**Intelligent Customer Retention: Using Machine Learning for Enhanced Prediction of Telecom Customer Churn**

**Data Collection & Preparation**

ML depends heavily on data. It is the most crucial aspect that makes algorithm training possible. So this section allows you to download the required dataset.

* Collect the Dataset
* Importing the Libraries
* Read the Dataset
* Data Preparation
* Handling Missing Values
* Handling Categorical Values
* Handling Imbalance Data

**Collect the Dataset**

* There are many popular open sources for collecting the data. E.g.: kaggle.com, UCI repository, etc.In this project we have used .cave data.
* This data is downloaded from kaggle.com. Please refer to the link given below to download the dataset.

**#Importing necessary Libraries**

**import pandas as pd**

**import numpy as np**

**import pickle**

**import matplotlib.pyplot as plt**

**%matplotlib inline**

**import seaborn as sns**

**import sklearn**

**from sklearn.preprocessing import LabelEncorder, OneHotEncoder**

**from sklearn.linear\_model import LogisiticRegression**

**from sklearn.tree import DecisionTreeClassifier**

**from sklearn.ensemble import RandomForestClassifier**

**from sklearn.neighbors import KNeighborsClassifier**

**from sklearn.svm import SVC**

**form sklearn.model\_selection import RandomizedSearchCV**

**import imblearn**

**from imblearn.over\_sampling import SMOTE**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.preprocessing import StandardScaler**

**from sklearn.metrics impor t accurancy\_score, classification\_report,**

**Read the Dataset**

* Our dataset format might be in .csv, excel files, .txt, .json, etc. We can read the dataset with the help of pandas.
* In pandas we have a function called read\_csv () to read the dataset. As a parameter we have to give the directory of the csv file.
* There are several techniques but in our project we are using manual encoding with the help of list comprehensions

**Data Preparation**

* As we have understood how the data is, let's pre- process the collected data.
* The download data set is not suitable for training the machine learning model as it might have so much randomness so we need to clean the dataset properly in order to fetch good results.
* This activity includes the following steps.

**Handling Missing Values**

* Let's find the shape of our dataset first.
* To find the shape of our data, the df.shape method is used.
* To find the data type, df.info () function is used.

**Handling Categorical Values**

* As we can see our dataset has categorical data we must convert the categorical data to integer encoding or binary encoding.
* To convert the categorical features into numerical features we use encoding techniques.
* There are several techniques but in our project we are using manual encoding with the help of list comprehension, into independent and dependent variables.
* The independent variable in the dataset would be considered as 'x' and gender, SeniorCitizen, Partner columns would be considered as independent variable.
* The dependent variable in the dataset would be considered as 'y' and the 'Churn' column is considered as dependent variable.
* Now we will split the.

**Label Encoding**

* Label Encoding refers to converting the labels into numeric form so as to convert it into the machine- readable form.
* Machine learning algorithms can then decide in a better way on how those labels must be operated.
* It is an important pre- processing step for the structured dataset in supervised learning

**Splitting the Dataset into Dependent and Independent variable**

* Let's split our dataset into independent and dependent variables.
* The independent variable in the dataset would be considered as 'x' and gender, Senior Citizen, Partner columns would be considered as independent variable.
* The dependent variable in the dataset would be considered as 'y' and the 'Churn' column is considered as

**Handling Imbalance Data**

* Data Balancing is one of the most important steps, which need to be performed for classification models, because when we train our model on imbalanced dataset, we will get biased results, which means our model is able to predict only one class element.
* For balancing the data we are using the SMOTE Method. SMOTE: Synthetic minority over sampling technique, which will create new synthetic data points