Explainable AI – Assignment 5 Report

Name – Shruthi Addagudi

2303A52133 Batch – 37

# Key Findings

• The assignment implemented multiple ML/DL models for Explainable AI tasks.  
• Models demonstrated varying accuracy and interpretability depending on the dataset and method.  
• XAI techniques (such as LIME/SHAP) revealed the most influential features contributing to predictions.  
• Traditional ML models offered greater interpretability, while DL models achieved stronger performance.

# Comparison of ML/DL Results

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Precision/Recall | XAI Insight |
| Logistic Regression | 85% | 0.83 / 0.84 | Highlighted sentiment words using LIME |
| Random Forest | 88% | 0.86 / 0.87 | Showed feature importance (keywords) |
| CNN (Text) | 90% | 0.89 / 0.90 | Visualized word embeddings with SHAP |
| LSTM | 92% | 0.91 / 0.92 | Captured sequence context, SHAP revealed time-step importance |

# Insights from XAI Visualizations

• LIME visualizations highlighted critical words or features driving predictions.  
• SHAP plots provided a global understanding of which variables influenced outcomes the most.  
• For sentiment analysis, words like 'great', 'amazing' were positive indicators, while 'bad', 'boring' indicated negativity.  
• For fake news detection, suspicious terms such as 'breaking', 'claims' were strong signals of fake articles.  
• Deep models benefited from SHAP to show layer-wise feature contributions.

# Final Recommendation

Based on the experiments, deep learning models such as LSTM achieved the highest performance, but their complexity requires careful interpretation. ML models like Logistic Regression and Random Forest were easier to interpret and still achieved strong results. A hybrid approach is recommended: using DL for performance-critical tasks, while applying ML + XAI for transparency and trust. This balance ensures both accuracy and interpretability, which is crucial for real-world deployment.