Explainability Report: Phishing Website Detection using Naive Bayes & SVM

1. Models Used

Model	Key Traits
Naive Bayes (GaussianNB)	Assumes feature independence, quick to train, good baseline classifier.
Support Vector Machine (SVM)	Linear kernel used; strong generalization and works well with high-dimensional data.

2. Model Performance

Metric	Naive Bayes	SVM
Accuracy	72%	93%
Precision, Recall, F1	Detailed in classification reports.	

> *SVM generally outperforms Naive Bayes in classification metrics, especially when linear decision boundaries exist.*

3. Explainability Techniques

SHAP (SHapley Additive Explanations)

Goal: Explains the prediction of a specific instance by computing the contribution of each feature.

SHAP for Naive Bayes:

- Used `naive_bayes.predict_proba` as model function.
- SHAP values computed on 10 sampled test instances.
- **Summary Plot**:
  - Shows global feature importance.
  - High contributors: likely to include `https_token`, `request_url`, `web_traffic`, etc.

SHAP for SVM:

- Linear SVM with `predict_proba`.
- Similar sampling and background logic.
- **Summary Plot**:
  - Generally crisper due to linearity.
  - Highlights linear contributions of features like `statistical_report`, `having_at_symbol`, etc.

Insight:

- SHAP clearly showed that both models rely on overlapping sets of high-impact features.
- Engineered features (like `obfuscation_score`) contributed to model interpretability and decisions.

LIME (Local Interpretable Model-agnostic Explanations)

Goal: Explains the prediction of individual instances by approximating the model locally.

How It Worked:

- For a randomly selected test instance, LIME provided a human-readable breakdown of how much each feature influenced the prediction.
- Visual explanation displayed in notebook (assumes Jupyter support).
- Top 10 most impactful features were shown with their direction (positive/negative influence).

Observed Behavior:

- For Naive Bayes: The explanation often included classic features like `https_token`, `dnsrecord`, `request_url`.
- For SVM: Slightly sharper influence observed due to the linear kernel — directionality was clear.

🔎 **Insight**:

- LIME complements SHAP by providing **instance-level** clarity with intuitive visuals.
- Helps debug model behavior or understand *why* a particular URL was flagged.