

Final Project — Choose One Case

Each group selects **one** of the cases below. All follow the same pipeline:

- **Mandatory** Data Preprocessing + EDA
- **Mandatory** Supervised ML (Classification/Regression)
- **Mandatory** Unsupervised ML (Clustering + Dimensionality Reduction)
- **Mandatory** Deep Learning (DNN + Specialized Architecture)
- **Mandatory** Streamlit Deployment
- **Clear Dataset Specification**
- **No Optional Tasks** - All components are required and integrated

Deep Learning

As a AI researcher

I want to develop deep learning models that capture complex audio patterns

So that we advance music understanding beyond traditional features

Requirements:

- Build DNN regressor with 5 hidden layers for popularity prediction
- Implement Autoencoders for unsupervised music representation learning
- Use cluster embeddings as additional model features
- Apply attention mechanisms for interpretable feature importance
- Compare against traditional audio analysis methods

Streamlit Deployment

As a music product manager

I want an interactive music analysis platform with 114,000+ tracks

So that our curation team can make data-driven decisions

Requirements:

- Song feature input with real-time popularity prediction
 - Music cluster explorer with similar track recommendations
 - Genre evolution visualization across decades
 - Artist similarity analysis and trend spotting
 - Playlist generator based on audio feature preferences
-

User Story 6: Traffic Prediction & Urban Mobility Analysis

Data Preprocessing

As a data engineer at SmartCity Traffic

I want to process 100,000 ride records from multiple transportation services

So that we build accurate urban mobility models

Dataset: Uber & Lyft Dataset - 100,000 records subset

- **Size:** 100,000 rides, 10 features
- **Target:** price (ride pricing prediction)

Requirements:

- Geospatial processing of pickup/dropoff coordinates
- Create temporal features: rush hour indicators, holiday effects
- Engineer distance features using Haversine formula
- Handle categorical variables with high cardinality (location pairs)
- Address pricing outliers and surge multiplier effects

Exploratory Data Analysis

As a traffic analyst

I want to analyze 100,000 rides to understand urban mobility patterns

So that I can identify transportation optimization opportunities

Requirements:

- Geographic heat maps of ride density and pricing variations
- Temporal analysis of ride patterns by hour, day, and season
- Service comparison: Uber vs Lyft pricing and availability
- Surge pricing analysis and triggering condition identification
- Route popularity and congestion correlation analysis

Supervised ML

As a data scientist

I want to predict ride prices using 100,000 records across multiple services

So that we optimize transportation pricing and availability

Requirements:

- Implement XGBoost with spatial-temporal cross-validation
- Build Random Forest with feature importance for business insights
- Train Gradient Boosting with categorical feature support
- Evaluate using quantile loss for pricing distribution prediction
- Feature engineering for spatial interactions and temporal effects

Unsupervised ML

As a urban planner

I want to identify mobility pattern types across 100,000 rides

So that we can develop targeted transportation solutions

Requirements:

- Apply KMeans clustering (k=10) on spatio-temporal ride patterns
- Use DBSCAN for anomaly detection in pricing and routing
- Implement PCA for dimensionality reduction of feature space
- Create mobility segment profiles with characteristic patterns

- Validate clusters with urban planning characteristics

Deep Learning

As a AI engineer

I want to develop neural networks for complex mobility pattern prediction

So that we improve transportation forecasting accuracy

Requirements:

- Build LSTM networks for temporal price pattern prediction
- Implement Autoencoders for ride pattern representation learning
- Apply attention mechanisms for interpretable predictions
- Compare against traditional spatial-statistical methods

Streamlit Deployment

As a transportation operations manager

I want a comprehensive mobility dashboard with 100,000+ ride insights

So that we can optimize urban transportation systems

Requirements:

- Real-time price prediction with route and time inputs
- Mobility pattern explorer with filtering capabilities
- Surge pricing alert system and trend analysis
- Service comparison dashboard with performance metrics
- Data export for urban planning and policy development