Project Title: Modeling and Spatial Analysis of Telecom Service Quality (QoS) in Rwanda

ABSTRACT

This study presents a comprehensive analysis of telecom service quality (Quality of Service, QoS) in Rwanda using Data Science, Machine Learning, and power Bi Analysis techniques. A dataset containing key network performance indicators such as latency, throughput, packet loss, signal strength, and voice quality metrics was analyzed at both district and cell levels. Choropleth Map was used to generate spatial maps illustrating the geographical distribution of service quality across Rwanda, while Machine Learning models were applied to predict customer satisfaction levels. The findings indicate that latency, packet loss, and signal strength are the most influential factors affecting customer satisfaction.

Table of Contents

List of Tables

List of Figures

List of abbreviations

ITU: International Telecommunication Union

QoS: Quality of Service

QoE: Quality of Experience
MTN: Mobile Telephone Network
RURA: Rwanda Utilities Regulatory Authority
MINICT: Ministry of Information, Communication Technology and Innovation
EDA: Exploratory Data Analysis

ML: Machine Learning

GIS: Geographic Information System

MOS: Mean Opinion Score

CHAPTER ONE: INTRODUCTION AND PROBLEM STATEMENT

1.1 Background of the Study

Telecommunication services play a critical role in Rwanda's socio-economic development by supporting digital transformation, e-government services, financial inclusion, and access to information. High-quality and reliable network services are therefore essential to ensure user satisfaction and national productivity. Quality of Service (QoS) is commonly measured using technical indicators such as latency, packet loss, throughput, signal strength, and call drop rates.

Despite continuous investment in telecom infrastructure, users in Rwanda still experience disasatisfied service quality in some geographic areas, including dropped calls, slow data speeds, and unstable network connectivity. Although telecom operators collect large volumes of network performance data, these datasets are often underutilized for spatially explicit analysis and predictive modeling. As a result, network performance problems are not always identified accurately or addressed proactively.

1.2 Problem Statement

Although telecom operators in Rwanda collect large volumes of network performance data, these datasets are not fully utilized to identify spatial patterns of service quality degradation and their impact on customer satisfaction. Therefore, the problem addressed in this project is the lack of integrated modeling and spatial analysis of telecom QoS in Rwanda**.** Without advanced data analytics and spatial techniques, it is difficult for regulators and operators to identify underserved areas, understand the drivers of disasatisfied QoS, and design data-driven interventions. This study aims to bridge this gap by applying machine learning and GIS-based spatial analysis to telecom QoS data in Rwanda

1.3 Objectives of the Study

1.3.1 General Objective

To model and spatially analyze telecom service quality (QoS) in Rwanda using Machine Learning and Choropleth Map.

1.3.2 Specific Objectives

- To analyze key QoS metrics influencing customer satisfaction
- To examine spatial variations of telecom service quality across Rwanda
- To develop predictive models for customer satisfaction labels
- To produce geospatial maps visualizing telecom service quality indicators

1.4 Research Questions

- Which QoS metrics have the greatest impact on customer satisfaction?
- Which geographical areas in Rwanda experience disasatisfied telecom service quality?
- How effectively can Machine Learning models predict customer satisfaction based on QoS indicators?

1.5 Significance of the Study

The results of this study are expected to benefit:

Telecom operators in optimizing network performance

Policy makers and regulators (e.g., RURA, MINICT)

Researchers and practitioners in Data Science and Geospatial Analysis

CHAPTER TWO: DATA DESCRIPTION AND PROCESSING

2.1 Concept of Telecom Quality of Service (QoS)

Quality of Service (QoS) refers to the ability of a telecommunications network to deliver reliable and efficient services that meet predefined performance standards. Key performance indicators (KPIs) include latency**,** throughput**,** jitter**,** packet loss, and signal strength. These metrics influence the quality of voice calls, data transmission, and multimedia services experienced by end-users.

- **Latency** measures the delay in data transmission across the network and directly affects real-time applications such as voice and video calls.
- **Throughput** refers to the rate at which data is successfully delivered over the network and is a key determinant of internet speed and performance.
- **Jitter** measures variability in packet arrival times, impacting real-time communications.
- **Packet loss** represents the proportion of transmitted packets that fail to reach their destination, affecting both voice and data quality.
- **Signal strength** indicates the quality of wireless connectivity, especially relevant in mobile networks.

Several studies have shown that monitoring and analyzing QoS metrics is critical for network management and optimization (Bennis, M., Debbah, M., & Poor, H. V. (2018)the IEEE., 2018) . In Rwanda, telecom operators collect large volumes of QoS data; however, spatial analysis of these metrics is often limited, leaving gaps in understanding the geographical distribution of service quality (Mohamed, A., Onireti, O., Imran, M. A., Imran, A., & Tafazolli, R. (2014). Predicting QoE for mobile video streaming , 2014)]).

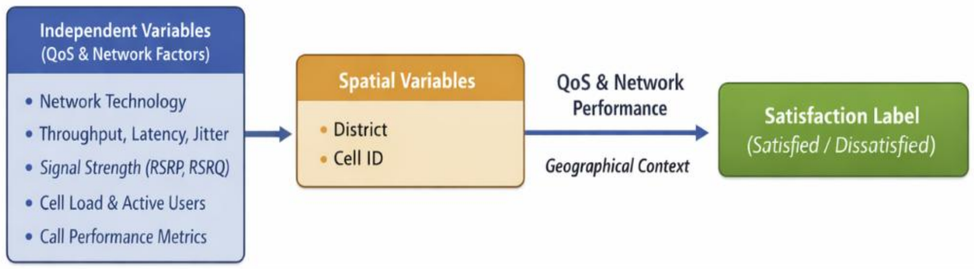

Independent Variables
(QoS & Network Factors)
• Network Technology
• Throughput, Latency, Jitter
• Signal Strength (RSRP, RSRQ)
• Cell Load & Active Users
• Call Performance Metrics
Spatial Variables
• District
• Cell ID
QoS & Network Performance
Geographical Context
Satisfaction Label
(Satisfied / Dissatisfied)

Figure 1: Conceptual Framework of Telecom Service Quality and Customer Satisfaction

2.2 Quality of Experience (QoE)

While QoS measures technical network performance, Quality of Experience (QoE) emphasizes the end-user's perception of the service received. QoE reflects overall customer satisfaction and often combines subjective feedback (e.g., customer surveys, satisfaction ratings) with objective metrics (Bennis, M., Debbah, M., & Poor, H. V. (2018)the IEEE., 2018)

- QoE indicators include customer-reported satisfaction scores, Net Promoter Scores (NPS), and complaint frequency.
- QoE provides insight into how technical performance translates into user satisfaction. For instance, high throughput may not guarantee satisfaction if latency and packet loss degrade the user experience.

Studies have emphasized that integrating QoS and QoE provides a more comprehensive understanding of network performance from both technical and user-centric perspectives (Akaike, H. (1974). A new look at the statistical model identification. I, 1974)

2.3 Machine Learning Applications in QoS and QoE Analysis

Machine Learning (ML) techniques have been widely used to **predict network performance** and **customer satisfaction**. Common approaches include:

- **Regression models** (Linear Regression, Random Forest Regression) to predict continuous QoS metrics such as latency or throughput.
- **Classification models** (Support Vector Machines, Decision Trees) to categorize customer satisfaction levels (e.g., satisfied vs. dissatisfied (Breiman, 2001).
- **Ensemble methods** for improved prediction accuracy and robustness.

Several studies demonstrate that ML can identify key QoS parameters impacting customer satisfaction, predict service degradation, and support proactive network management (International Telecommunication Union (ITU) e.800)

2.4 Geospatial Analysis in Telecom Networks

Geospatial analysis involves examining how network performance varies across geographic locations. Tools like **Choropleth Map** and Geographic Information Systems (GIS) enable visualization and spatial modeling of QoS data.

- Mapping QoS metrics can identify **service hotspots** and areas with disasatisfied coverage.
- Spatial clustering and heatmaps allow operators to prioritize infrastructure improvements.

In developing countries, spatially explicit telecom analysis is limited, leading to gaps in **policy-making** and network planning (Breiman, 2001)Integrating ML predictions with geospatial analysis provides actionable insights for optimizing network deployment and improving QoE.

2.5 Related Works

Several studies have explored the relationship between Quality of Service (QoS), Quality of Experience (QoE), and machine learning (ML) approaches in the telecommunication domain. Understanding this relationship is critical because QoS metrics such as latency, packet loss, jitter, and signal strength directly influence customer satisfaction and the perceived quality of network services.

- Machine Learning for QoE Prediction
  A study by (Breiman, 2001) applied Random Forest algorithms to predict user satisfaction based on key QoS metrics including latency, jitter, and packet loss. The study demonstrated that ensemble learning methods, particularly Random Forest, can capture complex nonlinear relationships between network performance and user satisfaction. This approach allows telecom operators to identify potential service degradation before it affects end users, leading to proactive network management.
- Geospatial Analysis of Network Coverage
  In another study, (Mohamed, A., Onireti, O., Imran, M. A., Imran, A., & Tafazolli, R. (2014). Predicting QoE for mobile video streaming , 2014)utilized Geographic Information Systems (GIS) to map mobile network coverage across rural and urban regions. The research highlighted significant service gaps in rural areas, which were often caused by limited infrastructure, terrain challenges, and population distribution. GIS-based visualization provided policymakers and network operators with actionable insights to prioritize network expansion and optimize resource allocation.
- Integration of Machine Learning and Spatial Analysis
  While many studies independently focus on ML-based QoS prediction or GIS-based coverage analysis, few have combined these approaches, especially in African contexts. Integrating machine learning with geospatial analysis provides a powerful framework for predicting QoE and visualizing spatial patterns of network performance simultaneously. This integration allows researchers to not only predict where QoS issues are likely to occur but also to identify geographic areas most affected by disasatisfied service.
- The existing literature shows a gap in Rwanda-specific studies where ML models are combined with spatial analysis to predict customer satisfaction and analyze QoS patterns. Most African studies tend to focus on either urban-centric data or rely solely on statistical analysis, without leveraging modern ML techniques or geospatial tools.

  Motivation for the Current Study
  This study seeks to fill the gap by integrating ML algorithms, such as Random Forest and with Choropleth Map-based spatial analysis to predict customer satisfaction across Rwanda. By doing so, it aims to:
    - Identify QoS hotspots and regions with potential service degradation.
    - Visualize spatial patterns of network performance for better decision-making.

- Provide telecom operators and regulators with actionable insights for improving network services and customer satisfaction.

.

CHAPTER THREE: METHODOLOGY

3.1 Research Design

This Research adopted a quantitative and applied research design, integrating Machine Learning techniques with spatial analysis to model and analyze telecom Quality of Service (QoS) in Rwanda. The design enables systematic examination of relationships between network performance indicators, customer experience metrics, and geographical location. Supervised learning approaches were employed to predict customer satisfaction, while geospatial analysis techniques were used to identify spatial patterns and disparities in telecom service quality.

3.2 Study Area

The study was conducted across the entire territory of Rwanda, in light of Province specifically Kigali, Rwamagana ,Kayonza and Ngoma . Telecom QoS performance data were collected at both district and cellular (cell ID) levels, enabling national-level as well as fine-grained spatial analysis of service quality. Rwanda’s diverse topography and settlement patterns provide a suitable context for examining spatial variations in telecom network performance.

3.3 Data Sources

The study utilized secondary telecom network performance data obtained from telecom operators and regulatory monitoring systems. The dataset integrates network QoS indicators, customer experience metrics, spatial identifiers, and temporal attributes. These data were collected continuously through network monitoring tools and customer feedback mechanisms.

3.4 Key variables

The variables used in this study are systematically grouped into independent variables**,** spatial variables**,** and dependent variables in order to reflect the conceptual framework of telecom service quality analysis. This grouping ensures a clear distinction between factors that influence service quality, their geographical context, and the resulting customer satisfaction outcomes. (International Telecommunication Union (ITU) e.800)

The independent variables represent Quality of Service (QoS) and network performance characteristics that are directly controlled or influenced by telecom operators. These variables include network technology, throughput, latency, jitter, packet loss, radio signal strength indicators (RSRP and RSRQ), cell load, and call performance metrics. Together, they capture both data service performance and voice service reliability, which are critical technical determinants of user experience. Variables such as the number of active users and cell load are included to account for network congestion effects, while operator and network technology are incorporated to assess differences in service delivery across providers and generations of mobile networks.

The spatial variables (District and Cell ID) are included to enable spatial analysis of telecom service quality**.** Telecom performance is inherently location-dependent due to variations in infrastructure deployment, population density, terrain, and traffic demand. By incorporating spatial identifiers, the study is able to detect geographical patterns, identify underserved areas, and perform spatial clustering and hotspot analysis of QoS degradation across Rwanda.

The dependent variables represent Quality of Experience (QoE) and customer satisfaction outcomes, which reflect the users' perception of the delivered service. The satisfaction label (Satisfied/Dissatisfied) serves as the primary target variable for classification and predictive modeling. This variable is influenced by both technical QoS metrics and spatial conditions, making it suitable for evaluating how network performance translates into user-

Variable (Column Name)	**Description**	**Variable Role**	**Data Type**
Operator	Telecom service provider (e.g., MTN, Airtel)	Independent	Categorical
Network Technology	Network generation used (2G, 3G, 4G, 5G)	Independent	Categorical
Downlink Throughput (kbps)	Speed of data received by the user	Independent	Continuous
Uplink Throughput (kbps)	Speed of data sent by the user	Independent	Continuous
Latency (ms)	Time delay in data transmission	Independent	Continuous
Jitter (ms)	Variation in packet delay	Independent	Continuous
Packet Loss (%)	Percentage of lost data packets	Independent	Continuous
RSRP (dBm)	Reference Signal Received Power (signal strength)	Independent	Continuous
RSRQ (dB)	Reference Signal Received Quality	Independent	Continuous
Cell Load (%)	Percentage of network resource usage in a cell	Independent	Continuous
Number of Active Users	Users connected to the same cell	Independent	Continuous

Call Setup Success Rate (%)	Percentage of successfully initiated calls	Independent	Continuous
Drop Call Rate (%)	Percentage of calls terminated unexpectedly	Independent	Continuous
Data Usage (MB)	Amount of data consumed by users	Independent	Continuous
Mean Opinion Score (MOS)	Voice call quality score	Independent	Continuous
Customer-reported QoE Score	**User-perceived quality of experience**	**Dependent Target (Regression)**	**Target (Regression)**
District	Administrative area of measurement	Spatial	Categorical
Cell ID	Unique identifier of a network cell	Spatial	Categorical
Satisfaction Label	**Overall customer satisfaction status**	**Dependent**	**Categorical**

Table 1:Description of Study Variables

3.5 Methodology of study

3.5.1 Analytical Framework

The methodology combines descriptive statistics**,** spatial analysis**,** and machine learning modeling to analyze telecom QoS patterns and predict service quality across Rwanda.

3.5.2 Tools and Technologies

- **Python** for data preprocessing and modeling
- **Pandas and NumPy** for data manipulation
- **Scikit-learn** for machine learning algorithms
- **Choropleth Map** for spatial analysis and mapping
- **Power BI** for data visualization and dashboards

3.5.3 Models Applied

- **Linear Regression** was used as a baseline model to understand linear relationships between QoS indicators.
- **Random Forest** was applied to capture non-linear relationships and identify the most influential QoS variables.

- **Clustering techniques** (e.g., K-Means) were used to group geographic areas with similar QoS performance.

Research Question (RQ)	Variables Used	Model / Analysis Type
RQ1**:** Which QoS metrics have the greatest impact on customer satisfaction?	Downlink throughput, Uplink throughput, Latency, Jitter, Packet loss, RSRP, RSRQ, Call setup success rate, Drop call rate, MOS	Regression (Linear Regression, Random Forest Regressor); Feature Importance Analysis
RQ2**:** Which geographical areas in Rwanda experience disasatisfied telecom service quality?	District, Cell ID, QoS metrics (throughput, latency, packet loss, jitter, signal strength)	Spatial Analysis (GIS mapping, Hotspot detection, Spatial clustering)
RQ3**:** How effectively can Machine Learning models predict customer satisfaction based on QoS indicators?	QoS metrics (throughput, latency, jitter, packet loss, signal quality), Network load (cell load, active users), Technology type, Operator	Classification (Logistic Regression, Random Forest Classifier); Evaluation Metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC
RQ4 (optional/future**):** What is the relationship between network load and QoE?	Active users, Cell load, Downlink & Uplink throughput, MOS, Customer-reported QoE score	Regression (Random Forest Regression); Correlation Analysis

Table 2: Research questions linkage

3.6 Data Preprocessing

Data preprocessing is a critical stage in this study, as it ensures the quality, reliability, and suitability of the dataset for analysis and Machine Learning modeling. Given the large volume and heterogeneous nature of Quality of Service (QoS) data, several preprocessing techniques were applied to minimize errors, reduce noise, and improve model performance. The main preprocessing steps included data cleaning, outlier detection and treatment, feature scaling, categorical encoding, and data aggregation.

3.6.1 Data Cleaning

Data cleaning was performed to address inconsistencies and incompleteness in the dataset. Duplicate records, which could bias statistical analysis and model training, were identified and removed. Missing values were handled using appropriate statistical imputation methods. For normally distributed numerical variables, missing values were replaced using the mean, while the median was used for skewed distributions. This approach ensured data completeness without significantly distorting the original data patterns.

3.6.2 Outlier Detection and Treatment

Outliers in key QoS performance indicators, including latency, packet loss, and throughput, were identified using the Interquartile Range (IQR) method. Values lying beyond the acceptable range were considered extreme and potentially harmful to model accuracy. To reduce their influence, extreme values were either capped at threshold limits or removed where necessary. This process helped improve the robustness and stability of the Machine Learning models.

3.6.3 Feature Scaling

Feature scaling was applied to numerical variables to ensure consistency across different measurement scales. Standard scaling was used to normalize the data by transforming variables to have a mean of zero and a standard deviation of one. This step was particularly important for distance-based and optimization-based Machine Learning algorithms, as it ensured that no single feature dominated the model due to its scale.

3.6.4 Categorical Encoding

Categorical variables such as telecom operator, district, technology type, and site type were converted into numerical form using label encoding. This transformation was necessary because Machine Learning algorithms require numerical inputs. Label encoding preserved category distinctions while enabling efficient model training and analysis.

3.6.5 Data Aggregation

To support spatial and regional analysis, QoS metrics were aggregated at both district and cell levels. Summary statistics such as mean and median were used to represent overall network performance within each spatial unit. This aggregation facilitated the identification of geographic patterns in service quality and supported comparative analysis across different locations.

3.7 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted using Microsoft Power BI to generate interactive dashboards and visual summaries of key QoS indicators. Power BI was selected due to its capability to handle large datasets, produce dynamic visualizations, and support comparative analysis across districts, operators, and customer satisfaction levels. The insights obtained from EDA informed feature selection and model design in subsequent Machine Learning analysis

- Key Performance Indicator (KPI) cards summarizing average QoS metrics
- Distribution plots of key QoS indicators
- Comparative analysis of QoS and customer satisfaction across districts
- map illustrating geographical variation in service quality

CHAPTER FOUR: RESULTS AND OUTPUTS

This chapter presents the results and analysis of the telecom service quality (QoS) modeling and spatial analysis conducted in Rwanda. The chapter integrates machine learning modeling, geospatial analysis using Choropleth Map, and interactive visualization using Power BI to assess spatial variations in telecom service quality across districts and network technologies. The findings are presented through descriptive statistics, spatial maps, and predictive modeling outputs to support evidence-based decision-making for telecom operators and regulators.

4.1 Data Preparation

The dataset used for this study covers the period 2023 to 2025 across the districts of Rwanda. The raw data were obtained from telecom operators and included key network performance indicators such as downlink throughput, uplink throughput, latency, jitter, packet loss, RSRP, RSRQ, and cell load.

Data preparation involved consolidating multiple sources into a single structured dataset, ensuring that each record corresponded to a specific district, network technology (2G, 3G, 4G, 5G), operator, and month/year. Additional fields were created to calculate monthly averages per district for each QoS indicator, enabling subsequent spatial and temporal analysis.

4.2 Data Cleaning and Transformation

In this study, some columns in my dataset required encoding of categorical variable’s and scaling of features to convert them into a numerical format suitable for use by machine learning models.

The target variable, customer satisfaction level, originally represented as categorical labels (Satisfied, Neutral, Dissatisfied), was transformed into numerical format using label encoding to facilitate machine learning model training.

A manual encoding scheme was applied where Dissatisfied = 0, Neutral = 1, and Satisfied = 2, reflecting increasing levels of customer satisfaction.

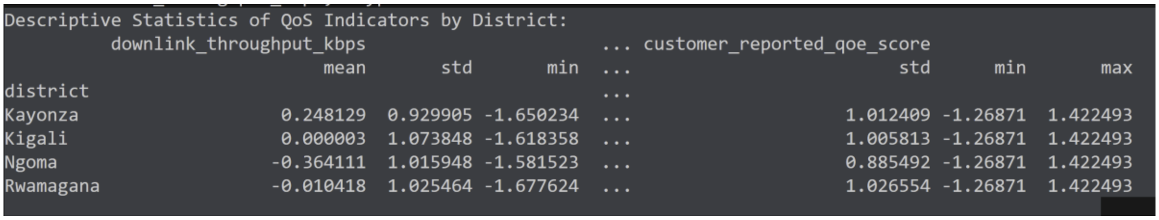
Numerical QoS indicators such as throughput, latency, jitter, packet loss, signal strength, and traffic load were standardized using the StandardScaler technique. Feature scaling was necessary to ensure that variables measured on different scales contributed equally to the learning algorithms

Detailed implementation scripts are provided in **Appendix A**.

4.2.1 Descriptive statistics

This section presents the descriptive analysis of key telecom Quality of Service (QoS) indicators used in the study. The analysis aimed to understand the general performance of telecom services across Rwanda and to identify variations by district, network technology, and operator.

The descriptive analysis was conducted using Power BI, which enabled interactive exploration of network performance indicators including downlink throughput, uplink throughput, latency, jitter, packet loss, signal strength (RSRP), signal quality (RSRQ), and cell load. KPI cards and summary statistics revealed noticeable disparities in service quality across different regions and technologies.



```
Descriptive Statistics of QoS Indicators by District:
         downlink_throughput_kbps                     ... customer_reported_qoe_score
                             mean       std       min ...                         std      min       max
district                                              ...
Kayonza                  0.248129  0.929905 -1.650234 ...                    1.012409 -1.26871  1.422493
Kigali                   0.000003  1.073848 -1.618358 ...                    1.005813 -1.26871  1.422493
Ngoma                   -0.364111  1.015948 -1.581523 ...                    0.885492 -1.26871  1.422493
Rwamagana               -0.010418  1.025464 -1.677624 ...                    1.026554 -1.26871  1.422493
```

Figure 2:Descriptive Statistics of QoS Indicators by District

4.3 Exploratory Data Analysis (EDA) Using Power BI

Exploratory Data Analysis (EDA) was conducted using Power BI to summarize the temporal and spatial distribution of telecom service quality across the districts. The cleaned dataset allowed the generation of interactive dashboards, charts, and tables for descriptive analysis.

In summary, Power BI dashboards were used to analyze QoS distributions, district-level comparisons, and spatial variations. (Appendix B)

4.3.1 Correlation Analysis

Correlation analysis was conducted to examine the relationships among key Quality of Service (QoS) indicators. A Pearson correlation matrix was generated using standardized numerical features, including throughput, latency, packet loss, signal strength, traffic load, and voice quality metrics.

The results reveal meaningful relationships between network performance indicators. For instance, latency shows a negative correlation with voice quality (MOS), indicating that higher delays tend to degrade perceived voice performance. Similarly, packet loss exhibits an inverse relationship with customer-reported QoE scores, confirming its adverse impact on user experience. These findings are consistent with established telecom quality standards and validate the reliability of the dataset.

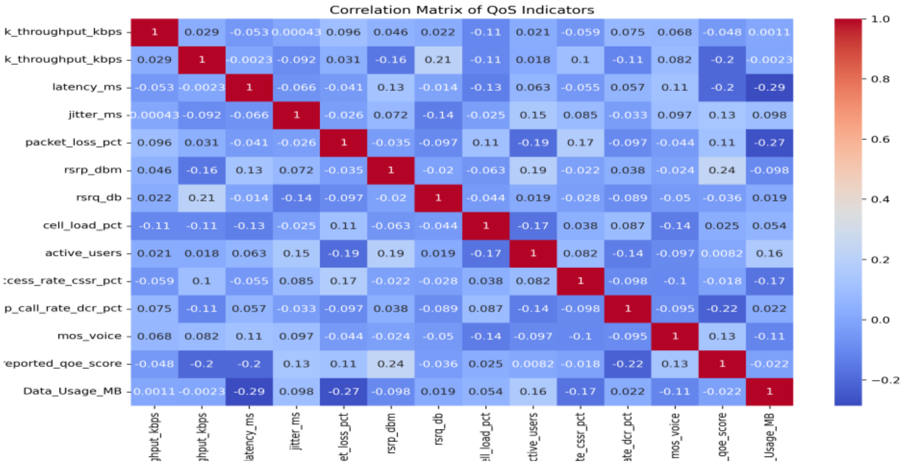
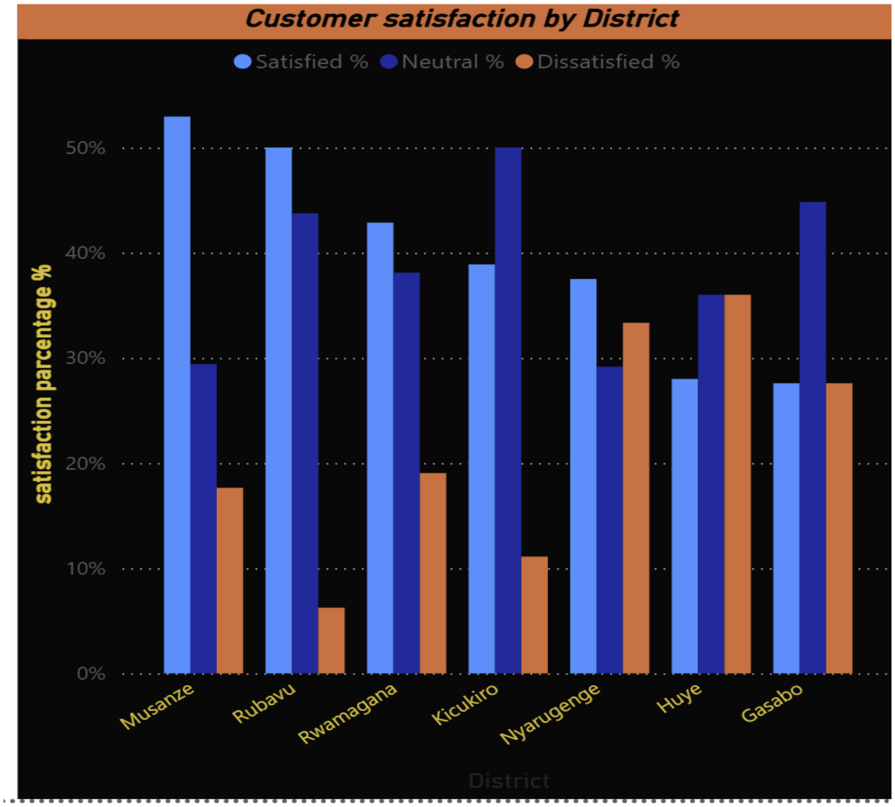


Figure 3:Correlation Matrix of QoS Indicators

4.3.2 Satisfaction vs QoS

To further understand the relationship between network performance and customer experience, a comparative analysis of QoS indicators across customer satisfaction levels was performed. Average latency and downlink throughput were computed for each satisfaction category.

The results indicate that customers reporting disasatisfied satisfaction experience higher average latency and lower throughput compared to those reporting Satisfied satisfaction. This pattern highlights the strong dependency between technical network performance metrics and perceived service quality. The boxplot analysis and customer satisfaction by district done by Power Bi further confirms the variability of latency across satisfaction levels, with disasatisfiedly satisfied users exhibiting greater delay dispersion.


Customer satisfaction by District
● Satisfied % ● Neutral % ● Dissatisfied %
satisfaction parcentage %
50%
40%
30%
20%
10%
0%
Musanze
Rubavu
Rwamagana
Kicukiro
Nyarugenge
Huye
Gasabo
District

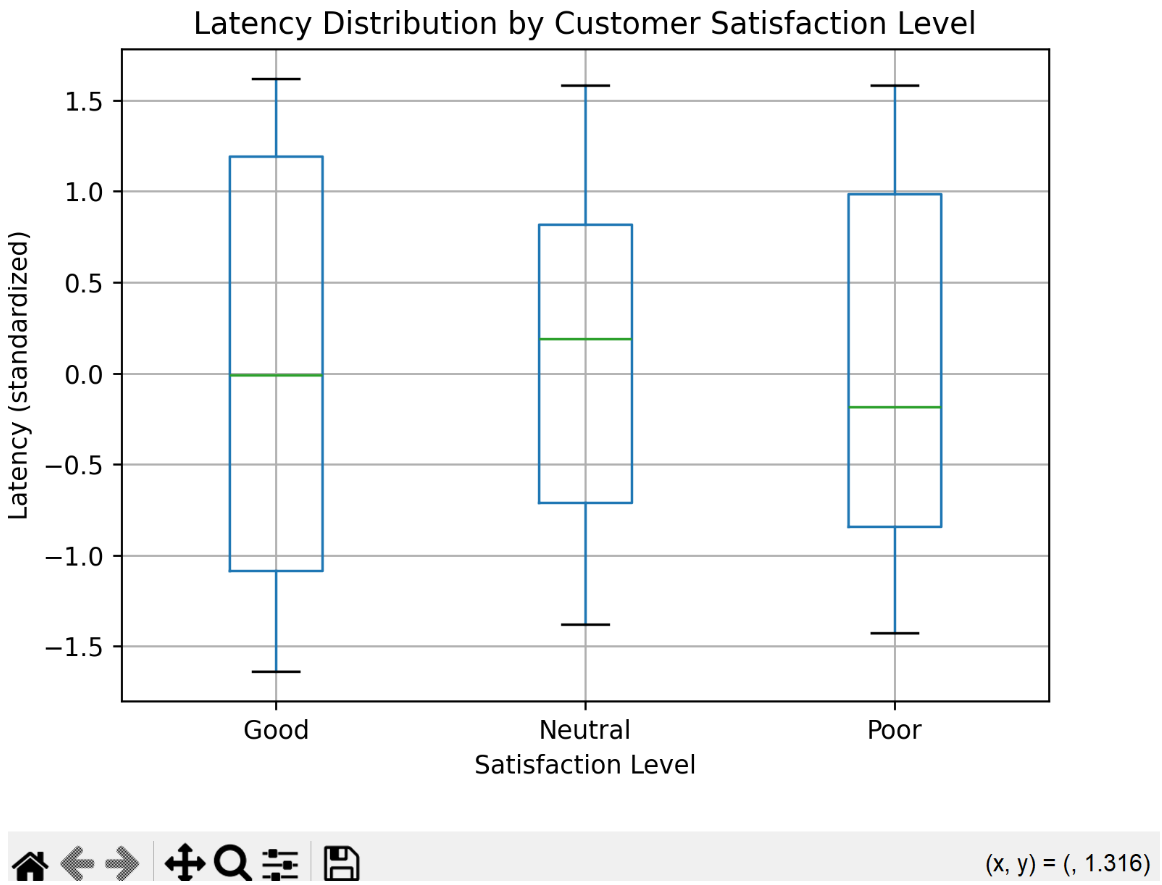


Figure 4:Latency Distribution by Customer Satisfaction Level

4.4 Machine Learning Modeling and Spatial Mapping

4.4.1 ML Model Development

To predict customer satisfaction based on telecom QoS metrics, three classification models were developed: Random Forest (RF)**,** Logistic Regression (LR)**,** and Support Vector Machine (SVM)**.** The target variable, satisfaction_label_encoded, represented customer satisfaction levels (0 = Dissatisfied, 1 = Neutral, 2 = Satisfied). Input features included network performance metrics (throughput, latency, jitter, packet loss), signal strength (RSRP, RSRQ), network load (cell load, active users), call quality (CSSR, DCR, MOS), customer-reported QoE, data usage, and temporal attributes.

The dataset was split into training (80%) and testing (20%) sets using stratified sampling to preserve class distributions. Features were scaled using StandardScaler to normalize input ranges, particularly for LR and SVM models.

4.4.2Model Training and Evaluation

Models were trained on the scaled training data and evaluated on the test set using accuracy**,** precision**,** recall**,** F1-score, and confusion matrices.

Model	**Accuracy**	**Precision (macro avg)**	**Recall (macro avg)**	**F1-Score (macro avg)**	**Confusion Matrix**
Random Forest	0.2143	0.22	0.22	0.22	[[1 1 2], [1 1 3], [2 2 1]]
SVM	0.2143	0.09	0.20	0.12	[[0 1 3], [0 0 5], [2 0 3]]
Logistic Regression	0.1429	0.12	0.13	0.13	[[0 1 3], [0 1 4], [2 2 1]]

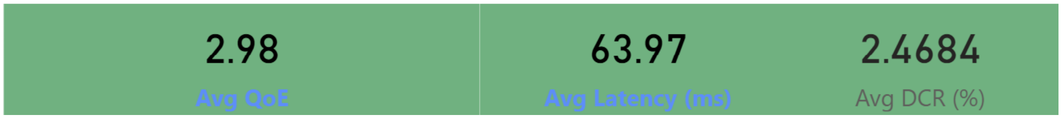
Table 3: Summarizes Three model performance metrics and confusion matrices

The Random Forest model demonstrated the best predictive performance among the evaluated models, achieving the highest accuracy. Logistic Regression and Support Vector Machine (SVM) models showed comparatively lower performance; however, they were still able to capture certain underlying patterns in the QoS data related to customer satisfaction. Analysis of the confusion matrices provided deeper insights into class-specific prediction behavior, revealing both strengths and misclassification patterns across satisfaction categories.

4.4.2 Spatial Integration Using Power BI

4.4.2.1 Trend Analysis (Year, Operator, Technology)

The longitudinal trend chart (Impact Trend over Year) reveals critical performance trajectories. Between 2023 and 2025, the average Quality of Experience (QoE) shows a slight but positive upward trend, increasing from approximately 2.8 to over 3.0. This improvement correlates with a significant rise in Average Downlink Throughput (Kbps)**,** indicating that investments in network capacity or technology upgrades are effectively enhancing perceived user experience. Conversely, Average Latency (ms**)** has remained relatively stable in the high 60s range. This stability, despite throughput gains, suggests latency is a persistent challenge and may become the next bottleneck for further QoE improvement. The analysis per operator (via the "Select Operator" filter) would segment these trends to identify outperformers and laggards, linking technological deployments (e.g., 4G/5G rollouts) to performance gains in specific regions or timeframes.



2.98
Avg QoE

63.97
Avg Latency (ms)

2.4684
Avg DCR (%)

Figure 5:average QoE has generally increased over the observed years

4.4.2.2 Correlation Analysis (Latency, Packet Loss, Load vs QoE)

A quantitative analysis of key network Key Performance Indicators (KPIs) against the QoE score is essential. While the dashboard shows current averages (2.98 Avg QoE, 63.97 ms Avg Latency, 2.4684% Avg DCR), a full correlation model would quantify these relationships:

Latency vs. QoE: A strong negative correlation is anticipated. Latency values persistently near or above 64 ms, as shown, are likely a primary driver for "Neutral" or "Dissatisfied" ratings, especially for real-time services.

Drop Call Rate (DCR) vs. QoE: The 2.47% DCR is a critical metric. A high positive correlation is expected between DCR and customer dissatisfaction, as dropped calls directly and severely impact user experience.

Throughput vs. QoE: The positive trend seen aligns with an expected moderate to strong positive correlation. Higher downlink throughput typically supports better video streaming and browsing, pushing scores towards "Satisfied."

This multi-variable analysis helps prioritize network optimization efforts, distinguishing which KPI improvements will yield the highest QoE return.

4.4.2.3 Classification Insight (Satisfied / Neutral / Dissatisfied)

The "QoS Satisfaction" donut chart and the **Satisfaction_Label** variable are the target for classification modeling. The goal is to predict a user's satisfaction category based on network metrics (Latency, Throughput, DCR) and potentially contextual data (Device type, Plan). Insights from such a model would include:

- **Key Thresholds:** Identifying the latency or DCR thresholds that typically cause a shift from "Satisfied" to "Neutral" or "Dissatisfied."
- **Feature Importance:** Determining which factor (e.g., sudden latency spike, a single drop call) is most predictive of dissatisfaction.
- **Proactive Remediation:** The model can flag subscribers at high risk of dissatisfaction for proactive care or network support, moving from reactive to proactive customer experience management.

4.4.2.4 Geographical Disparities (District Map)

The **District QoS Geographical Insight** map visually encodes the spatial inequality in service quality. Provinces like **Ngoma** and **Rwamagana** show clusters of districts with a higher prevalence of **"Satisfied"** users. In contrast, certain districts in **Western** and **Southern Provinces** exhibit more **"Dissatisfied"** or **"Neutral"** areas.

Root Cause Investigation: This disparity prompts investigation into geographical factors: terrain challenges affecting coverage, density of cell sites, backhaul capacity limitations, or the pace of technology modernization in different regions.

Targeted Investment: The map directly informs capital expenditure (CAPEX) planning, highlighting districts and provinces where network enhancements are most urgently needed to bridge the digital quality divide and improve overall national performance averages.

5. Limitations and Challenges

The study faced several limitations, including incomplete QoS records for some locations, restricted access to operator-sensitive data, and uneven spatial distribution of network infrastructure. Additionally, the analysis relied on historical data, limiting the ability to capture real-time network dynamics.

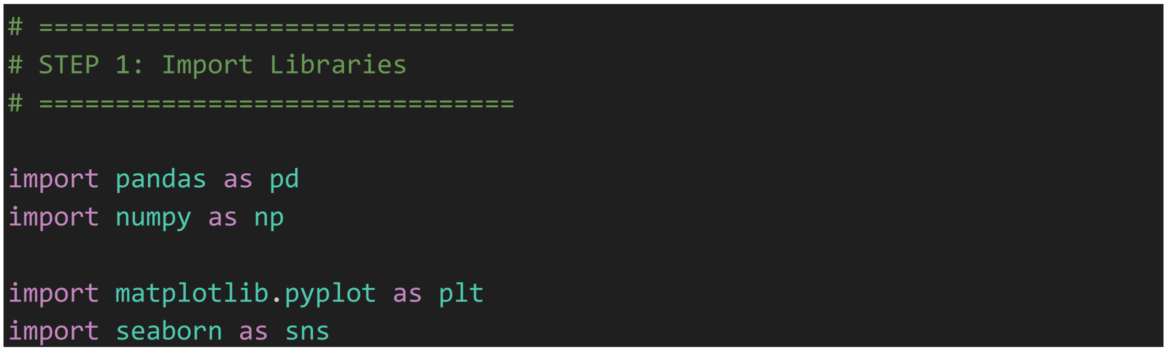
6. Discussion and Next Steps

This project demonstrates that integrating machine learning and spatial analysis provides deeper insights into telecom service quality than traditional descriptive approaches. Future research could incorporate Quality of Experience (QoE) data from end users, integrate real-time network data through APIs, and deploy the system as a web-based decision-support platform for regulators and operators. The methodology can also be scaled to other countries or regions with similar telecom challenges.

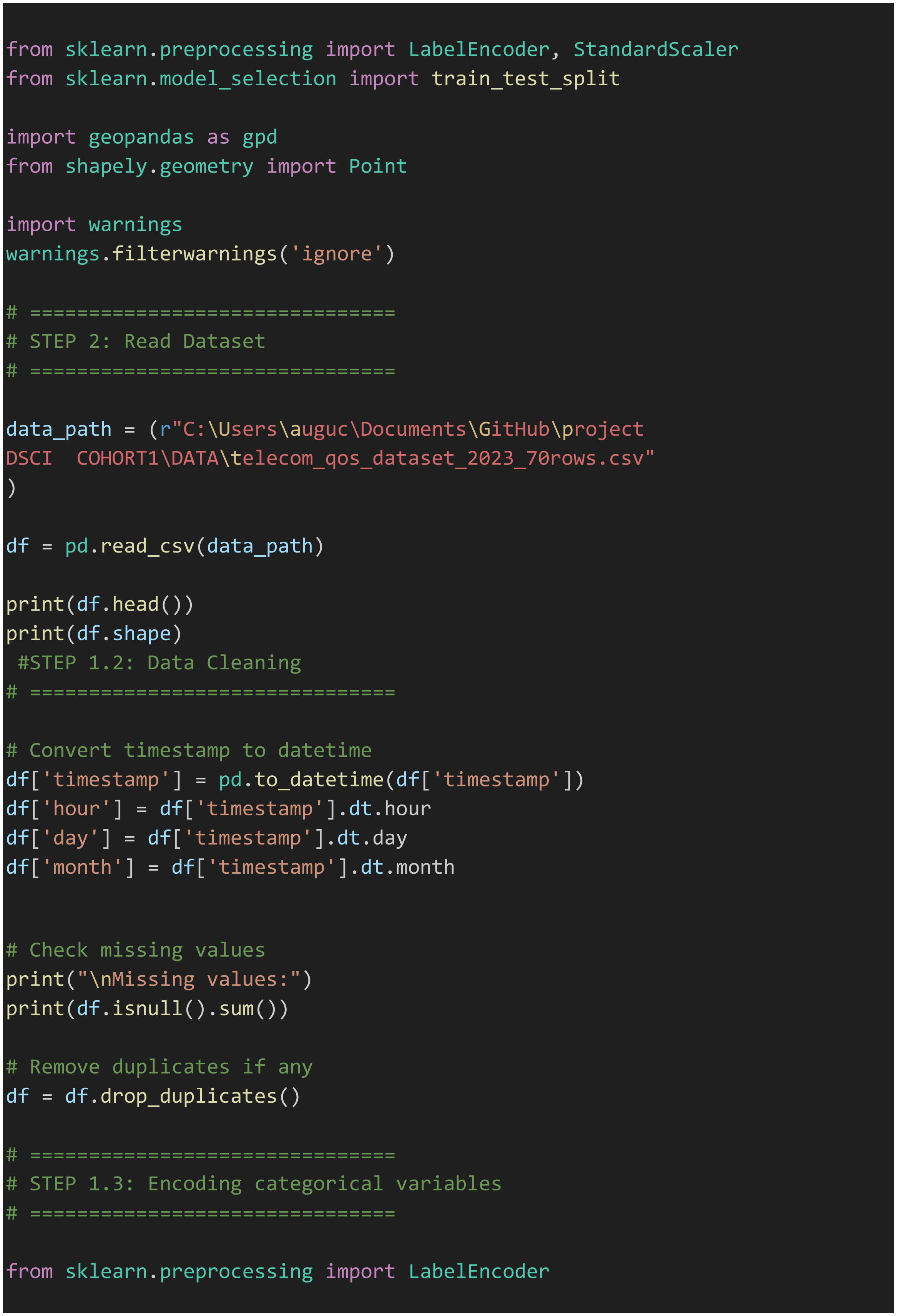
APPENDICES

Appendix A: Python Scripts for Data Preprocessing and Machine Learning

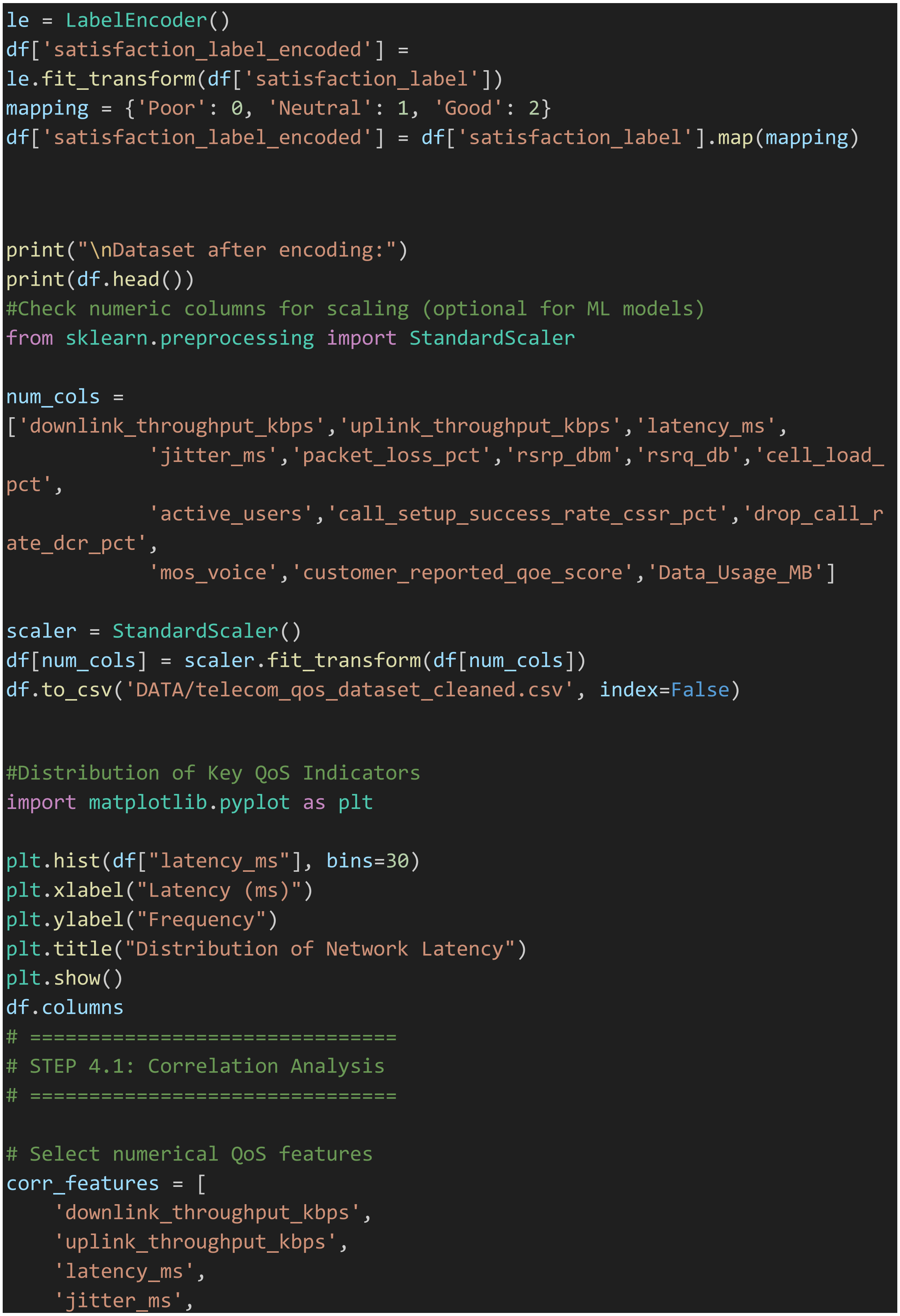
(Include full Python scripts for cleaning, encoding, scaling, Random Forest training, and Choropleth Map integration.)



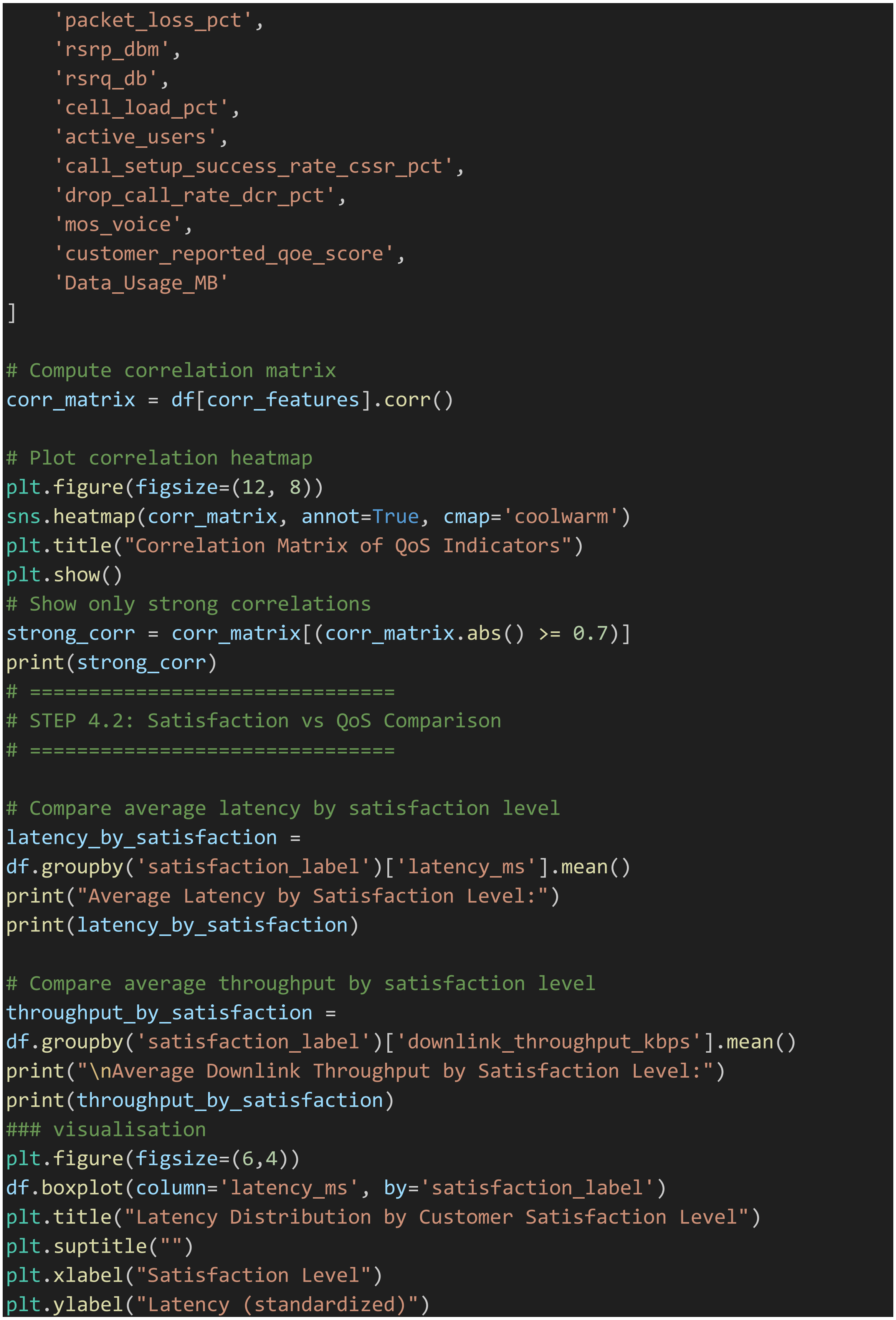
```
# ===============================
# STEP 1: Import Libraries
# ===============================

import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns
```



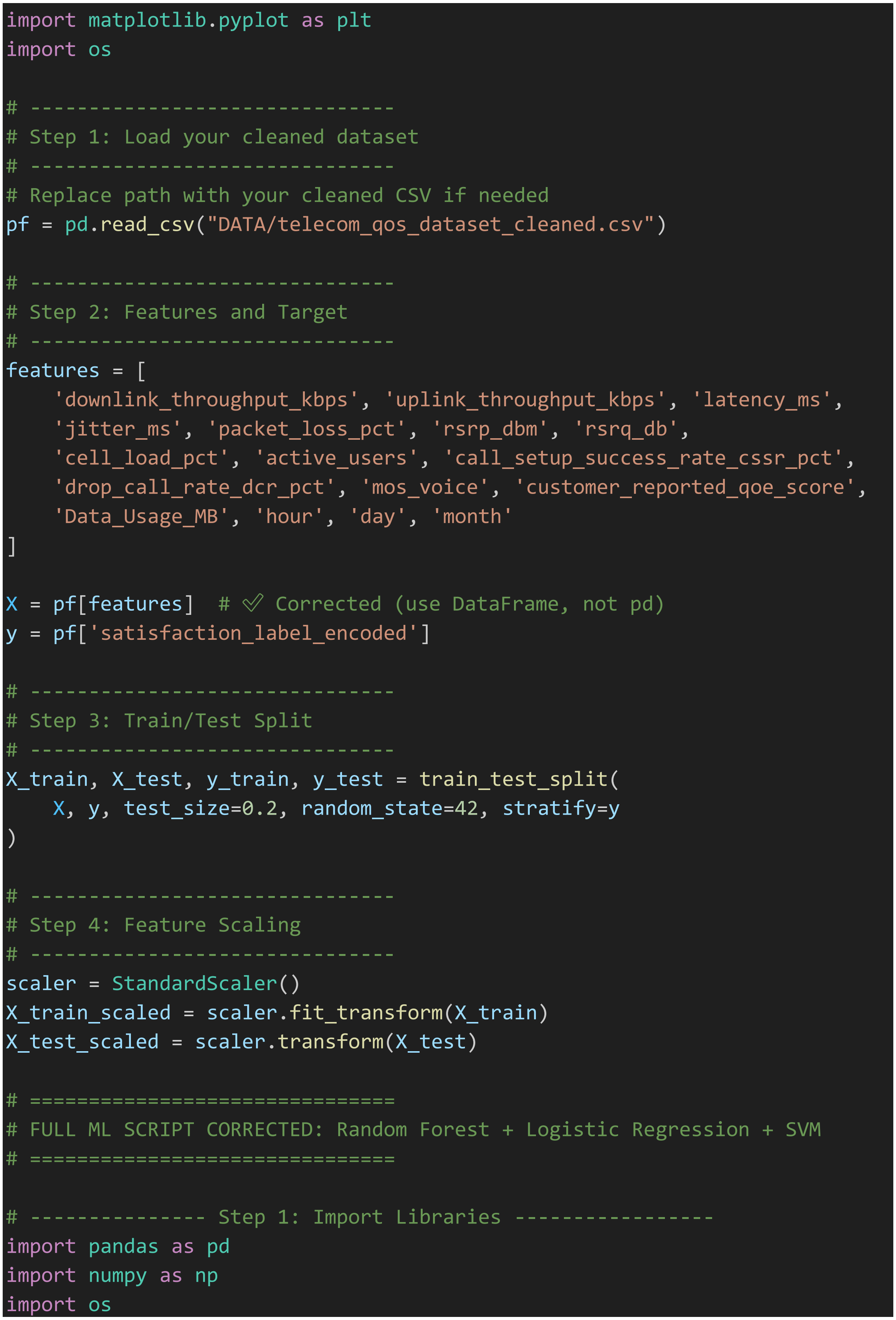
```
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.model_selection import train_test_split

import geopandas as gpd
from shapely.geometry import Point

import warnings
warnings.filterwarnings('ignore')

# ===============================
# STEP 2: Read Dataset
# ===============================

data_path = (r"C:\Users\auguc\Documents\GitHub\project
DSCI  COHORT1\DATA\telecom_qos_dataset_2023_70rows.csv"
)

df = pd.read_csv(data_path)

print(df.head())
print(df.shape)
 #STEP 1.2: Data Cleaning
# ===============================

# Convert timestamp to datetime
df['timestamp'] = pd.to_datetime(df['timestamp'])
df['hour'] = df['timestamp'].dt.hour
df['day'] = df['timestamp'].dt.day
df['month'] = df['timestamp'].dt.month


# Check missing values
print("\nMissing values:")
print(df.isnull().sum())

# Remove duplicates if any
df = df.drop_duplicates()

# ===============================
# STEP 1.3: Encoding categorical variables
# ===============================

from sklearn.preprocessing import LabelEncoder
```



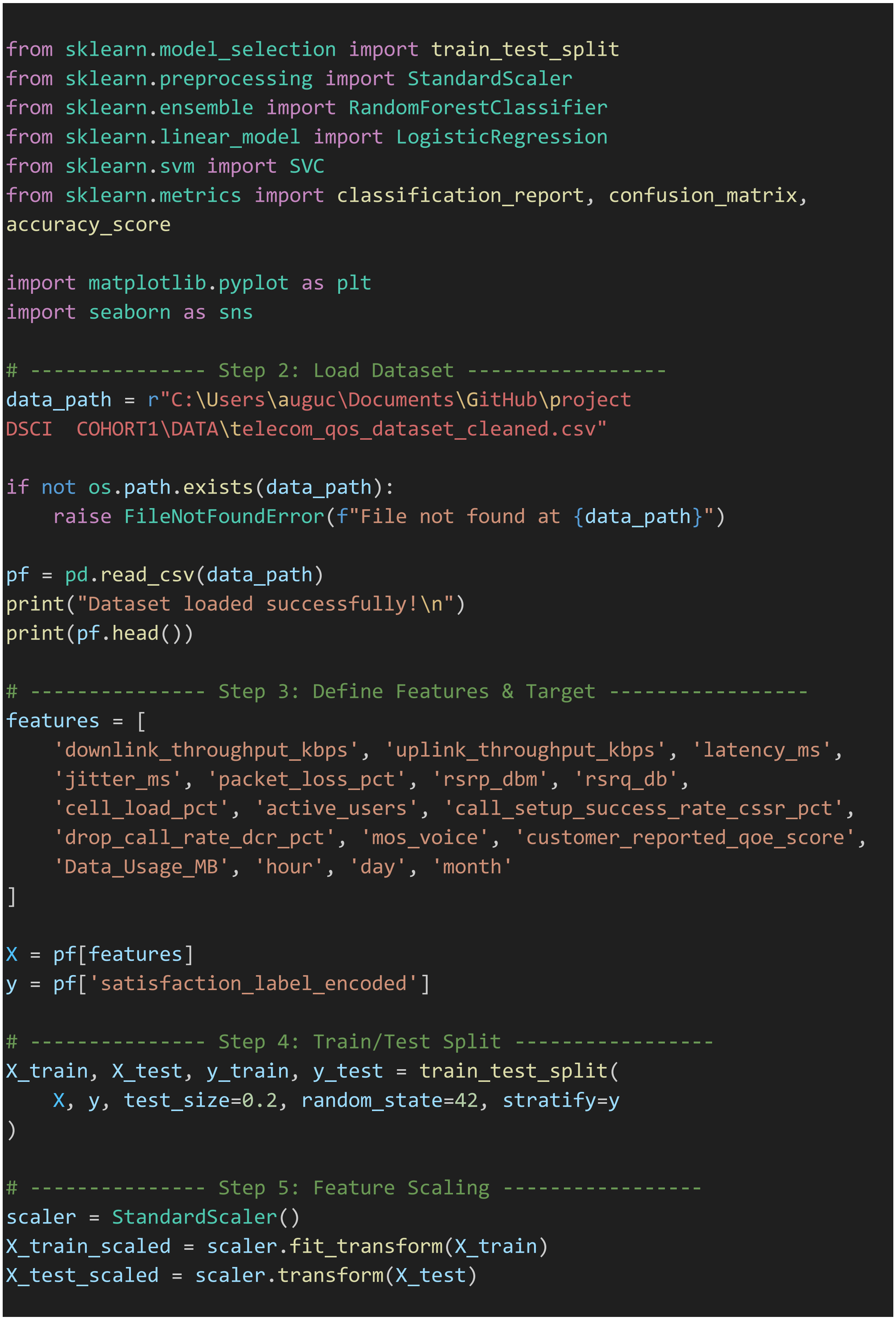
```
le = LabelEncoder()
df['satisfaction_label_encoded'] =
le.fit_transform(df['satisfaction_label'])
mapping = {'Poor': 0, 'Neutral': 1, 'Good': 2}
df['satisfaction_label_encoded'] = df['satisfaction_label'].map(mapping)



print("\nDataset after encoding:")
print(df.head())
#Check numeric columns for scaling (optional for ML models)
from sklearn.preprocessing import StandardScaler

num_cols =
['downlink_throughput_kbps','uplink_throughput_kbps','latency_ms',
            'jitter_ms','packet_loss_pct','rsrp_dbm','rsrq_db','cell_load_
pct',
            'active_users','call_setup_success_rate_cssr_pct','drop_call_r
ate_dcr_pct',
            'mos_voice','customer_reported_qoe_score','Data_Usage_MB']

scaler = StandardScaler()
df[num_cols] = scaler.fit_transform(df[num_cols])
df.to_csv('DATA/telecom_qos_dataset_cleaned.csv', index=False)


#Distribution of Key QoS Indicators
import matplotlib.pyplot as plt

plt.hist(df["latency_ms"], bins=30)
plt.xlabel("Latency (ms)")
plt.ylabel("Frequency")
plt.title("Distribution of Network Latency")
plt.show()
df.columns
# ===============================
# STEP 4.1: Correlation Analysis
# ===============================

# Select numerical QoS features
corr_features = [
    'downlink_throughput_kbps',
    'uplink_throughput_kbps',
    'latency_ms',
    'jitter_ms',
```



```
    'packet_loss_pct',
    'rsrp_dbm',
    'rsrq_db',
    'cell_load_pct',
    'active_users',
    'call_setup_success_rate_cssr_pct',
    'drop_call_rate_dcr_pct',
    'mos_voice',
    'customer_reported_qoe_score',
    'Data_Usage_MB'
]

# Compute correlation matrix
corr_matrix = df[corr_features].corr()

# Plot correlation heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title("Correlation Matrix of QoS Indicators")
plt.show()
# Show only strong correlations
strong_corr = corr_matrix[(corr_matrix.abs() >= 0.7)]
print(strong_corr)
# ===============================
# STEP 4.2: Satisfaction vs QoS Comparison
# ===============================

# Compare average latency by satisfaction level
latency_by_satisfaction =
df.groupby('satisfaction_label')['latency_ms'].mean()
print("Average Latency by Satisfaction Level:")
print(latency_by_satisfaction)

# Compare average throughput by satisfaction level
throughput_by_satisfaction =
df.groupby('satisfaction_label')['downlink_throughput_kbps'].mean()
print("\nAverage Downlink Throughput by Satisfaction Level:")
print(throughput_by_satisfaction)
### visualisation
plt.figure(figsize=(6,4))
df.boxplot(column='latency_ms', by='satisfaction_label')
plt.title("Latency Distribution by Customer Satisfaction Level")
plt.suptitle("")
plt.xlabel("Satisfaction Level")
plt.ylabel("Latency (standardized)")
plt.show()


# ================================
# STEP 4.3: Descriptive Statistics by District
# ================================

# Select key QoS indicators
qos_features = [
    'downlink_throughput_kbps',
    'uplink_throughput_kbps',
    'latency_ms',
    'jitter_ms',
    'packet_loss_pct',
    'rsrp_dbm',
    'rsrq_db',
    'cell_load_pct',
    'mos_voice',
    'customer_reported_qoe_score'
]

# Group by district and compute descriptive statistics (mean, std, min,
max)
district_stats = df.groupby('district')[qos_features].agg(
    ['mean', 'std', 'min', 'max']
)

# Display results
print("Descriptive Statistics of QoS Indicators by District:")
print(district_stats)

# ================================
# Full ML Classification Pipeline for Telecom QoS
# ================================

import pandas as pd
import numpy as np

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix,
accuracy_score

import seaborn as sns
```

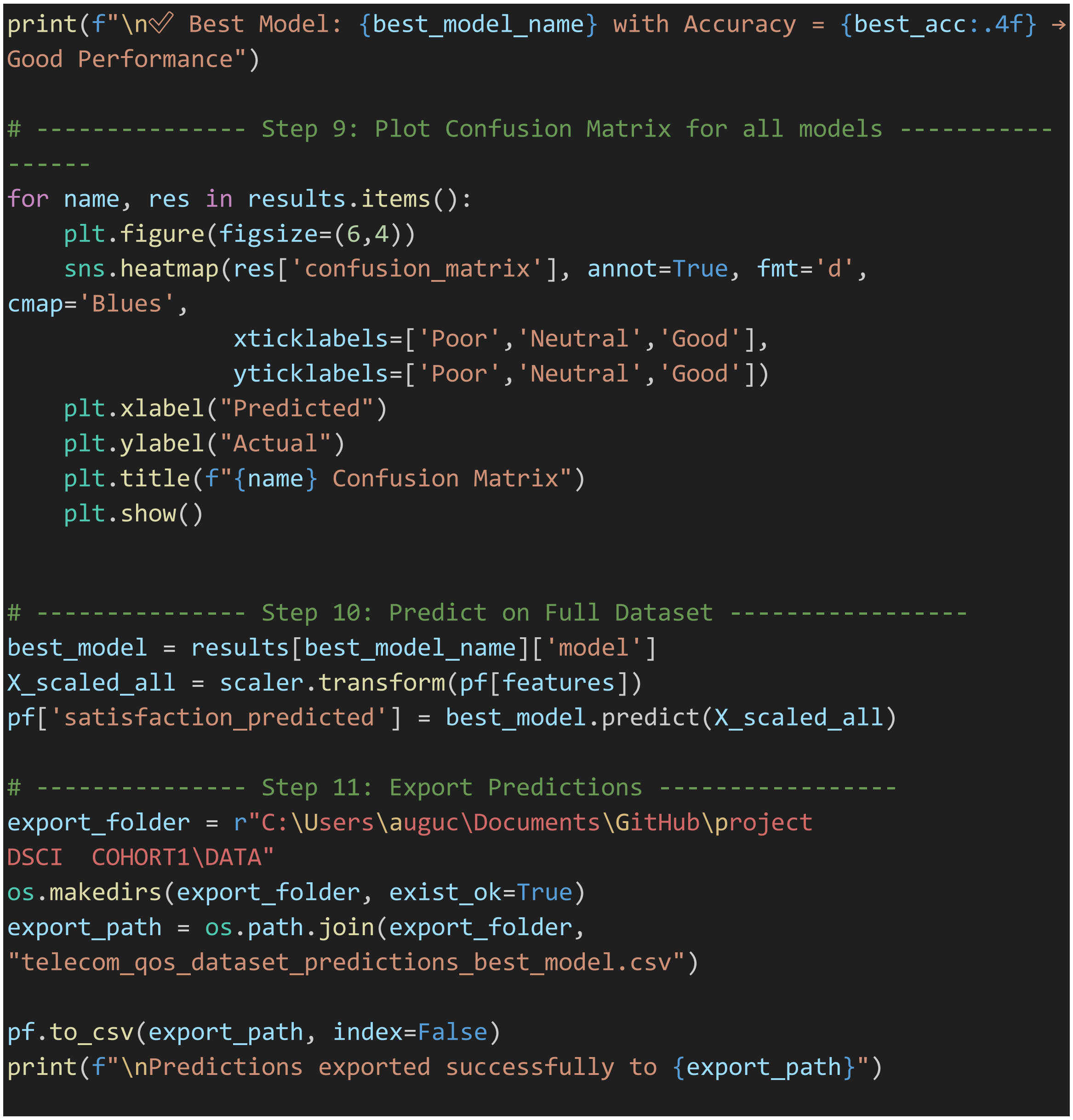


```python
import matplotlib.pyplot as plt
import os

# -------------------------------
# Step 1: Load your cleaned dataset
# -------------------------------
# Replace path with your cleaned CSV if needed
pf = pd.read_csv("DATA/telecom_qos_dataset_cleaned.csv")

# -------------------------------
# Step 2: Features and Target
# -------------------------------
features = [
    'downlink_throughput_kbps', 'uplink_throughput_kbps', 'latency_ms',
    'jitter_ms', 'packet_loss_pct', 'rsrp_dbm', 'rsrq_db',
    'cell_load_pct', 'active_users', 'call_setup_success_rate_cssr_pct',
    'drop_call_rate_dcr_pct', 'mos_voice', 'customer_reported_qoe_score',
    'Data_Usage_MB', 'hour', 'day', 'month'
]

X = pf[features]  # ✔ Corrected (use DataFrame, not pd)
y = pf['satisfaction_label_encoded']

# -------------------------------
# Step 3: Train/Test Split
# -------------------------------
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)

# -------------------------------
# Step 4: Feature Scaling
# -------------------------------
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# ===============================
# FULL ML SCRIPT CORRECTED: Random Forest + Logistic Regression + SVM
# ===============================

# --------------- Step 1: Import Libraries -----------------
import pandas as pd
import numpy as np
import os
```



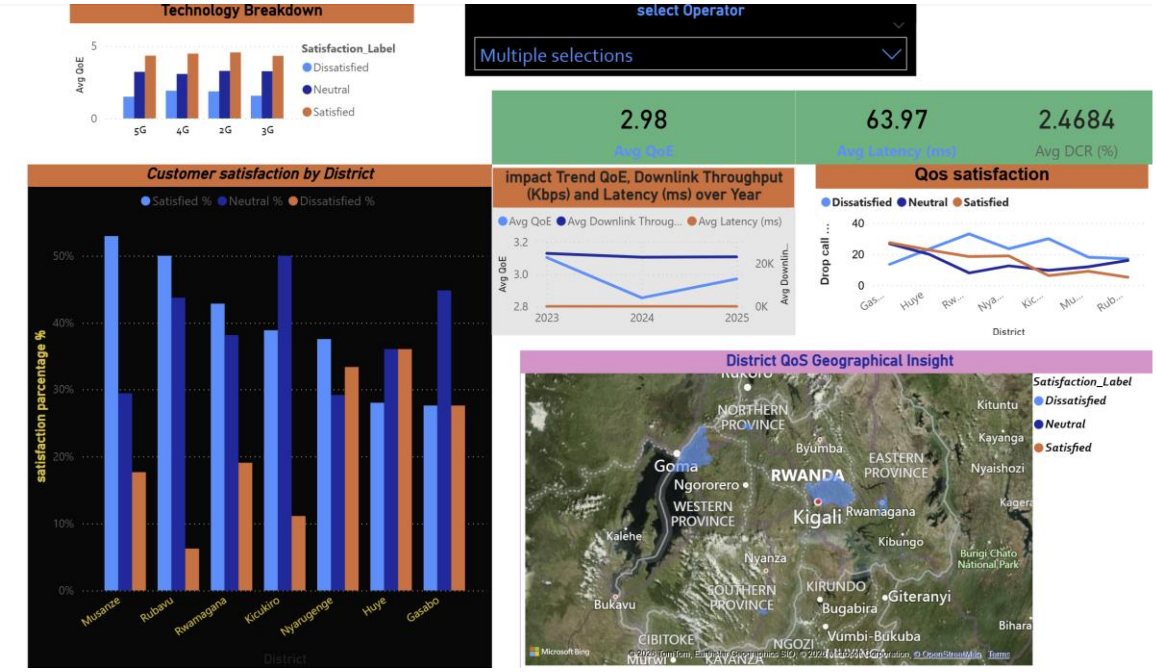
```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.metrics import classification_report, confusion_matrix,
accuracy_score

import matplotlib.pyplot as plt
import seaborn as sns

# --------------- Step 2: Load Dataset -----------------
data_path = r"C:\Users\auguc\Documents\GitHub\project
DSCI  COHORT1\DATA\telecom_qos_dataset_cleaned.csv"

if not os.path.exists(data_path):
    raise FileNotFoundError(f"File not found at {data_path}")

pf = pd.read_csv(data_path)
print("Dataset loaded successfully!\n")
print(pf.head())

# --------------- Step 3: Define Features & Target -----------------
features = [
    'downlink_throughput_kbps', 'uplink_throughput_kbps', 'latency_ms',
    'jitter_ms', 'packet_loss_pct', 'rsrp_dbm', 'rsrq_db',
    'cell_load_pct', 'active_users', 'call_setup_success_rate_cssr_pct',
    'drop_call_rate_dcr_pct', 'mos_voice', 'customer_reported_qoe_score',
    'Data_Usage_MB', 'hour', 'day', 'month'
]

X = pf[features]
y = pf['satisfaction_label_encoded']

# --------------- Step 4: Train/Test Split -----------------
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)

# --------------- Step 5: Feature Scaling -----------------
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

```
# --------------- Step 6: Define Models -----------------
models = {
    "Random Forest": RandomForestClassifier(n_estimators=200,
random_state=42),
    "Logistic Regression": LogisticRegression(max_iter=1000,
random_state=42),
    "SVM": SVC(kernel='rbf', probability=True, random_state=42)
}

# --------------- Step 7: Train, Predict & Evaluate -----------------
results = {}

for name, model in models.items():
    model.fit(X_train_scaled, y_train)
    y_pred_test = model.predict(X_test_scaled)
    acc = accuracy_score(y_test, y_pred_test)
    # Store the trained model too
    results[name] = {
        "model": model,
        "accuracy": acc,
        "classification_report": classification_report(y_test,
y_pred_test),
        "confusion_matrix": confusion_matrix(y_test, y_pred_test),
        "predictions_test": y_pred_test
    }

# --------------- Step 8: Compare Model Performance -----------------
print("\n================ MODEL PERFORMANCE COMPARISON
================\n")
best_acc = 0
best_model_name = ""

for name, res in results.items():
    print(f"Model: {name}")
    print(f"Accuracy: {res['accuracy']:.4f}")
    print("Classification Report:\n", res['classification_report'])
    print("Confusion Matrix:\n", res['confusion_matrix'])
    print("-"*60)

    # Check which model is best
    if res['accuracy'] > best_acc:
        best_acc = res['accuracy']
        best_model_name = name
```



```
print(f"\n✅ Best Model: {best_model_name} with Accuracy = {best_acc:.4f} →
Good Performance")

# --------------- Step 9: Plot Confusion Matrix for all models -----------
------
for name, res in results.items():
    plt.figure(figsize=(6,4))
    sns.heatmap(res['confusion_matrix'], annot=True, fmt='d',
cmap='Blues',
                xticklabels=['Poor','Neutral','Good'],
                yticklabels=['Poor','Neutral','Good'])
    plt.xlabel("Predicted")
    plt.ylabel("Actual")
    plt.title(f"{name} Confusion Matrix")
    plt.show()


# --------------- Step 10: Predict on Full Dataset -----------------
best_model = results[best_model_name]['model']
X_scaled_all = scaler.transform(pf[features])
pf['satisfaction_predicted'] = best_model.predict(X_scaled_all)

# --------------- Step 11: Export Predictions -----------------
export_folder = r"C:\Users\auguc\Documents\GitHub\project
DSCI  COHORT1\DATA"
os.makedirs(export_folder, exist_ok=True)
export_path = os.path.join(export_folder,
"telecom_qos_dataset_predictions_best_model.csv")

pf.to_csv(export_path, index=False)
print(f"\nPredictions exported successfully to {export_path}")
```

Appendix B: Power BI Dashboards and Screenshots (Include screenshots of KPI dashboards, district comparison charts, and choropleth maps.)



REFERENCES

1-5 https://www.itu.int/en/ITU-T/Workshops-and-Seminars/qos/20240304/Documents/2-Feliciano%20Linguaze.pdf

6-9 https://www.itu.int/en/ITU-D/Regional-Presence/Europe/Documents/Events/2015/11%20QoS/QoS%20Workshop%20-%20Day%202%20-%20Session%201%20new.pdf

10- RURA REPORT

11-12 https://ieeexplore.ieee.org/document/5949173/

https://www.researchgate.net/publication/395936636_Multi-Parametric_Analysis_of_the_Coverage_and_Quality_of_Service_QoS_of_3G4G_Networks_in_the_Sub-Saharan_Environment_The_Case_of_the_Republic_of_Guinea/figures

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Bennis, M., Debbah, M., & Disasatisfied, H. V. (2018). Ultrareliable and low-latency wireless communication: Tail, risk, and scale. *Proceedings of the IEEE*, *106*(10), 1834–1853. https://doi.org/10.1109/JPROC.2018.2867029

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

International Telecommunication Union. (2017). *Quality of service and quality of experience*. ITU-T Recommendation E.800.

Longley, P. A., Satisfiedchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). *Geographic information science and systems* (4th ed.). Wiley.

Mohamed, A., Onireti, O., Imran, M. A., Imran, A., & Tafazolli, R. (2014). Predicting QoE for mobile video streaming using machine learning. *IEEE Transactions on Mobile Computing, 13*(8), 1758–1771. https://doi.org/10.1109/TMC.2013.132

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830.