**KHARAGPUR DATA SCIENCE HACKATHON REPORT**

# **Automated Publish-ability Prediction and Conference Selection for Research Papers**

## Team Name:

Data Wizards

## Event:

KDSH Hackathon 2025 - Round 2

## Submission Details:

Date: January 14, 2025

## Team Members & Contact Information:

* Aditi Chaudhary – Team Lead, Report and Presentation Designer, Data Preprocessing Specialist

Email: aditi2310003@akgec.ac.in  
Phone: +91-7011701032

* Divya Pal – Machine Learning Engineer, NLP Expert

Email: divyapal065@gmail.com  
Phone: +91-7017646327

### Abstract

The Kharagpur Data Science Hackathon presented a challenge to develop an AI-driven system for automating the process of research paper evaluation and conference selection. The first task involved the classification of research papers as "Publishable" or "Non-Publishable" based on content analysis, while the second task required the recommendation of the most suitable conferences for the "Publishable" papers. Our team successfully completed Task 1 by implementing two models: Random Forest and BERT, achieving good accuracy with the former and decent results with the latter. Despite the success in the first task, Task 2, which focused on conference selection using Pathway resources, presented significant challenges. Issues in real-time data streaming, integrating Pathway connectors, and implementing the Pathway framework for conference selection impacted the overall progress. This report provides an overview of the methodologies employed, achievements, and challenges faced during the hackathon.

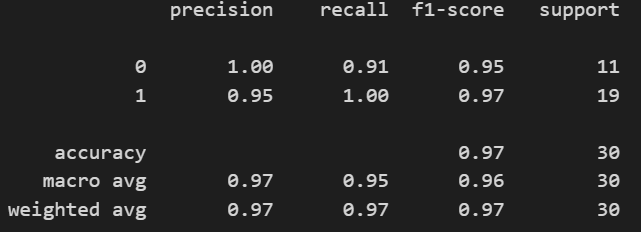
# Introduction

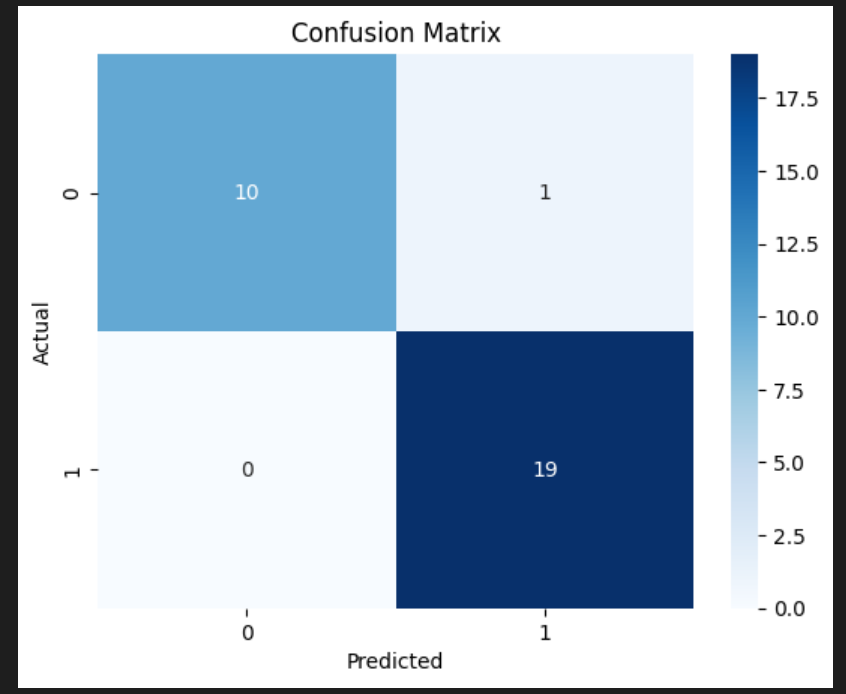
Pathway Technology Inc. Provided us the opportunity to participate in the Kharagpur Data Science Hackathon 2025. The challenge involved using Pathway's Python framework to automate the evaluation of research papers and determine their suitability for publication, as well as to select the most appropriate conference for each paper.  
This report discusses our approach towards solving the two primary tasks: Research Paper Publish-ability Assessment and Conference Selection. Our team focused on developing an AI-driven solution leveraging machine learning models like Random Forest and BERT, as well as the Pathway framework for efficient data processing.

# Task 1: Research Paper Publishability Assessment

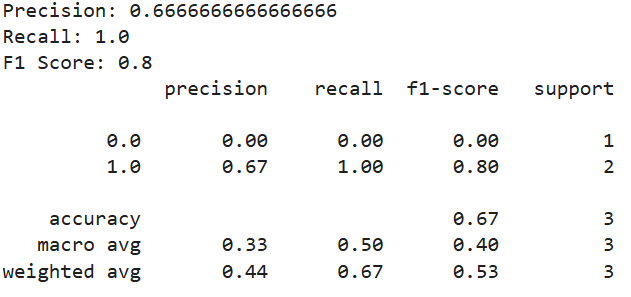
In Task 1, our goal was to classify research papers as either 'Publishable' or 'Non-Publishable' based on their content. We utilized Random Forest and BERT models to achieve this. The Random Forest model yielded excellent accuracy, while the BERT model, although producing a more contextual understanding, provided decent results.  
To classify the papers, we focused on the following key features:  
- Methodology suitability  
- Logical coherence of arguments  
- Adequacy of evidence and validation of claims  
The model was trained using a dataset of 150 research papers, with 15 labeled papers serving as a reference.  
  
**Accuracy and Evaluation Metrics:**

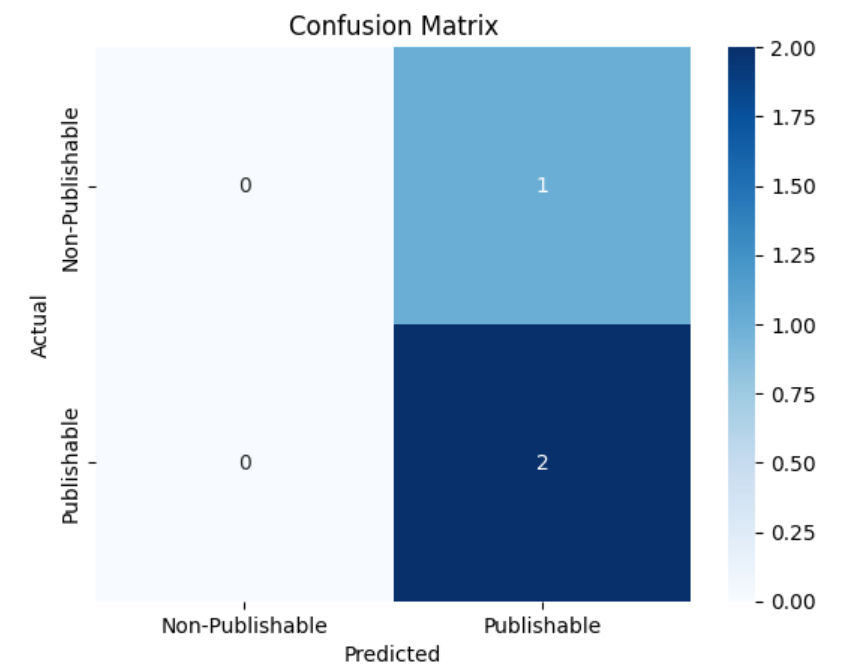
**Random Forest Accuracy: 0.9666666666666667**





**BERT Accuracy:** **0.6666666666666666**





# **Task 2: Conference Selection**

Task 2 aimed to develop a framework for conference selection, where the system had to evaluate the research papers classified as 'Publishable' and suggest the most appropriate conference for each paper. However, implementing this task was more challenging, as it required accurate comparative analysis between the submitted paper and the reference papers from various conferences. We also faced challenges in using Pathway’s resources, such as data connectors and integration with the Pathway framework, which hindered progress in completing this task.  
Despite this, our team was able to establish a basic structure for conference selection, but the integration of the Pathway connector and vectorstore for real-time data streaming was a work in progress.

# Challenges Faced

1. Pathway Integration: The primary challenge was the integration of Pathway's data streaming and vector store capabilities. While the documentation was helpful, the real-time data processing required careful handling, which slowed down the development process.  
2. Conference Selection Logic: The logic for matching research papers to conferences was complex. Comparing subject matter, methodology, and findings with conference profiles required a deep understanding of each paper's nuances, which was not trivial.  
3. Limited Time for Task Completion: Given the constraints of the hackathon, time was a significant limiting factor in fully implementing Task 2.

# Solution and Approach

For Task 1, we employed a machine learning pipeline that involved data preprocessing, feature extraction, and model training. The Random Forest classifier was chosen for its robustness and ability to handle the feature set effectively, while BERT was explored for its capacity to handle natural language more comprehensively. We followed a typical supervised learning approach with cross-validation to ensure reliable performance metrics.  
  
For Task 2, we initially attempted to integrate the Pathway framework to streamline the selection process using a comparative analysis of papers. However, due to technical issues with the Pathway system, this task was not fully implemented. We focused on the basic logic of paper-to-conference matching but faced challenges in effectively using Pathway connectors for real-time data analysis.

# Conclusion

In conclusion, our team successfully completed Task 1 on Publishability Classification, achieving good results with Random Forest and decent results with BERT. While Task 2, Conference Selection, remains unfinished due to integration challenges with Pathway, we made significant progress in developing a foundation for the framework. The experience provided valuable insights into both AI-based classification techniques and real-time data streaming systems, offering a learning opportunity for further enhancement of the system in future iterations.