using the main factor analysis method, four main common components of digital literacy of students in vocational education and digital literacy of students were calculated [12].

In this paper, principal component analysis (PCA) is used to obtain a high compression ratio (CR) in ECG data with low reconstruction error (under 2 percentage root mean squared difference, or, PRD). To reduce the dimensions one of the variants of PCA that is lead-wise PCA is deployed for each heartbeat and the reduced components are fed to the MLP-NN [1].

In 2017, Canada's first national fire information database (NFID) enabled big data analytics to explore fire and firefighter-related issues since firefighters are the most critical resources for protecting the public and reacting to emergencies. This research proposes using principal component analysis and deep neural networks to investigate the factors that influence firefighter injuries. The approaches were validated using NFID data. The findings are useful in aiding multicriteria decision-making and decision support systems [18].

The majority of diseases can be cured when they are diagnosed early lung cancer is one of them. CAD(Computer Aided Diagnosis) methods help to diagnose diseases early. This work investigates the robustness of deep learning features from pre-trained deep learning architectures. These characteristics were compared to the widely used Gray-level Co-occurrence Matrix (GLCM). Deep features produced the maximum accuracy of 100 percent, whilst GLCM features produced 93.52 percent. This study also compared deep feature classification with five different classifiers [9].

Regression models are the most basic machine learning models and have various applications in disease diagnosis. The major challenges in predicting the results using regression models are collinearity. Collinearity in the data adds redundancy and affects the results. One way to solve the problem is removing the features before predicting the results but this is not an effective solution. The other method is using the machine learning models which have the in-built capability of removing the collinearity. Those models are Lasso and Ridge regression. In this study, we have proposed machine learning regression models to predict type-I and type-II diabetes [17].

The sales of new automobiles are increasing day by day and the trend is the same for second-hand vehicles also. In this study, we are implementing 5 different machine learning regression models to predict the prices of autos. The dataset is collected from Kaggle; it has 205 automobile brands and 26 unique features. To improve the performance data cleaning is done effectively [6].

India has become a country with water scarcity And the demand for freshwater has increased. The Water quality has declined in both urban and rural areas as a result of recent trends in urbanization and industrialization. In this project, the water quality is predicted using pH, dissolved oxygen (DO), and biochemical oxygen demand (BOD) parameters. To predict the water quality multiple regression models are deployed. The mean square error was used to compare the predictions made using the various models [15].

Ensemble methods are an efficient solution to complex data. This method can be applied to classification as well as regression problems. In ensemble regression, the process can be done in 3 stages: first, a collection of weak learner models is constructed, in the second stage an ensemble of best models is chosen, and in the end, individual model predictions are aggregated to obtain the best prediction on a test sample. In this paper, the best strategies to choose diverse models are explained [10].

Coronavirus is a global epidemic that has afflicted practically every country on the planet. The number of infected cases and dying individuals is rapidly increasing. As a result, to construct effective short-term prediction models for estimating the number of future instances, extensive investigations exhibiting the following trend of COVID-19 are necessary. Forecasting skills are often taught to aid in the development of better plans and the making of sound decisions. Forecasting techniques evaluate past events, allowing forecasts about future events to be made [4].

The influence of varied ventilation conditions on the infection potential of COVID-19 in a metro wagon is numerically investigated. Two major indications are utilized to quantify this potential. Based on the numerical data, a regression analysis is performed to determine the best regression model for these critical parameters. The presented regression models are useful in assessing the infection risk under various ventilation conditions [13].

The primary goal of this article is to forecast automobile sales using sentiment analysis from diverse sources on the internet. The online presence of a car, as well as its brand, are important factors in automobile sales. Many more criteria, however, are essential and will be covered in this work. In today's industry, sales forecasting benefits not only the manufacturer but also several other companies that make vehicle parts or accessories. It is also beneficial to retailers, showroom owners, and service mechanics. We used linear regression for sentiment analysis and polynomial regression for sales prediction in this application [7].

This study presents a generalized fuzzy regression model called the fuzzy changing coefficient regression model. We present a fuzzy changing coefficient regression model based on the Huber loss function and a kernel function in this paper. Unlike Shen et al.'s method, our method is robust in the presence of outlier data. A numerical example is used to demonstrate this benefit [2].

Many third-party merchants of automobiles and electronics often dupe customers who try to buy secondhand things. Many buyers will be uninformed about the characteristics that must be evaluated before purchasing a product and they try to sell things that are no longer functional. To address this issue, this research attempts to create three distinct regression models capable of forecasting the selling price of used cars based on input characteristics such as actual pricing, km traveled, and so on. Three different regression models: Linear regression, lasso regression, and ensemble regression are built. The ensemble regression method is used [16].

V. Data Description

The dataset is collected from Kaggle open-source repository. The main aim is to predict the percentage of Silica that is impure in iron ore. The percent of the silica column is sampled every 20 seconds and a few columns are sampled on an hourly basis. There is a timestamp for each row. Data is collected from March 2017 - September 2017. There are 737453 records and 24 features in the dataset.

S.No.	Feature	Description
1.	percent of Iron Feed	percent of Iron (ore)
2.	percent Silica Feed	percent of silica (impurity)
3.	Starch Flow	Starch flow measured in m3/
4.	Amina Flow	Amina flow measured in m3/
5.	Ore Pulp Flow	t/h
6.	Ore Pulp pH	pH scale from 0 to 14
7.	Ore Pulp Density	Density scale from 1 to 3 kg/c
8.	Flotation Column 02 Air Flow	Airflow in Nm³/h

Table I
Dataset feature description

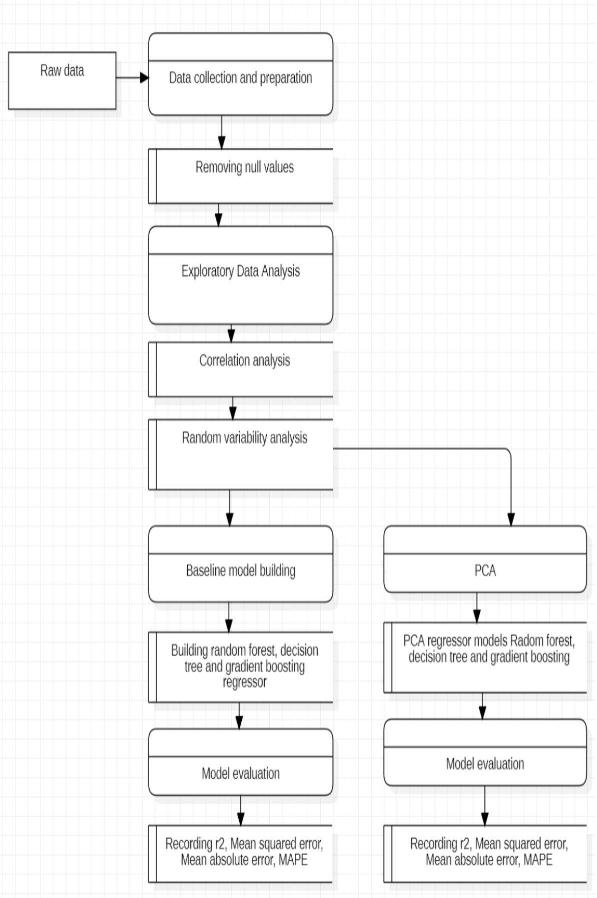

Raw data
Data collection and preparation
Removing null values
Exploratory Data Analysis
Correlation analysis
Random variability analysis
Baseline model building
PCA
Building random forest, decision tree and gradient boosting regressor
PCA regressor models Radom forest, decision tree and gradient boosting
Model evaluation
Model evaluation
Recording r2, Mean squared error, Mean absolute error, MAPE
Recording r2, Mean squared error, Mean absolute error, MAPE

Figure 1. Workflow

VI. Proposed Framework

A. Data collection and preparation

Raw data is collected and cleaned. In the cleaning process, null values are removed. There are no null values in the data. In the preliminary stage statistical analysis of the data is observed using pandas data. describe the function to gain a high-level understanding of the features.

B. Exploratory Data Analysis

In this step, the relationship of the features is analyzed using various visualization techniques. In this step in addition to the general exploration of data correlation analysis and random variability are checked. In the correlation analysis, highly correlated features are extracted using a heatmap. There is a presence of multicollinearity in the data. With time series analysis random variability of the data is observed in Fig. 2..There is a presence of random variability in the data.

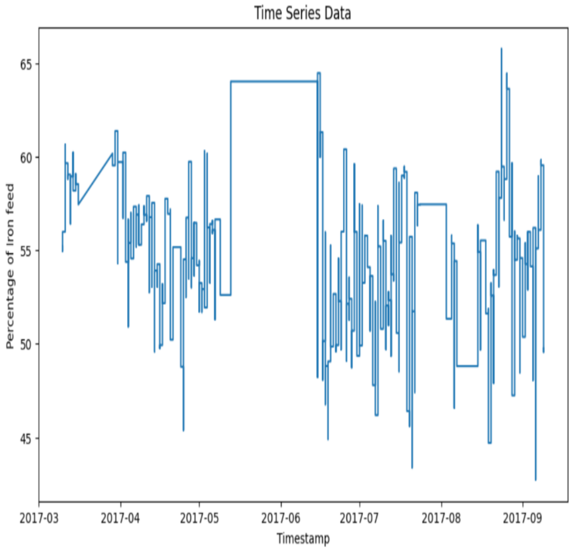


Figure 2. Time series analysis of Iron feed

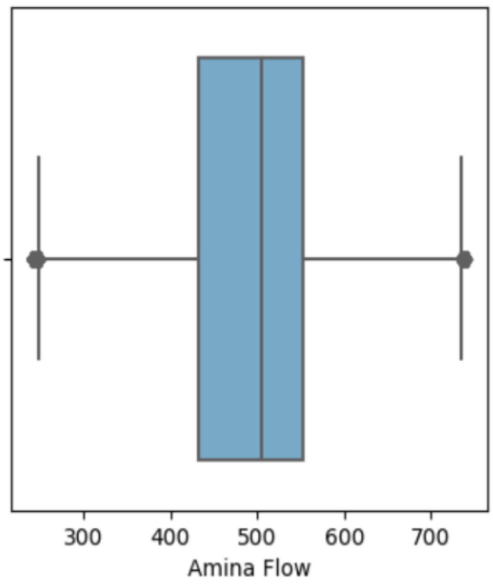


Figure 3. Distribution of Amina flow

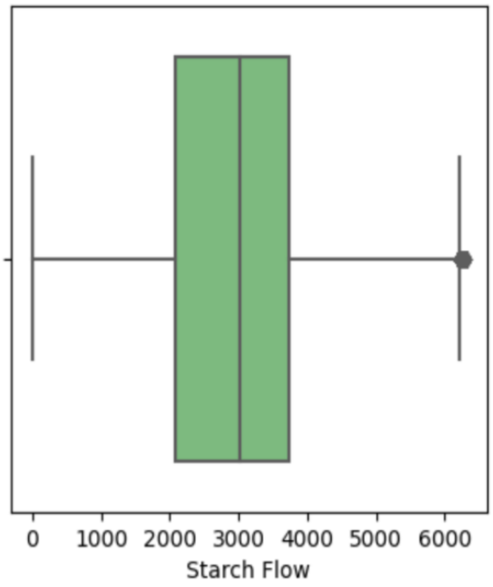
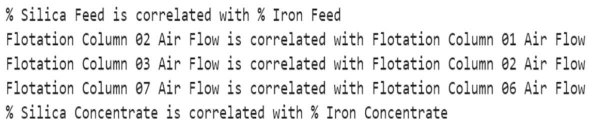


Figure 4. Distribution of starch flow



```
% Silica Feed is correlated with % Iron Feed
Flotation Column 02 Air Flow is correlated with Flotation Column 01 Air Flow
Flotation Column 03 Air Flow is correlated with Flotation Column 02 Air Flow
Flotation Column 07 Air Flow is correlated with Flotation Column 06 Air Flow
% Silica Concentrate is correlated with % Iron Concentrate
```

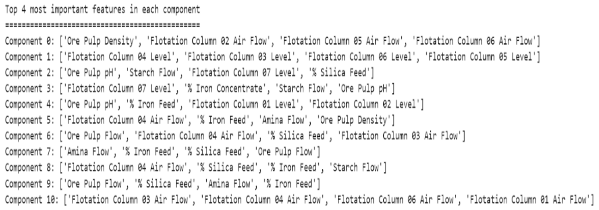
Figure 5. Multicolinearity of the data

C. Baseline models

Baseline regression models are constructed using RandomForest regressor, Decision Tree regressor, and Gradient Boosting regressor. Models are evaluated using various regression metrics r2 value, Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, and MAPE.

D. Principal Component Analysis(PCA)

To withstand the multicollinearity and random variability data is transformed using PCA. PCA object is initialized using the scikit-learn module. Before we pass the data to the PCA object data is scaled using a standard scaler. Scaled data is transformed using PCA. Cumulative variance and explained variance are visualized in fig.8,9. From the visualization, we observed that 85 of the variance is covered using 10 principal components. The PCA object is trained using 10 principal components. The top 4 most important features in each component are



```
Top 4 most important features in each component
=================================================
Component 0: ['Ore Pulp Density', 'Flotation Column 02 Air Flow', 'Flotation Column 05 Air Flow', 'Flotation Column 06 Air Flow']
Component 1: ['Flotation Column 04 Level', 'Flotation Column 03 Level', 'Flotation Column 06 Level', 'Flotation Column 05 Level']
Component 2: ['Ore Pulp pH', 'Starch Flow', 'Flotation Column 07 Level', '% Silica Feed']
Component 3: ['Flotation Column 07 Level', '% Iron Concentrate', 'Starch Flow', 'Ore Pulp pH']
Component 4: ['Ore Pulp pH', '% Iron Feed', 'Flotation Column 01 Level', 'Flotation Column 02 Level']
Component 5: ['Flotation Column 04 Air Flow', '% Iron Feed', 'Amina Flow', 'Ore Pulp Density']
Component 6: ['Ore Pulp Flow', 'Flotation Column 04 Air Flow', '% Silica Feed', 'Flotation Column 03 Air Flow']
Component 7: ['Amina Flow', '% Iron Feed', '% Silica Feed', 'Ore Pulp Flow']
Component 8: ['Flotation Column 04 Air Flow', '% Silica Feed', '% Iron Feed', 'Starch Flow']
Component 9: ['Ore Pulp Flow', '% Silica Feed', 'Amina Flow', '% Iron Feed']
Component 10: ['Flotation Column 03 Air Flow', 'Flotation Column 04 Air Flow', 'Flotation Column 06 Air Flow', 'Flotation Column 01 Air Flow']
```

Figure 6. First 4 principal components

VII. PCA REGRESSION MODELS

To check the impact of PCA on regression models. PCA-transformed data is passed to the regression models and evaluated using regression metrics.

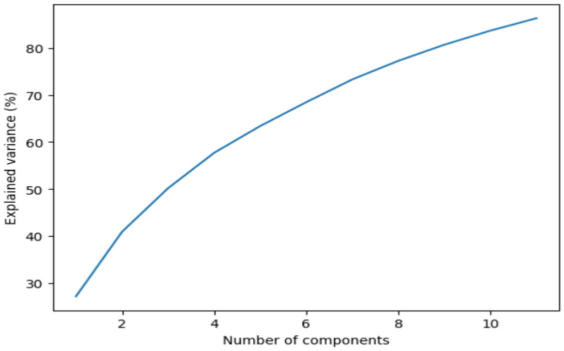


Figure 7. Cumulative variance plot

VIII. Results summary

In the first stage, baseline models are constructed using a gradient boosting regressor, Decision Tree regressor, and Random Forest regressor. Input values are standardized using standard scalar and then passed to the regression models. Each model is trained and evaluated using RMSE, R2, and MSE values are calculated. With the thumb rule of low is best PCA-Random Forest regression model performed best compared to Gradient boosting and decision tree algorithms.

The models are implemented using scikit-learn python module. The accuracy of the each modules is evaluated using different evaluation metrics.

Metric	Training	Testing
R2	0.98	0.98
MSE	0.6	0.6
RMSE	0.78	0.78
MSE	0.53	0.53

Table II
Accuracy results of Gradient Boosting

Metric	Training	Testing
R2	0.9	0.9
MSE	3.9	3.9
RMSE	1.9	1.9
MAE	1.5	1.5

Table III
Accuracy results of PCA-Gradient Boosting

Metric	Training	Testing
R2	1.0	0.94
MSE	5	2.4
RMSE	1.5	2.7
MAE	2.7	0.5

Table IV
Accuracy results PCA-DT

Metric	Training	Testing
R2	0.99	0.98
MSE	0.10	0.76
RMSE	0.3	0.8
MSE	0.16	0.43

Table V
Accuracy results of PCA-RF

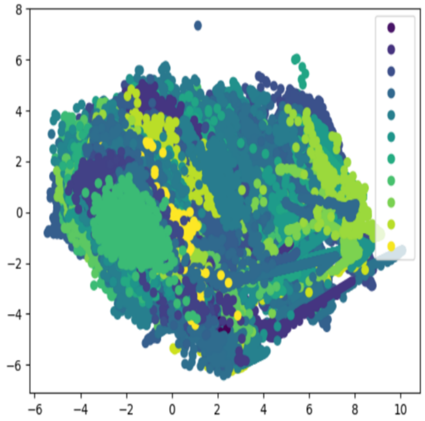


Figure 8. PCA transformed data

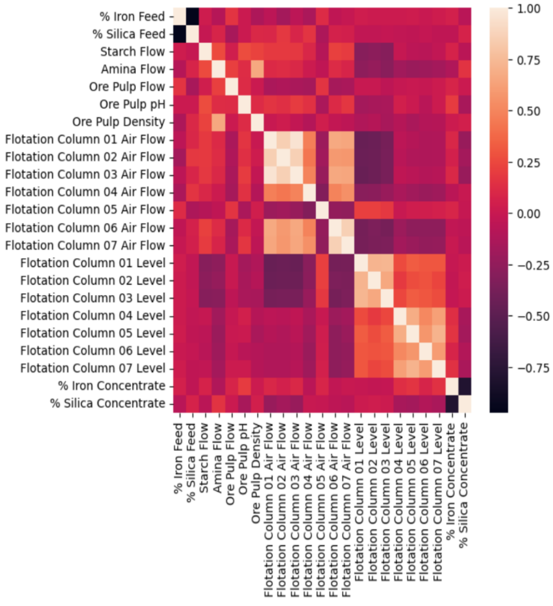


Figure 9. Correlation matrix

References

[1] Soumyendu Banerjee, Rajarshi Gupta, and Jayanta Saha. Compression of multilead electrocardiogram using principal component analysis and machine learning approach. In *2018 IEEE Applied Signal Processing Conference (ASPCON)*, pages 24–28, 2018.

[2] Thupakula Bhaskar, S Arumai Shiney, S. Babitha Rani, K. Maheswari, Samrat Ray, and V. Mohanavel. Usage of ensemble regression technique for product price prediction. In *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1439–1445, 2022.

[3] Ali Cüvitoğlu and Zerrin Işik. Classification of cad dataset by using principal component analysis and machine learning approaches. In *2018 5th International Conference on Electrical and Electronic Engineering (ICEEE)*, pages 340–343, 2018.

[4] Mostafa El-Salamony, Ahmed Moharam, and Amr Guaily. Regression modeling for the ventilation effect on covid-19 spreading in metro wagons. In *2021 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pages 55–58, 2021.

[5] Jingkai Guo and Jing Gao. Comparison of different machine learning algorithms on cell classification with scrna-seq after principal component analysis. In *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pages 1476–1479, 2022.

[6] Rupesh Gupta, Avinash Sharma, Vatsala Anand, and Sheifali Gupta. Automobile price prediction using regression models. In *2022 International Conference on Inventive Computation Technologies (ICICT)*, pages 410–416, 2022.

[7] Amirhamzeh Khammar, Mohsen Arefi, and Mohammad Ghasem Akbari. Robust fuzzy varying coefficient regression model based on huber loss function. In *2020 8th Iranian Joint Congress on Fuzzy and intelligent Systems (CFIS)*, pages 077–079, 2020.

[8] Mohammad Masum, Hossain Shahriar, and Hisham M. Haddad. Epileptic seizure detection for imbalanced datasets using an integrated machine learning approach. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 5416–5419, 2020.

[9] Joel Than Chia Ming, Norliza Mohd Noor, Omar Mohd Rijal, Rosminah M. Kassim, and Ashari Yunus. Lung disease classification using different deep learning architectures and principal component analysis. In *2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS)*, pages 187–190, 2018.

[10] Alireza Mohammadi, Dmytro Chumachenko, and Tetyana Chumachenko. Machine learning model of covid-19 forecasting in ukraine based on the linear regression. In *2021 IEEE 12th International Conference on Electronics and Information Technologies (ELIT)*, pages 149–153, 2021.

[11] Maria Nutu, Radu Martis, Horia F. Pop, and Claudia Martis. Principal component analysis for computation of the magnetization characteristics of synchronous reluctance machine. In *2018 AEIT International Annual Conference*, pages 1–6, 2018.

[12] Fei Peng, Mengke Guo, Chenglong Zheng, Shang Wang, Xuelei Wang, and Meide Xu. An assessment model of digital literacy for the students in vocational education based on principal component analysis in machine learning. In *2023 IEEE 6th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC)*, volume 6, pages 1382–1386, 2023.

[13] Sunil K Punjabi, Vikyhat Shetty, Shreemun Pranav, and Abhishek Yadav. Sales prediction using online sentiment with regression model. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 209–212, 2020.

[14] Guowen Qi, Yi Jin, and Jun Yan. Rssi-based floor localization using principal component analysis and ensemble extreme learning machine technique. In *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, pages 1–5, 2018.

[15] Bhawna Sharma and Harsimran Kaur. Parameters of water to be predicted using regression analysis. In *2022 11th International Conference on System Modeling Advancement in Research Trends (SMART)*, pages 970–976, 2022.

[16] Nantinee Soodtoetong, Eakbodin Gedkhaw, and Montean Rattanasiriwongwut. The performance of crop yield forecasting model based on artificial intelligence. In *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 683–686, 2020.

[17] Chetneti Srisa-An. Guideline of collinearity - avoidable regression models on time-series analysis. In *2021 2nd International Conference on Big Data Analytics and Practices (IBDAP)*, pages 28–32, 2021.

[18] Zijiang Yang and Youwu Liu. Investigating the influential factors on firefighter injuries using statistical machine learning. In *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 2, pages 422–427, 2018.

[19] Ricardo Yauri, Rubén Acosta, Marco Jurado, and Milton Rios. Evaluation of principal component analysis algorithm for locomotion activities detection in a tiny machine learning device. In *2021 IEEE Engineering International Research Conference (EIRCON)*, pages 1–4, 2021.

[20] Yuan Zhou, Le Yu, Mingshan Liu, Yuanyuan Zhang, and Helin Li. Network intrusion detection based on kernel principal component analysis and extreme learning machine. In *2018 IEEE 18th International Conference on Communication Technology (ICCT)*, pages 860–864, 2018.