Risk Data Masters

> *Rules:* The projects will be done by groups of 4 students. You will need to upload 3 files (the scripts, the requirements and the csv output). The script for the projects will be named exclusively `predict_risk.py`. A requirement file (`requirements.txt`) is also demanded. Requirement files can be generated automatically (your task is to find out how). A `requirements.txt` sample is provided at the end of this document. The csv output file must be named `predictions.csv`. Any change in file names will be penalized. Other files besides the script, the requirements file and the csv file will not be considered. The files must be uploaded on the platform before the deadline. The deadline will be clearly marked on the upload section created for this project. Any delay will be sanctioned with 0/20 for the missing uploads.

In this project, you are tasked with the role of a financial data analyst who must compile training data from various sources to create a comprehensive dataset.

In the realm of financial analytics, the '**Risk Data Masters**' project presents a compelling challenge that requires the application of machine learning to predict potential fiscal distress. The dataset, a meticulous compilation of financial metrics across various companies, offers a window into the intricate world of corporate finance. Each record encapsulates an array of financial indicators: revenue, expenses, profit, employee headcount, and several key ratios that signify a company's leverage, valuation, and market presence. The dataset's cornerstone is the 'Risk' classification—a binary target that signifies the likelihood of a company's financial hardship.

The dataset comprises variables including 'Company ID', 'Revenue', 'Expenses', 'Profit', 'Employee Count', 'Debt-to-Equity Ratio', 'Price-to-Earnings Ratio', 'Research and Development Spend', 'Market Capitalization', and 'Credit Rating'.

The train data can be found across different data sources: 'Company ID', 'Revenue', 'Employee Count', 'Credit Rating', and 'Risk' are stored in a MongoDB collection; 'Company ID', 'Expenses', and 'Research and Development Spend' reside in a

MySQL database table; and 'Company ID', 'Profit', 'Debt-to-Equity Ratio', 'Price-to-Earnings Ratio', and 'Market Capitalization' are presented in an HTML table available at https://h.chifu.eu/data.html.

The common thread linking each data source is the 'Company ID' column, which you will use to join or merge data across different storage systems—MongoDB, MySQL, and an HTML page.

The variable description is here:

- `Company ID`: A unique identifier for each company within the dataset.
- `Revenue`: The total revenue of the company, reflecting the gross income generated from sales or services.
- `Expenses`: The total expenses incurred by the company, encompassing costs such as operating expenses, cost of goods sold, and other outlays necessary for maintaining business operations.
- `Profit`: The net profit of the company, calculated as the difference between total revenue and total expenses.
- `Employee Count`: The number of employees working for the company, indicating the size of its workforce.
- `Debt-to-Equity Ratio`: A measure of a company's financial leverage, calculated as total debt divided by shareholder equity, indicating how much debt is used to finance assets relative to the value of shareholders' equity.
- `Price-to-Earnings Ratio`: A valuation metric comparing a company's current share price to its per-share earnings, often used to gauge whether a stock is overvalued or undervalued.
- `Research and Development Spend`: The amount of money that a company spends on research and development activities, showcasing the company's investment in innovation and future growth.
- `Market Capitalization`: The total market value of a company's outstanding shares, representing the public opinion of a company's net worth and is calculated by multiplying the current stock price by the total number of outstanding shares.
- `Credit Rating`: An assessment of the creditworthiness of a corporation, reflecting the company's ability to repay its debts and the risk of default.
- `Risk` (target variable): A binary indicator representing the classification outcome, where '1' signifies a high risk of financial distress and '0' indicates a low risk. This

variable is the one that models will attempt to predict based on other financial indicators.

Connection details for the data sources are as follows:

- MongoDB server: 208.87.130.253:27017, database: mag1_project, authentication database: mag1_project, collection: project, username: mag1_student, password: Gogo1gogo2.
- MySQL server: 144.24.194.65:3999, database: mag1_project, table: project, username: mag1_student, password: Gogo1gogo2.
- HTML table: https://h.chifu.eu/data.html

Your primary objective is to train a supervised classification model using the compiled training data. This model should be capable of predicting the 'Risk' of financial distress for unseen companies, as represented by the test data located at https://h.chifu.eu/final_test.csv. Remember, the test data does not include the 'Risk' column; it is your model's job to predict these labels.

After training your model, you must evaluate its performance on the training data by generating a classification report, which provides key metrics that give insight into the accuracy and robustness of your model. These metrics include precision, recall, f1-score, and support for each class. The classification report will be displayed on screen, when running the python script.

Here is an example of what a classification report might look like:

```
              precision    recall  f1-score   support

           0       0.95      0.98      0.97       150
           1       0.89      0.75      0.81        28

    accuracy                           0.94       178
   macro avg       0.92      0.86      0.89       178
weighted avg       0.94      0.94      0.94       178
```

In this example, class '0' represents low-risk companies, and class '1' represents high-risk companies. Precision is the ratio of correctly predicted positive observations to the total predicted positives. Recall (or sensitivity) measures the ratio of correctly predicted positive observations to all actual positives. The f1-score is a weighted harmonic mean of precision and recall, and the support is the number of actual occurrences of each class in the specified dataset.

Your final deliverable will be a CSV file containing the 'Company ID' and the predicted 'Risk' labels for the test dataset. You will also provide the classification report for the training data as part of your model evaluation. Here is an example of the CSV output:

```
Company ID,Predicted Risk
100001,0
100002,1
100003,0
100004,0
100005,1
100006,0
100007,1
100008,0
100009,1
100010,0
...
```

Your script must accept the output file as an argument using the `argparse` module. The output file should contain the predicted labels for the test dataset along with the corresponding company IDs.

Here is an example run command for your script:

```
python predict_risk.py --output predictions.csv
```

A list of useful modules for this project includes:

- pandas
- numpy
- scikit-learn
- pymongo
- sqlalchemy
- pymysql
- argparse
- requests
- BeautifulSoup (bs4)

These modules provide a range of functionalities from data manipulation to machine learning and database connectivity, which will be instrumental in accomplishing the tasks required for the 'Risk Data Masters' project.

A `requirements.txt` sample:

```
pandas==1.3.5
numpy==1.21.4
scikit-learn==1.0.1
pymongo==3.12.1
sqlalchemy==1.4.27
pymysql==1.0.2
argparse==1.4.0
requests==2.26.0
beautifulsoup4==4.10.0
lxml==4.6.3
html5lib==1.1
```

Good luck, and may your data insights be profound and your models robust!

> 💡 Bonus points will be awarded for the best performing model.

> 💡 Bonus points may be awarded for relevant improvements and extra features.