International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Online subsidy management system Using machine learning (algorithm- logistic regression, random forest, decision tree)

Soundrya Mallappa Biradar*, Nikhil Gurudev Lonari, Aniket Ramesh Bhandare, Vishwaraj Pradip Pawar, Mrs. Pallavee Bavane-Patil
D Y Patil Technical campus talsande Kolhapur, Maharashtra, India.

Abstract- — Government subsidy programs play a crucial role in socio-economic development by supporting vulnerable populations in sectors such as agriculture, education, healthcare, energy, and food security. However, traditional subsidy management systems are often plagued by inefficiencies, fraud, leakage, lack of transparency, and poor targeting. The advent of digital governance and data-driven technologies has opened new avenues for reforming subsidy allocation and monitoring mechanisms. Machine learning (ML), in particular, offers powerful tools for automating eligibility assessment, predicting beneficiary behavior, detecting anomalies, and optimizing policy outcomes. This review paper presents a comprehensive analysis of online subsidy management systems integrated with machine learning techniques, with a specific focus on Logistic Regression, Decision Tree, and Random Forest algorithms. The paper discusses system architecture, data sources, preprocessing methods, algorithmic frameworks, evaluation metrics, real-world use cases, challenges, ethical considerations, and future research directions. The review aims to serve as a ready reference for researchers, policymakers, and system designers working toward intelligent, transparent, and efficient subsidy management platforms.

Keywords: Online Subsidy Management System, Machine Learning, Logistic Regression, Decision Tree, Random Forest, E-Governance, Fraud Detection, Welfare Distribution

I. INTRODUCTION

Subsidies are financial assistance mechanisms provided by governments to reduce inequality, promote economic stability, and ensure access to essential goods and services. In developing countries, subsidy programs constitute a significant portion of public expenditure. Despite their importance, conventional subsidy distribution frameworks rely heavily on manual verification, rule-based decision-making, and fragmented databases, which often result in delayed disbursement, inclusion and exclusion errors, and large-scale misuse.

With the rapid digitization of public services, online subsidy management systems (OSMS) have emerged as a transformative solution. These platforms integrate beneficiary registration, document verification, eligibility assessment, fund disbursement, and grievance redressal into a unified digital ecosystem. However, merely digitizing legacy processes does not eliminate structural inefficiencies. Intelligent decision-making is required to handle large-scale, heterogeneous data and complex eligibility criteria.

Machine learning provides adaptive, data-driven approaches that can learn patterns from historical subsidy data and improve decision accuracy over time. Classification algorithms such as Logistic Regression, Decision Trees, and Random Forests are particularly suitable for subsidy eligibility prediction and fraud detection tasks. This review explores how these algorithms are applied within OSMS, evaluates their comparative performance, and highlights research gaps.

II. BACKGROUND AND RELATED WORK

2.1 Traditional Subsidy Management Systems

Traditional systems are largely paper-based or semi-digitized, involving multiple government departments and intermediaries. Common limitations include:

- Manual verification of beneficiary data
- Lack of real-time monitoring
- High administrative overhead
- Vulnerability to corruption and duplicate beneficiaries

Several studies have reported leakage rates ranging from 10–40% in large subsidy programs, emphasizing the need for automation and transparency.

2.2 Evolution to Online Subsidy Management Systems

Online subsidy management systems leverage web portals, centralized databases, biometric identification, and direct benefit transfer (DBT) mechanisms. While these systems improve efficiency, rule-based eligibility checks often fail to adapt to evolving socio-economic conditions. Recent research has therefore explored the integration of artificial intelligence and machine learning for intelligent subsidy governance.

2.3 Role of Machine Learning in E-Governance

Machine learning has been successfully applied in e-governance domains such as tax fraud detection, smart policing, healthcare policy planning, and social welfare analytics. In subsidy systems, ML enables:

- Automated eligibility classification
- Fraud and anomaly detection
- Predictive analytics for budget planning
- Policy impact assessment

III. SYSTEM ARCHITECTURE OF AN ML-BASED ONLINE SUBSIDY MANAGEMENT SYSTEM

A typical ML-enabled OSMS consists of the following layers:

1. **Data Collection Layer:**
   - Demographic data (age, gender, income)
   - Socio-economic indicators (occupation, landholding, education)
   - Transactional data (previous subsidies, payment history)
   - Government databases (tax, census, identity systems)
2. **Data Preprocessing Layer:**
   - Data cleaning and normalization
   - Handling missing values
   - Feature encoding (categorical to numerical)
   - Class imbalance correction
3. **Machine Learning Layer:**
   - Logistic Regression
   - Decision Tree
   - Random Forest
4. **Application Layer:**
   - Eligibility prediction
   - Fraud detection alerts
   - Dashboard and reporting tools
5. **Security and Governance Layer:**
   - Data encryption
   - Access control
   - Audit logs

IV. MACHINE LEARNING ALGORITHMS FOR SUBSIDY MANAGEMENT

4.1 Logistic Regression

Logistic Regression is a supervised learning algorithm widely used for binary classification problems. In subsidy systems, it is applied to predict whether an applicant is eligible (yes/no) based on input features.
Advantages:

- Simple and interpretable
- Computationally efficient

- Suitable for large datasets

Limitations:

- Assumes linear relationship between features and log-odds
- Limited performance with complex, non-linear data

Mathematically, the probability of eligibility is given by:

$$P(y=1|x) = 1 / (1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)})$$

4.2 Decision Tree

Decision Trees classify data by recursively splitting it based on feature values. They mimic human decision-making and are highly intuitive for policy interpretation.

Advantages:

- Easy to understand and visualize
- Handles non-linear relationships
- Requires minimal data preprocessing

Limitations:

- Prone to overfitting
- Sensitive to noisy data

Decision Trees are particularly useful for rule extraction, enabling policymakers to understand key eligibility determinants.

4.3 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and robustness.

Advantages:

- High accuracy
- Reduced overfitting
- Handles high-dimensional data well

Limitations:

- Reduced interpretability compared to single trees
- Higher computational cost

Random Forest models are widely used for fraud detection by identifying anomalous beneficiary patterns.

V. DATA PREPROCESSING AND FEATURE ENGINEERING

Effective ML performance depends heavily on data quality. Common preprocessing steps include:

- Removing duplicate beneficiary records
- Normalizing income and asset values
- Encoding categorical variables (one-hot encoding)
- Feature selection using correlation analysis or importance scores

Feature engineering may include derived indicators such as income-to-family-size ratio or subsidy dependency index.

VI. MODEL EVALUATION METRICS

The performance of ML models in subsidy systems is evaluated using:

- Accuracy
- Precision and Recall
- F1-Score
- ROC-AUC

In fraud detection scenarios, recall is often prioritized to minimize false negatives.

VII. COMPARATIVE ANALYSIS OF ALGORITHMS

Algorithm	Interpretability	Accuracy	Scalability	Use Case
Logistic Regression	High	Moderate	High	Eligibility screening

Decision Tree	Very High	Moderate	Moderate	Rule extraction
Random Forest	Moderate	High	High	Fraud detection

VIII. CHALLENGES AND LIMITATIONS

Despite their advantages, ML-based subsidy systems face several challenges:

- Data privacy and security concerns
- Bias in training data leading to unfair decisions
- Lack of explainability in complex models
- Integration with legacy government systems

IX. Ethical and Legal Considerations

The use of ML in public welfare requires strict adherence to ethical principles such as fairness, accountability, transparency, and explainability. Regulatory compliance with data protection laws is essential to maintain public trust.

X. FUTURE RESEARCH DIRECTIONS

Future work may explore:

- Deep learning for multi-class subsidy prediction
- Federated learning for privacy-preserving analytics
- Explainable AI (XAI) frameworks
- Real-time anomaly detection systems

XI. CONCLUSION

Machine learning-enabled online subsidy management systems represent a significant advancement in e-governance. Algorithms such as Logistic Regression, Decision Tree, and Random Forest offer complementary strengths for eligibility assessment and fraud detection. While challenges related to ethics, privacy, and interpretability remain, continued research and policy collaboration can ensure equitable and efficient subsidy distribution.

REFERENCES

1. Bishop, C. M. Pattern Recognition and Machine Learning. Springer, 2006.
2. Breiman, L. “Random Forests.” Machine Learning, 45(1), 2001, pp. 5–32.
3. Mitchell, T. M. Machine Learning. McGraw-Hill, 1997.
4. OECD. Digital Government Review, OECD Publishing, 2020.
5. Witten, I. H., Frank, E., Hall, M. A. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2016.
6. World Bank. Improving Social Protection through Digital Technologies, 2019.