Applied Data Science Externship

Project Report

Project Title:

Estimating the Presence of Impurities in Iron Ore using IBM Watson Machine Learning

Team Members:

Shantanu Dandekar
Konish Bagchi
Satish Sarma
Adhyan Vijeta Mittal
Rahul P

Index

1. Introduction

a. Overview

Ores are natural rocks or sediments that contain one or more valuable minerals, typically continuing metals that can be mined, treated and used for various purposes. Iron ores are rocks and minerals from which metallic iron can be extracted economically. They are usually rich in iron oxides. Iron is usually found in the form of *magnetite, hematite, goethite, limonite* or *siderite* forms. Each form of ore contains some quantity of impurities that need to be filtered out, to obtain pure or high concentration metals. Usually, one of the most commonly used iron ores, *magnetite*, has the common impurity *silica*, which ranges from 3% to 7% of the total iron ore. Estimation of the precise amount of silica impurity is a complex process which is both expensive and time-consuming. Hence, in this project, we design a Machine Learning model, which uses a dataset consisting of the percentage of silica concentrate present in iron ore pulp from a large number of samples, and accurately predicts the amount of impurity for given iron ores.

b. Purpose

This project aims to construct a Machine Learning model which can accurately predict the amount of silica impurity in a given iron ore. The dataset used helps in the training and testing of the model being created, and helps to increase the accuracy of the model due to the large number of samples considered in the data. Physical estimation of silica quantity involves complex chemical analysis. It involves conversion of the silica into potassium fluosilicate by digestion process using various other chemicals, followed by volumetric determination of the impurity. It is a labor and cost intensive process. Thus we aim to build an efficient and effective model to overcome these issues related to impurity determination.

2. Literature Survey

a. Existing Problem

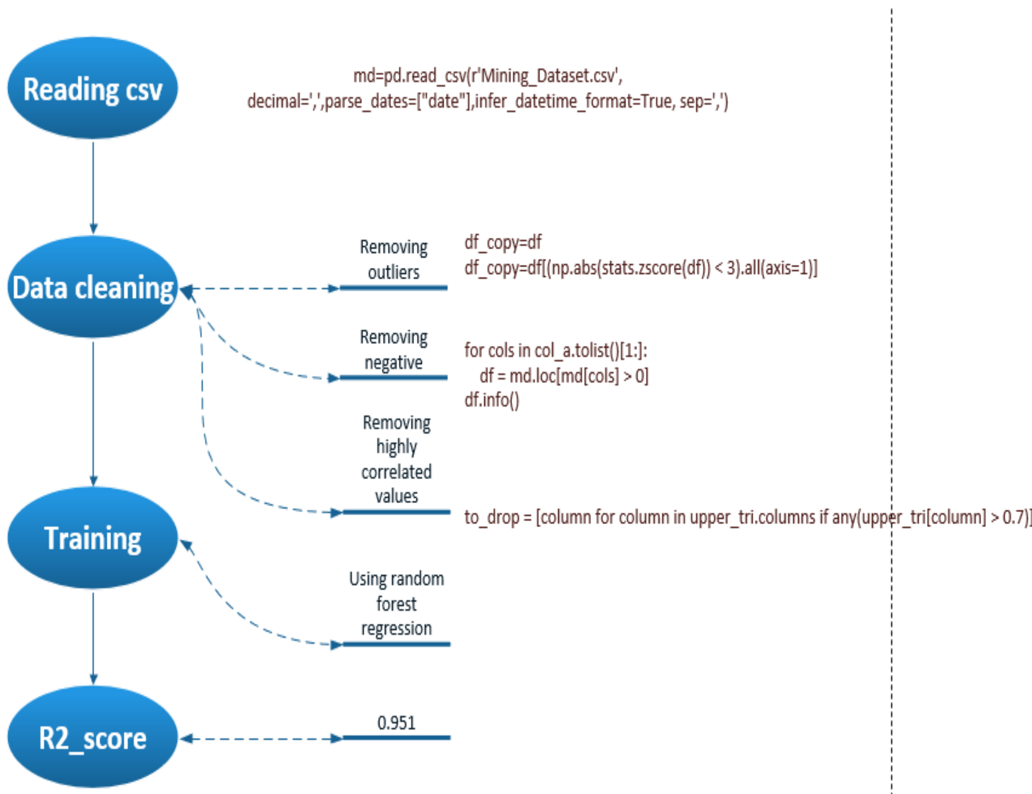
Current methods of determination of impurities in iron ores prove to be inefficient. To precisely determine the concentration of compounds such as silica, significant amounts of chemical compounds, as well as energy and time, are being extinguished. This is an important problem to resolve in the modern world, where the demands of quick and efficient production of good quality metals need to be met. Hence, a proper solution is required to enhance the process of good quality metal extraction and thus, a better way of impurity determination is needed for faster impurity extraction.

b. Proposed Solution

A better method of determining the percentage of impurity accurately, via the use of machine learning, is proposed. Using a massive dataset of concentrations of silica (%) in various iron ores, a model is built, trained and tested, which can analyse a given iron ore and accurately predict the amount of silica in the given ore. Various Regression/Classification algorithms are available to train the model. A basic application is created using *HTML* and *Python* coding, which is used to take in the inputs of the iron ore that are to be analysed, such as the density of the ore and air flow through the ore. The model is then successfully able to predict the amount of silica present in the iron ore with high accuracy, using the given input data. Basic python coding is done for exploratory analysis of the dataset used, and IBM Cloud is used to deploy and train the created Machine Learning model.

3. Theoretical Analysis

a. Block Diagram



b. Hardware/Software Design

Hardware Requirements: A Computer/Laptop.

Software Requirements:
1. Python Programming Language
2. Python packages - Numpy, Pandas, Matplotlib, Scikit-learn, Flask
3. Jupyter Notebooks
4. IBM Cloud

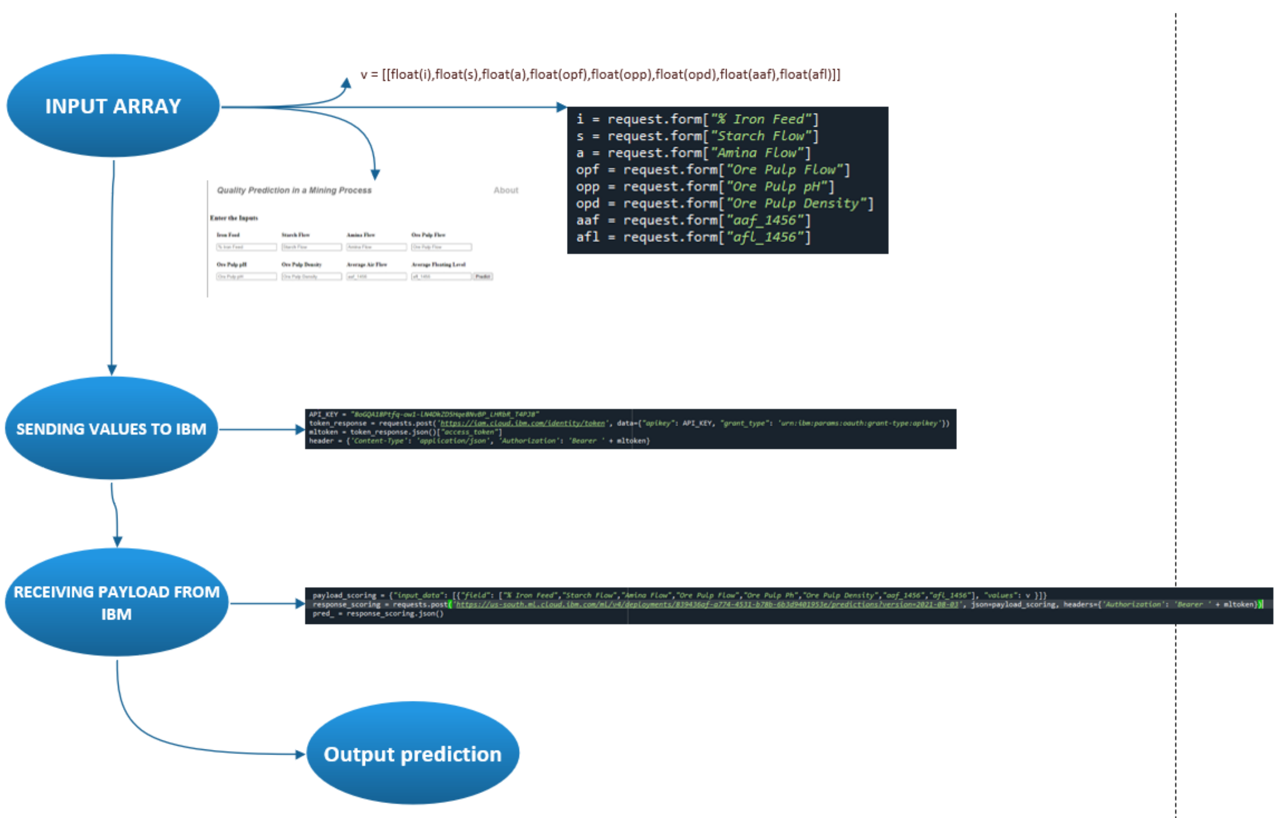
4. Experimental Investigations

a. The first step we took was to train the raw data, i.e., we removed outliers and negative values. After that we realized that the model was over fitted as the r2 score was 0.99.

b. We observed that some of the columns were highly correlated. In order to remove them, we just added a small filter that removed all the variables with high correlation (>0.7).

c. After doing the above, we obtained an r2 score of 0.95 and an rmse of 0.24.

5. Flowchart



6. Results

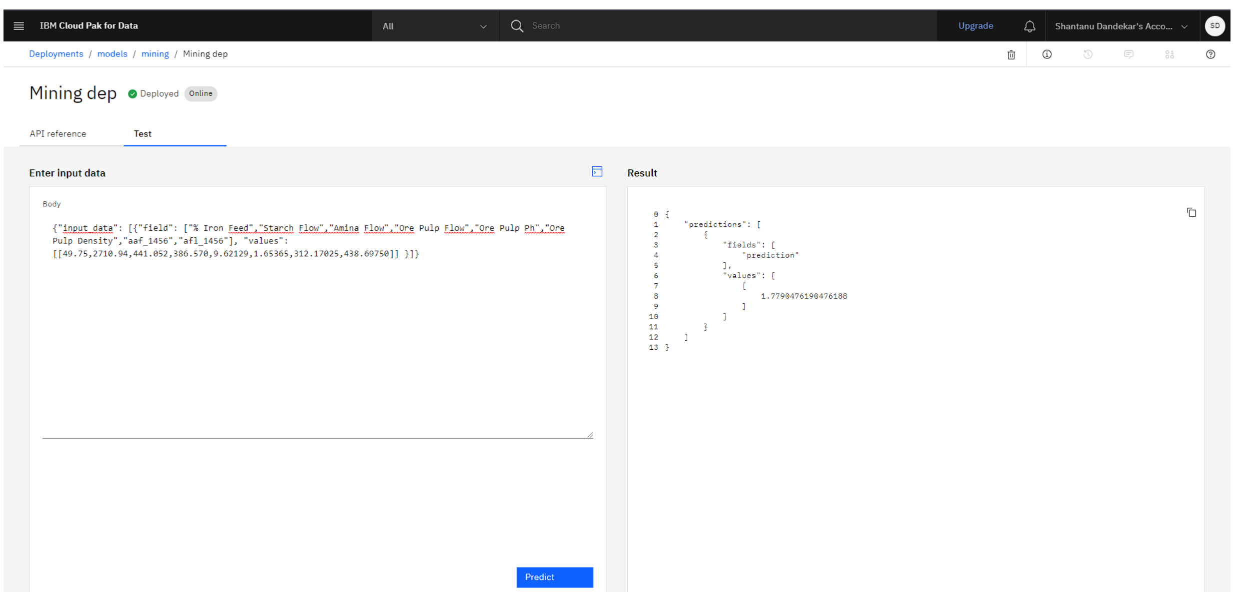
We have successfully designed our required model and trained it using IBM Cloud platform. The following images illustrate our progress and work done in building and testing the model:

```
In [27]: r2_score(y_test,y_pred)
Out[27]: 0.9502979067184666

In [28]: mse=mean_squared_error(y_test,y_pred)
         rmse=np.sqrt(mse)
         rmse
Out[28]: 0.24833201809608352

In [29]: pickle.dump(model,open('mining.pkl','wb'))

In [30]: model.predict([[55.20,3033.69,558.167,400.254,10.06970,1.74000,275.42050,453.09775]])
Out[30]: array([1.32547619])
```



7. Advantages and Disadvantages

Listed below are the various advantages and disadvantages of using our method over conventional methods for impurity detection in metal ores:

Advantages

1. Prediction of quantity of impurities (%) with high accuracy.
2. A well trained model reduces time consumption that generally occurs in conventional processes. Production of results usually takes up a great amount of time, and the process of measuring silica concentrate is stopped or delayed to test the impurity with various chemicals. This appears to be a continuous batch process, where raw material is fed into a flotation system, processed, removed, and the process repeated.
3. The use of Machine Learning eliminates such a process which consists of delays and breaks, and thus greatly improves time efficiency of impurity detection.
4. Since the prediction in our method does not require the use of various chemicals for testing the impurities, our work proves to be cost efficient in comparison to conventional methods. The cost is minimal as most of the impurity determination is done simply by the model and only basic values such as the density and air flow through the ore pulp have to be measured manually.

Disadvantages

1. Accurate values of the required inputs must be calculated. Minor changes in the values of air flow, floating level and other constraints could result in incorrect values of impurity percentage in the output.
2. Even with the regression/classification algorithm that has the best accuracy, there is a maximum of approx. 95 percent accuracy, which indicates that there is always a chance that 5 percent of the iron ore extracted could be more impure than the rest.

8. Other applications

Our model can be also used for predicting various other metals' compositions and determining the impurities present in them. It can also be used to predict the amount of pure metal actually present in the ore pulp using other data related to the ore. Since our project is a model which is specifically designed to predict the amount of silica in iron ore, to use this model for other applications would involve the training of the model using other datasets which are related to the application needed. Thus the model would have to be rebuilt to suit the specific needs and trained accordingly for other applications.

9. Conclusion

We have successfully designed a model that can accurately predict the amount of silica impurities in a metal ore. The model has proved to be highly accurate and quick in determining the percentage of impurities present in the given input ore.

10. Future Scope

The model designed in our project is an extremely versatile model, as it works based on prediction using regression or classification. Now, the model designed here is specifically trained to recognise and predict the metal ore's impurity (%) quantity. This can be expanded for future aspects to calculate the quantities of other impurities, or rather calculate the percentage of pure metal which can be extracted. The work can be further expanded to include the estimation of the rate of production of these metals which can help in the annual or periodic analysis and monitoring of metal production.

11. Bibliography

- Kaggle

12. Appendix

a. SourceCode-
[smartinternz02/SI-GuidedProject-4920-1627464653 (github.com)]

b. UI Output (Screenshot)