LAMARR
Institute for Machine Learning and Artificial Intelligence

UA RUHR CENTER FOR TRUSTWORTHY DATA SCIENCE AND SECURITY RESEARCH ALLIANCE

technische universität dortmund

Case Study: Automated Data Analytics Using LLMs

Simon Klüttermann, Emmanuel Müller

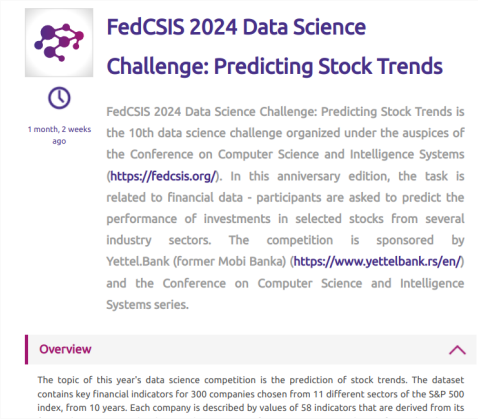
17.07.2024

TU Dortmund University - Department of Computer Science

Data Mining Cup

Every Winter Semester: Data Mining Cup

Solve Data Mining Challenges together & compete against other groups



”My Students” Code

Last Week I was talking to a thesis student

And he wanted to change a plot a bit

To do this, he copied all his code into Gemini (Google Chatgpt) and let it change it

```python
import json
import matplotlib.pyplot as plt


with open('results_per_algorithm.json', 'r') as f:
    data = json.load(f)


datasets = list(data['results'].keys())
percent_corr_values = [data['results'][dataset]['percent_corr'] for dataset in datasets]


plt.figure(figsize=(10, 6))
plt.scatter(datasets, percent_corr_values, color='blue', marker='o')


plt.xlabel('Dataset - Removed Algorithm')
plt.ylabel('Percentage of Combined AUC > 0.5')
plt.title('Scatter Plot of Percentage Values for Each Dataset and Removed Algorithm')
plt.xticks(rotation=90)
plt.grid(True)
plt.tight_layout()


plt.savefig('plots/scatter_plot.png')


plt.show()
```

My Case Study

So, do I even need the student?

Or how much of his work can we automatize here?

Let's find out together!

In this Case Study: Try to solve Data Mining Challenges as automatically as possible

Searching for about 12 Students

Separate the task into subtasks

- Data handling
- Exploratory data analysis
- Feature Engineering
- ...

See how much of this a LLM can do

Possibility to write (and present) a research paper

Any Questions?

Write me: Simon.Kluettermann@cs.tu-dortmund.de