Evaluation of Modern Tools for Data Scientists

Jozef Ivančo
Master of Management in Data Science
Hasselt University
Supervised by Prof. dr. Koenraad Vanhoof

INTRODUCTION

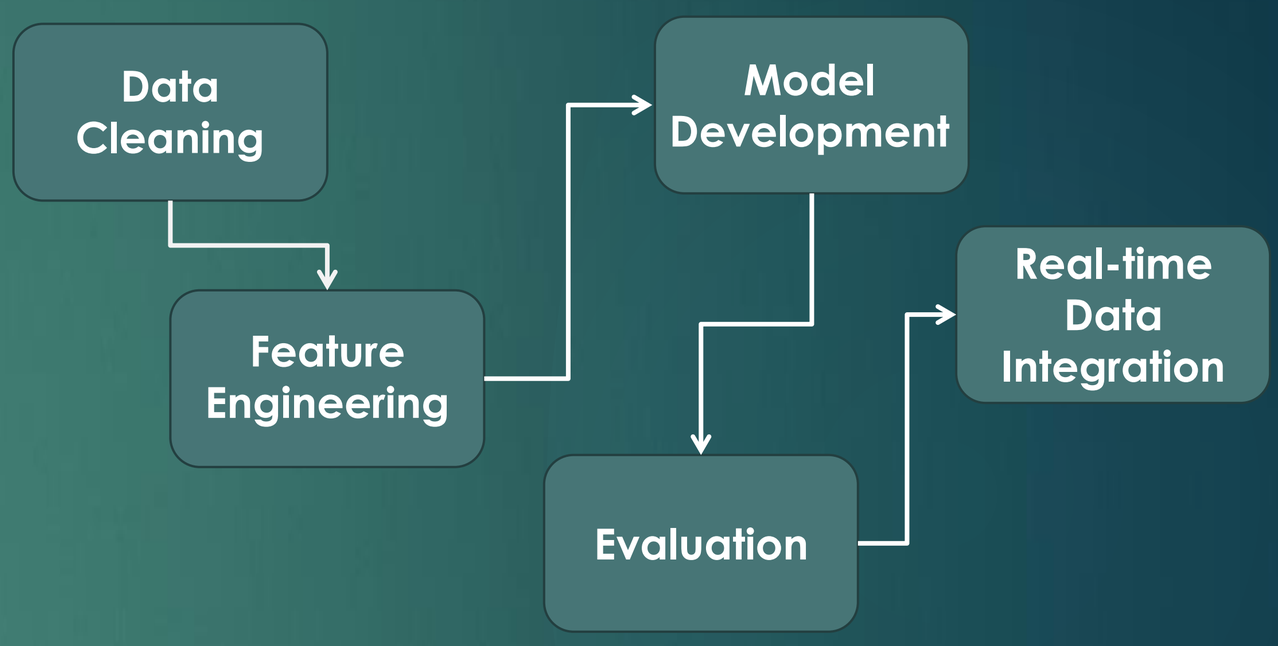
The increasing adoption of **artificial intelligence (AI)** across industries has created opportunities to democratize complex **data science workflows.** However, building machine learning models for **flight delay prediction** remains challenging for **non-expert users** due to the technical skills required. This research investigates whether **AI chatbots** like ChatGPT, DeepSeek, Google Gemini, Microsoft Copilot, and Meta AI can assist non-experts in developing predictive models using **real-time weather** data. The study assesses usability, effectiveness, and contextual understanding of these chatbots in guiding the development process from data cleaning to model evaluation.

OBJECTIVES

- To evaluate the usability, contextual intelligence, and performance of these AI chatbots in predictive modelling.
- To examine the feasibility of integrating real-time data streams for improved model accuracy.
- To assess the interpretability and reliability of chatbot-guided model development in high-stakes domains like aviation.

METHODS

- **Chatbots Evaluated**: ChatGPT, DeepSeek, Google Gemini, Microsoft Copilot, and Meta AI.
- **Task**: Assist non-expert users in building predictive models for flight delay forecasting using real-time weather and air traffic data.
- **Analysis Criteria**: Usability, contextual memory, error handling, and output quality.
- **Environment**: Python-based, visualized in Visual Studio Code.



RESULTS

Chatbot	Data Cleaning Ability	Classification Performance	Regression Performance	Error Handling	Usability & Memory	Friendliness	Conversation Evolution
ChatGPT	Automated with feature selection, missing value handling, and merging	Accuracy: 61.72%, Macro F1: 0.41, Bias towards class 2, iterative improvement	MAE: 20.15, MSE: 1481.66, R^2: 0.145	Excellent, interactive debugging with iterative improvement	High contextual awareness and memory within session	Very friendly and step-by-step	Adapted and evolved with feedback
DeepSeek	Basic dropna, failed to clean full dataset, required external help	Failed to complete classification, memory issues	Did not complete	Lost track during debugging, confusing replies	Verbose, lacks continuity in responses	Neutral but overwhelming	Fragmented and lost memory
Google Gemini	Partial, hallucinated columns, caused errors, high memory usage	Accuracy: 99.88% (overfit), Macro F1: 0.73, Precision 0.44 for class 0	MAE: 10.35, MSE: 1636.59, R^2: 0.98 (overfit)	Provided faulty code multiple times, context lost easily	Low, limited to short prompts, easily forgets context	Fast but impersonal	Quickly forgot earlier context
Microsoft Copilot	Incorrect assumptions, required manual column selection	Failed to deliver classification model	MAE: 19.67, R^2: 0.20	Handled some errors but needed manual correction and resetting	Poor memory, frequent restarts needed	Mechanical and slow	Disconnected conversations
Meta AI	Not capable; provided unusable code with syntax errors	Not evaluated due to input length limits and syntax errors	Not supported due to interface limits	Failed due to token limit, could not handle long inputs	Extremely limited message length, no continuity	Courteous but unhelpful	No real conversation possible

- **ChatGPT**: Outperformed others with strong usability, contextual understanding, and iterative improvement.
- **DeepSeek**: Struggled with memory gaps, repetitive errors, and inability to maintain workflow continuity.
- **Google Gemini**: Achieved high accuracy in metrics but was prone to overfitting and hallucinated data.
- **Microsoft Copilot**: Demonstrated some integration potential but struggled with data size limits and failed to build classification models effectively.
- **Meta AI**: Limited by input size constraints, syntax errors, and poor conversation evolution, making it impractical for complex workflows.

CONCLUSION

- **AI chatbots** (like ChatGPT) can assist non-experts in building predictive models, making complex data science tasks more accessible.
- **Best Performance**: ChatGPT demonstrated strong usability, contextual understanding, and iterative improvement.
- **Limitations**: Other chatbots struggled with memory gaps, hallucinations, and inconsistent outputs.
- **Real-time Predictive Modelling**: Possible with chatbot guidance, but human oversight is essential to ensure accuracy and reliability.
- **Takeaway**: AI chatbots are promising companions in democratizing data science, though not yet a full replacement for expert intervention.

REFERENCES