INTRODUCTION TO DATA SCIENCE

Homework 10: Project Report

Project Title: Drinking Water Quality Prediction (Project A2: Kaggle)

Team: Neha Sharma, Muhammad Ali, Victoria Chinonyerem Udemezue

GitHub: data_science_project_2022_ A2_group

Task 2. Business understanding

- ### Business Goals Identification

This project on “Drinking water quality prediction” was adopted from Kaggle, an online platform where data analysts and machine learning enthusiasts can collaborate with each other to find and share datasets as well as compete to solve real-life problems using data science. For this project, we have joined one of the Kaggle competitions. It is essential to state that concluding this project serves as part of the requirements for completing the course on “introduction to data science,” hence instead of business goals, we have come up with other goals that, although aim to solve the project problem, also speak to our growth as data science students.

The main goal of this project is to try and develop a predictive water quality model for Estonian water inspectors. The Estonian health board has been making tremendous efforts to measure the water quality across Estonia in order to ensure that citizens get the freshest tap water. However, their need to bring water measurement to the next level and to automate the working process of the water inspectors is what we aim to solve. Also, this project serves as an avenue for us as a group to apply the knowledge we have gained from the data science course, as it will require the skills taught therein.

We selected this project because it addresses a development concern in Estonia, which, upon solved, would result in the advancement of the water management sector of the country and ensure cleaner water for the residents. We will consider this project a success when we appear in the top 50% of the Kaggle leaderboard, obtain at least 95% assessment results from the data science course, and, by extension, have the measurement model considered for implementation by the Estonian Health Board.

- **Situation Assessment**

  Working on this project for successful delivery is a team of three students and a teacher available for supervision and guidance when necessary. We have access to the project overview and competition data on Kaggle. We also have all the data science course materials, including lecture videos and slides, available for reference and implementation.

  We are required to work on the project as a team within a timeline of approximately 1.5 months. We have two-time constraints (deadlines), which include submitting the concluded project by the 9th of December 2022 on Kaggle and presenting the project on the 15th of December 2022 for course assessment. The potential risks in this project will be the inability to complete the project before the deadlines or not attaining the top 50% on the Kaggle leaderboard. However, we believe that these risks can be mitigated by maximizing the allotted timeline for the project and affording time to review the models to make necessary corrections and increase accuracy. Most of the terms used in this project are terms that we have been familiarized with during the data science course. Hence, there will be no need to define them. In terms of cost, the only one we will incur during this project would be our time which is estimated to be about 80 hours. However, our significant benefits would be an excellent grade at the end and an opportunity to put into practice all the knowledge we have gathered so far. Winning the presentations will be an added perc.

- **Data-mining Goals**

  Our data-mining goal is centered around identifying the feature(s) that affect drinking water quality the most and subsequently designing a model that best predicts the quality of drinking water in Estonia with optimal accuracy. We will report our findings and do a presentation using poster. The Kaggle evaluation will serve as the success criteria for our data-mining goal.

Task 3. Data understanding

- **Collect initial data:**

  Data were collected from the Kaggle competition page[data]. The information is missing from the owners of this competition about how and from where they collected the data. The data has been split into two groups: *Training set(train.csv)* and *Test set(test.csv)*. The training set will be used to build our machine-learning model. It has both features and the result variable. The Test set will be used to see how well our model performs on unseen data. It contains all the features which are present in the train set except the result variable, which we have to predict. There is also a *sample_submission.csv* file which is an example of what the submission in the competition should look like.

- **Describe the data:**

The shape of the training data is (440,58), meaning there are 440 observations, 57 features, and 1 result variable. Similarly, the shape of the test data is (189,57), meaning there are 189 observations, 57 features, and no result variable(We have to predict it). The total size of the data is 104.53 kB.
The features include various numeric metrics (e.g., color, smell, ph-level, coli-type bacteria) from previous water tests that are used by the inspectors to define whether the quality is compliant with the regulation. For each water station, we have information about the test results of the years 2019 and 2020. The corresponding year number is added at the end of each variable name. In addition, we have information about the compliance of the station in 2019 (*compliance_2019*) and 2020 (*compliance_2020*). The result variable is *compliance_2021.*

- **Explore the data:**

This section of the report will give a brief idea about the training dataset. As we know, the training dataset from the train.csv file contains 440 observations, 57 features, and 1 result variable named compliance_2021.
The result variable contains binary values (0,1) where 1 represents the water quality that is compliant with the regulation, and 0 means it's not. There are a total 374(85%) “0 class” and 66(15%) “1 class”. These values show that our prediction classes are highly imbalanced, which could cause some bias/overfitting/underfitting issues in our prediction model.
All 57 features are numeric (int and float data type). Feature “station_id” has 440 unique values representing the water stations. There are features for each type of mineral and chemical present in the water for a given water station, such as “aluminum, boron, chloride, coli-like-bacteria, fluoride, iron, manganese, etc.” these features are present for the year 2019 as well as 2020. There are also 2 features present for the years 2019 and 2020 compliance.
Compliance features(including the result variable) and station_id do not have any missing values. The remaining features contain null values. where almost 30 features have more than 50% missing values. This could be an issue for our model as we need to fix these null values by some methods, and missing a lot of data is not a good sign.

- **Verify data quality:**

From the previous section, we believe that data quality is not excellent. The reasons are:

  - Data is very small (can be an issue for inadequate training of our model as there is not enough data for the model to train)

- Data has a lot of missing values (almost all the features have missing values, and 30 features have more than 50% of missing data). Data is already tiny; then there are a lot of missing values.
- There is no metadata for further information. For example, there are 57 features but no explanation of what they mean.
- Prediction classes are highly imbalanced (85%:15%).

Overall The data quality is not ideal.

Task 4. Planning your project

This section introduces the roadmap for the project. The roadmap is as follows:

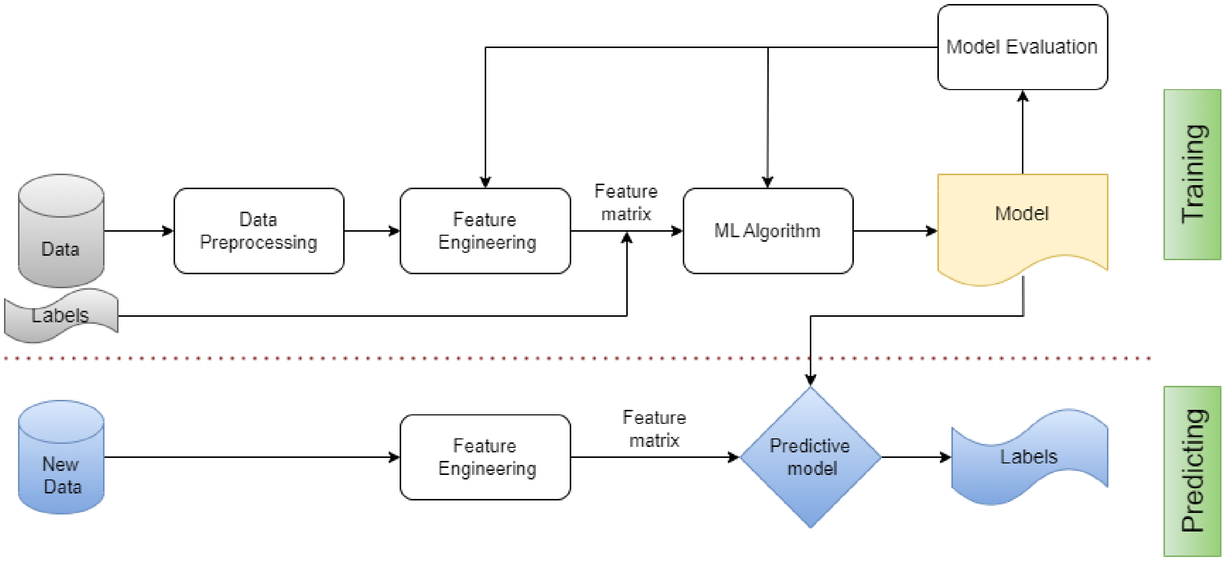

Figure 1: Project Roadmap

- ### Data Preprocessing:

  This step is a crucial part of any predictive modeling. Data Preprocessing will help us to understand our data and its features better. This step includes exploratory data analysis, data visualization, data cleaning, and data scaling. Basically, users have to understand the data and make it readable for the machine learning model. As we know that our data is imbalanced, we will try some overfitting/underfitting techniques to reduce the biases.

- ### Feature Engineering:

  Feature engineering refers to the process of designing artificial features into an algorithm. These artificial features are then used by that algorithm in order to improve its performance or, in other words, reap better results. A few techniques for feature engineering are as follows: one-hot encoding, scaling, log transformation, etc. Machine Learning Algorithm:

After preparing our data for the machine learning model, it will be time to train our models. As this problem is a classification problem, so we will be using classifiers from sklearn library. We will start from the baseline model: logistic regression. We will move forward with tree algorithms and then boosting methods. If we have enough time, we can try some deep-learning algorithms for classification. We will also tune the hyperparameters of our models for better performance.

- **Model evaluation and testing:**

  Training a model is not sufficient. We need to evaluate it based on a proper matrix. Based on our problem, We will choose our evaluation matrix from recall, precision, or f1 score.

  The final step would be testing our model on the test data set and making submissions in the Kaggle competition.

The goal is to spend at least 20 hours per person on this project. Also, there are going to be frequent group discussions and meetups. As all team members agreed to work together on this project and help out each other, it can be assumed everyone has equal responsibility and load for this project. Also, The project is going to be done in Jupyter notebook using python.

Link for the competition: Competition
Link for the GitHub repo: github