**Predictive Analysis Using Machine Learning**

**Objective**

The primary goal of this project is to build a predictive machine learning model capable of classifying whether a customer is likely to churn based on features such as usage, demographics, and service preferences.

**Dataset**

We used the **Telco Customer Churn** dataset available from open data repositories like Kaggle. The dataset contains information about:

* Customer demographics
* Account information
* Subscription details
* Churn status (target variable)

**Data Preprocessing**

1. **Missing Values Handling**: Checked for nulls and replaced them with mode (for categorical variables) or median (for numerical variables).
2. **Encoding Categorical Variables**:
   * Used LabelEncoder for binary categories.
   * Applied OneHotEncoding for multi-class categorical variables.
3. **Feature Scaling**: Normalized numerical features using StandardScaler.

**Feature Selection**

* Correlation matrix analysis for numerical data.
* Recursive Feature Elimination (RFE) to identify top predictors.
* Domain-based feature understanding to retain relevant business features.

**Model Selection**

We experimented with multiple classification algorithms:

| **Model** | **Reason** |
| --- | --- |
| Logistic Regression | Baseline model |
| Random Forest | Handles imbalance well |
| XGBoost | High accuracy & interpretable |
| SVM | Effective in high-dimensional space |
| KNN | Instance-based learning |

**Final Model Training**

* **Chosen Model**: Random Forest Classifier
* **Hyperparameter Tuning**:  
  Used GridSearchCV to tune:
  + Number of estimators
  + Maximum depth
  + Minimum samples split

**Evaluation Metrics**

| **Metric** | **Score** |
| --- | --- |
| Accuracy | 83.5% |
| Precision | 78.6% |
| Recall | 80.3% |
| F1 Score | 79.4% |
| ROC-AUC Score | 88.2% |

Visuals Used:

* Confusion Matrix
* ROC Curve
* Feature Importance Plot

**Key Insights**

* Tenure, contract type, monthly charges, and payment method significantly influence churn.
* Customers on month-to-month contracts have the highest churn rate.
* Long-term customers (higher tenure) show greater loyalty.

]

**Deliverables**

* **Jupyter Notebook (.ipynb)** demonstrating:
  + Data preprocessing
  + Feature selection
  + Model training
  + Evaluation and visualization
* **Python Script (.py)** version of the notebook
* **Project Report (.docx)** – this document

**Tools & Libraries**

* Python: pandas, numpy, scikit-learn, matplotlib, seaborn, xgboost
* Jupyter Notebook
* Visual Studio Code

**Conclusion**

This project successfully demonstrates how machine learning can be used to predict customer behavior using historical data. By applying preprocessing techniques, selecting relevant features, training multiple models, and tuning hyperparameters, we built a robust predictive model. These insights can support customer retention strategies and improve business decision-making.