💼 CELEBAL TECHNOLOGIES PVT. LTD.

🧑‍🎓 DATA SCIENCE INTERN – BATCH 1

📂 PROJECT

📊✨ Project Insights: Spam Email Classification Using Machine Learning

---

🔍 1. Feature Correlation Analysis

- A deep correlation study showed that only the **top 29 features** had a strong direct correlation with the target (Class).
- However, when the model was trained using **only these 29 features**, the **F1 score and accuracy dropped** significantly.
- In contrast, using **all 57 features** yielded much better performance — clearly indicating that even weakly correlated features were contributing valuable information to the model.

🧪 2. Data Quality

- ✅ **No missing values** in the dataset — ensuring a clean and complete foundation for training.
- 📉 Outliers were analyzed using **Z-score technique**:
  - Outliers were present in most features but **not severe enough** to harm model performance.
  - So, **no outliers were removed**, and **no feature selection** was applied — preserving the dataset's natural structure.

🧠 3. Model Experimentation

- Several models were evaluated including:
  - **Logistic Regression**, **SVM**, **Random Forest**, **XGBoost**, and a **Neural Network (MLP)**.
- The top-performing models were:
  - ✅ **XGBoost**
  - ✅ **Random Forest**

⚙️ 4. Hyperparameter Tuning & Ensembling

- Surprisingly, **hyperparameter tuning** on Random Forest and XGBoost led to **slight decreases** in F1 and accuracy — likely due to overfitting or an already optimal base configuration.
- A **stacking ensemble** of tuned Random Forest and XGBoost was attempted but:
  - Resulted in **no significant improvement** over standalone XGBoost.
- A **1D Neural Network (Dense MLP)** was also trained, but it did **not outperform** XGBoost on tabular data.

🚀 5. Final Model Choice & Deployment

- Based on performance, simplicity, and interpretability, **XGBoost** was chosen as the final model.
- The model was then:
  - 🧠 **Saved**
  - 🧪 **Integrated with a StandardScaler**
  - 🌐 **Deployed via Streamlit for real-time spam detection**

🖥️ 6. Streamlit App Features

- 📥 **Default values** are pre-filled for user convenience.
- 🔁 One-click button to **generate random values** for quick testing.
- 🔄 Option to **reset inputs to default** anytime.
- 🚀 A single **Predict** button:
  - Instantly tells you if the email is **Spam** or **Not Spam** 📪
  - Displays the 📊 **spam probability score** to show model confidence

---

✅ Conclusion

By leveraging all available features, avoiding unnecessary filtering, and focusing on model strength, this project achieved highly accurate and reliable spam detection. The XGBoost model, combined with a clean Streamlit UI, delivers a practical and intelligent email classification tool.

🌐 Deployed Streamlit Application

🔗 Visit the live application here: https://spam-classifier-muskan2003.streamlit.app/

Prepared by: Muskan

🧑‍💻 Data Science Intern

📍 Celebal Technologies Pvt. Ltd., Jaipur, Rajasthan