Internet Traffic Performance Analysis Project

Wroclaw university of science and technology

Team members:

Name	ID
Mostafa Abdelmohsen	**288862**
Mohamed Issa	**288885**

1. Introduction

This project analyzes internet traffic measurement data to understand performance metrics including download speeds, server performance, and protocol efficiency (IPv4 vs. IPv6). The comprehensive analysis includes:

- Exploratory Data Analysis (EDA)
- Data visualization
- Preprocessing
- Predictive modeling
- Comparative evaluation of machine learning approaches

2. Project Structure

The analysis is organized into four Jupyter notebooks:

Notebook 1: Data Exploration & EDA (202102 Dataset)

Objective: Perform comprehensive exploratory analysis of February 2021 network performance data.

Dataset Overview:

- 19 CSV files containing diverse network metrics
- Measurements collected in February 2023
- Key categories:
    - Download speed metrics (HTTP GET tests)
    - Upload speed metrics
    - Network quality measurements (DNS, ping, traceroute)
    - Protocol-specific tests (IPv4 vs IPv6)

    - Data usage statistics

Key Analysis of HTTPGETMT Data:

- Examined multi-threaded HTTP download speeds
- Clean dataset with 724,511 records and no null values
- Converted bytes/sec to Mbps for analysis:
    - Mean speed: 204.98 Mbps
    - Median speed: 204.63 Mbps
    - Most speeds cluster around 200Mbps

Visualizations Created:

- Histograms of download speed distributions
- Temporal trends in performance
- Comparative charts of IPv4 vs IPv6 performance

Recommendations for Expansion:

- Time-of-day analysis
- Geographic performance mapping
- Multi-threaded vs single-threaded comparison
- Network quality correlation studies
- IPv4/IPv6 performance benchmarking
- ISP performance comparisons

Notebook 2: XGBoost Modeling (202102 Dataset)

Objective: Develop and evaluate XGBoost predictive models.

Key Tasks:

- Feature engineering and selection
- Model training and validation
- Performance evaluation (accuracy, precision, recall)
- Comparison against baseline models

Notebook 3: Data Exploration & EDA (202302 Dataset)

Objective: Analyze February 2023 dataset for comparison.

Key Tasks:

- Data cleaning and transformation
- Visualization of temporal trends

- Statistical analysis of key metrics
- Year-over-year comparison

Notebook 4: SVM & Random Forest Models (202102 Dataset)

Objective: Implement and compare alternative ML approaches.

Key Tasks:

- SVM implementation with hyperparameter tuning
- Random Forest model development
- Cross-validation and performance evaluation
- Feature importance analysis

3. Key Findings

Protocol Performance:

- **IPv6 Advantages:**
    - Higher mean/median download speeds
    - Lower average fetch times (~9353 ms vs IPv4's ~10000 ms)
- **IPv4 Advantages:**
    - Higher success rates (99.5% vs 92.4%)

Server Performance:

- Significant variability across different servers
- Some IPv6 servers outperform IPv4 counterparts
- Inconsistent performance indicates potential network issues

Visual Insights:

- Clear protocol comparisons via bar charts
- Success rate vs speed tradeoffs visualized
- Identifiable temporal patterns in performance

4. Data Preprocessing

Key Steps:

1. **Missing Data Handling:** Strategic imputation and removal
2. **Feature Engineering:**
    - Mbps conversion from bytes/sec
    - Derived temporal features
3. **Data Normalization:** Scaled numerical features

4. **Categorical Encoding:** Prepared server/protocol data for modeling

5. Model Performance

XGBoost Results:

- High prediction accuracy
- Key important features identified:
    - bytes_sec
    - fetch_time
    - success rates

SVM vs Random Forest:

- **Random Forest:**
    - Strong performance
    - High interpretability
    - Robust to overfitting
- **SVM:**
    - Required careful tuning
    - Effective for classification tasks
- **Comparison:** Random Forest generally superior

6. Challenges & Solutions

Key Challenges:

1. **Data Imbalance:** Addressed with resampling techniques
2. **Feature Selection:** Used correlation analysis and importance ranking
3. **Overfitting:** Mitigated via:
    - Cross-validation
    - Regularization
    - Pruning (for tree-based models)

7. Future Work

Proposed Enhancements:

- **Expanded Temporal Analysis:** Incorporate additional months/years
- **Advanced Modeling:**
    - LSTM/RNN for time-series prediction
    - Ensemble methods

- **Operational Deployment:**
    - Real-time monitoring system
    - Automated alerting for performance degradation
- **Enhanced Visualization:**
    - Interactive dashboards
    - Geographic performance mapping

8. Conclusion

This comprehensive analysis revealed significant insights about network performance characteristics, particularly the tradeoffs between IPv4 and IPv6 implementations. The machine learning approaches demonstrated strong predictive capability, with Random Forest emerging as the most robust model. The project establishes a foundation for ongoing network performance monitoring and optimization.

Appendix

- **GitHub Repository:** [https://github.com/Mohammed-abdulaziz-eisa/WBS-MBA]
- **Data Sources:** 202102 and 202302 network measurement datasets
- **Tools & Technologies:**
    - Python
    - Pandas, NumPy
    - Matplotlib, Seaborn, Plotly
    - Scikit-learn
    - XGBoost