SKILL SHIKSHYA & VRIT TECHNOLOGIES

Data Science and Machine Learning (DSML) Course

PROJECT PROPOSAL

Customer Churn Prediction System with Deployment

Student Name:	Suresh Karki
Batch:	DSML Batch-08
Date:	February 2026

1. Executive Summary

This project aims to develop a comprehensive Customer Churn Prediction System that leverages machine learning algorithms to identify customers at risk of leaving a subscription-based service. The system will incorporate end-to-end data science workflow including data preprocessing, exploratory data analysis, feature engineering, multiple ML model implementation, hyperparameter tuning, and deployment via a web interface.

The project demonstrates practical application of concepts learned throughout the DSML course including supervised learning, deep learning, model evaluation, and web deployment using Flask/Streamlit.

2. Project Objectives

- Build a robust classification model to predict customer churn with accuracy above 85%
- Implement and compare multiple ML algorithms (Logistic Regression, Random Forest, XGBoost, Neural Networks)
- Perform comprehensive EDA to identify key churn indicators and patterns
- Apply advanced preprocessing techniques, handling imbalanced data using SMOTE
- Deploy the final model through an interactive web interface for real-time predictions
- Generate actionable insights and recommendations for customer retention strategies

3. Dataset Description

Dataset Source: Telco Customer Churn dataset from Kaggle or UCI Machine Learning Repository

Sample Size: ~7,000 customer records

Key Features:

- Customer Demographics: Gender, age, partner status, dependents
- Service Details: Phone service, internet service, online security, tech support
- Account Information: Tenure, contract type, payment method, billing preferences
- Financial Metrics: Monthly charges, total charges

4. Methodology and Technical Approach

4.1 Data Preprocessing

- Handle missing values using appropriate imputation techniques
- Encode categorical variables using Label Encoding and One-Hot Encoding
- Feature scaling using StandardScaler for numerical features
- Address class imbalance using SMOTE (Synthetic Minority Over-sampling Technique)

4.2 Exploratory Data Analysis

- Univariate analysis: Distribution plots, box plots for outlier detection
- Bivariate analysis: Correlation heatmaps, churn rate by features
- Feature importance analysis using Random Forest and statistical tests

4.3 Model Development

The following algorithms will be implemented and compared:

1. **Logistic Regression** (baseline model)
2. **Decision Tree Classifier** (interpretability)
3. **Random Forest** (ensemble method)
4. **XGBoost/AdaBoost** (gradient boosting)
5. **Artificial Neural Network** (deep learning with Keras)

4.4 Model Evaluation

- Performance metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC
- K-Fold Cross-validation for robust performance estimation
- Confusion matrix analysis for understanding prediction errors
- Hyperparameter tuning using GridSearchCV

4.5 Deployment

- Web interface development using Streamlit or Flask
- Model serialization using pickle/joblib
- User-friendly input form for real-time churn prediction

5. Tools & Technologies

Category	Technologies
Programming Language	Python 3.8+
Data Analysis	NumPy, Pandas, Matplotlib, Seaborn
Machine Learning	Scikit-learn, XGBoost, imbalanced-learn
Deep Learning	TensorFlow, Keras
Deployment	Streamlit / Flask
Development Environment	Jupyter Notebook, VS Code

6. Expected Deliverables

6. **Jupyter Notebooks:** Separate notebooks for EDA, preprocessing, modeling, and evaluation
7. **Trained Models:** Serialized models (.pkl files) for all algorithms
8. **Web Application:** Fully functional prediction interface
9. **Documentation:** README file with setup instructions and project overview
10. **Final Report:** Comprehensive report with findings and recommendations
11. **Presentation:** PowerPoint slides for project demonstration

7. Project Timeline

Phase	Activities	Duration
Week 1	Data collection, EDA, preprocessing	5 days
Week 2	Feature engineering, model development	5 days
Week 3	Hyperparameter tuning, model comparison	4 days
Week 4	Web deployment, testing, documentation	5 days

8. Expected Outcomes & Business Impact

- Achieve minimum 85% accuracy in churn prediction
- Identify top 5 features contributing to customer churn
- Reduce false negatives to minimize revenue loss from undetected churners
- Provide actionable retention strategies based on model insights
- Deploy a production-ready system for real-world application

9. Conclusion

This project integrates multiple concepts from the DSML curriculum including Python programming, data preprocessing, machine learning algorithms, deep learning with neural networks, and web deployment. The comprehensive approach ensures practical understanding of the complete data science pipeline from problem definition to deployment.

By successfully completing this project, I will demonstrate proficiency in applying theoretical knowledge to solve real-world business problems, which aligns with the course objectives of preparing industry-ready data science professionals.

Student Signature
Date: 9th February 2026