ML Mini Project

Team Member 1 : SHIVAM ANAND
Team Member 1 (SRN) : PES2UG23CS549

Team Member 2 : SHASHWAT SOLANKI
Team Member 2 (SRN) : PES2UG23CS545

Title : Predicting Conference Paper Acceptance

Problem Statement:
Predicting whether a research project will be accepted by a conference is often subjective and time-consuming. This project aims to automate the process by developing a web application that predicts the likelihood of acceptance for major NLP and AI conferences (ACL, CoNLL, ICLR) or rejection, based on a project's title and introduction.

Approach:
We formulated the problem as a multi-class text classification task. The dataset consists of research papers labeled by conference or rejection status. Titles and introductions were combined and preprocessed by converting to lowercase, removing numbers and punctuation, and normalizing whitespace. TF-IDF vectorization was used to convert text into numerical features suitable for modeling.

A dense neural network with two hidden layers (256 → 128 neurons) and dropout regularization was employed to predict probabilities for four classes: ACL, CoNLL, ICLR, and Rejected. Class imbalance was addressed using weighted loss during training. Early stopping was implemented to prevent overfitting.

Implementation:

Users can input the title and introduction through the web interface, and the backend returns the predicted probabilities for all classes.
Example prediction:

```
{'000': 0.12, '001': 0.68, '010': 0.05, '100': 0.15}

000 - Rejected From Conferences
001 - ICLR
010 - CoNLL
100 - ACL
```

The model was implemented using **TensorFlow/Keras** as a dense neural network classifier with two hidden layers (256 and 128 neurons) and dropout regularization to prevent overfitting. Text data was vectorized using **TF-IDF** from **scikit-learn**, then fed into the network for training. The dataset was split 80-20 for training and testing, and **class weights** were applied to handle imbalance. The model was trained with the **Adam optimizer** and early stopping for efficient convergence, achieving strong generalization performance.

Challenges Faced:
The main challenge was collecting and preparing a suitable dataset for training. Since no direct dataset existed for conference-specific paper classification, the data had to be parsed, cleaned, and labeled manually from multiple sources before training the model.

Conclusion:
The project successfully demonstrates how machine learning can predict the likelihood of research paper acceptance using textual features. It provides a practical baseline for automated conference prediction and can be further improved with larger datasets and advanced NLP models.