**Customer Churn Telcom Data mining**

**Fredrick Kimani Njuguna 14301612**

**Kelvin Chidi 14348359**

**Ugoh Chukwuhma Leonard 14282392**

**22nd March 2024**

# Introduction

The project at hand delves into the application of data mining techniques to a dataset that is centered around customer churn in the Iranian telecom industry. This dataset is a rich source of information, encompassing details about customers' usage of services, their demographic characteristics, and crucially, whether they have churned or not. Customer churn, also known as customer attrition, refers to the phenomenon of customers ceasing their usage of a company's products or services.

In the context of the telecom industry, this could mean customers discontinuing their telecom subscriptions or switching to a different service provider. Churn is a critical metric for any business, as retaining existing customers is often more cost-effective than acquiring new ones. Therefore, understanding and predicting customer churn is of paramount importance for effective customer relationship management and strategic decision-making.

The primary objective of this project is to construct a predictive model with the capability to accurately identify customers who are likely to churn. This involves the use of data mining techniques to uncover patterns and relationships in the dataset that can help predict churn behavior. The model's predictions can then be used to inform proactive customer retention strategies.

For instance, if the model identifies a particular customer as being at high risk of churn, the telecom company could engage this customer with targeted interventions, such as personalized offers or improved service packages, in an attempt to retain them. By enabling such proactive measures, the predictive model could potentially help the telecom company reduce churn rates, enhance customer loyalty, and ultimately, drive business growth.

The project will involve various stages, including data preprocessing, exploratory data analysis, feature selection, model training, and model evaluation. Each of these stages will require careful execution and rigorous validation to ensure the robustness and reliability of the final predictive model. The project will also necessitate a deep understanding of both the technical aspects of data mining and machine learning, as well as the business context of customer churn in the telecom industry.

# Description on Dataset

The data set we are working with is a customer churn data set, specifically for a telecommunications company. Customer churn, also known as customer attrition, refers to when a customer stops doing business with a company. The goal of analyzing churn is to understand the reasons why customers are leaving, and use this information to develop strategies to retain customers, thereby increasing company revenue.

This data set contains 7043 rows, each representing a unique customer, and 21 columns, each representing a different attribute or feature of the customer. The features in this data set can be broadly categorized into three types: services that the customer has signed up for, customer account information, and demographic information about the customer.

The services-related features include whether the customer has signed up for phone service, multiple lines, internet service, online security, online backup, device protection, tech support, and streaming TV and movies. These features provide insight into the types of services that customers find valuable, and can help the company understand which services are most influential in a customer's decision to stay or leave.

The account-related features include the length of time the customer has been with the company, the type of contract the customer has, their payment method, whether they use paperless billing, their monthly charges, and their total charges. These features can help the company understand the financial aspects of customer churn, such as whether higher monthly charges or total charges are associated with higher churn rates.

The demographic features include the customer's gender, age range, and whether they have partners and dependents. These features can help the company understand if certain demographic groups are more likely to churn than others, which can inform targeted marketing and retention strategies. The target variable in this data set is the "Churn" column, which indicates whether the customer left the company within the last month. This is the variable that we will try to predict using the other features in the data set. By building a model that can accurately predict customer churn based on these features, the company can proactively address customer issues and implement retention strategies to reduce churn.

# Preprocessing:

The preprocessing stage involved handling missing values, transforming categorical variables, and normalizing numerical variables. Missing values were imputed using appropriate strategies like mean imputation for numerical variables and mode imputation for categorical variables. Categorical variables were transformed using one-hot encoding to convert them into binary variables that can be processed by the algorithm. Numerical variables were normalized to ensure that they are on the same scale and that the algorithm's performance is not affected by the scale of the variables.

Below are the various steps that were used in the entire project when it comes to preparing and processing the data

* Data Loading: The data is loaded into a pandas DataFrame from a CSV file using the pd.read\_csv() function.
* Data Cleaning: The notebook checks for missing values in the dataset using the isnull().sum() function. If any missing values are found, they are handled appropriately (although this specific notebook doesn't show that part).
* Data Encoding: The notebook uses the LabelEncoder() function from the sklearn.preprocessing module to convert categorical data into numerical data. This is necessary because machine learning algorithms work better with numerical data.
* Data Splitting: The dataset is split into training and testing sets using the train\_test\_split() function from the sklearn.model\_selection module. This is done to evaluate the performance of the machine learning model on unseen data.
* Data Balancing: The notebook uses the SMOTE() function from the imblearn.over\_sampling module to balance the dataset. This is done because the dataset is imbalanced, meaning one class has significantly more samples than the other. Balancing the dataset helps improve the performance of the machine learning model.
* Feature Scaling: The notebook uses the StandardScaler() function from the sklearn.preprocessing module to standardize the features by removing the mean and scaling to unit variance. This is done because machine learning algorithms perform better when the input numerical variables fall within a similar scale.
* Dimensionality Reduction: The notebook uses the PCA() function from the sklearn.decomposition module to reduce the dimensionality of the dataset. This is done to reduce the computational complexity of the model and to avoid the curse of dimensionality.
* Model Saving: The trained model is saved using the pickle module. This allows the model to be reused later without needing to be retrained.

# DM Area and DM Algorithm:

The field of data mining encompasses a wide range of techniques and methodologies aimed at extracting meaningful information from large datasets. In this project, the focus is on predictive modeling, a data mining technique used to forecast outcomes based on historical data. Predictive modeling can be used in various fields, such as finance, marketing, healthcare, and many others, to make informed decisions about future events.

The specific type of predictive modeling used in this project is classification. Classification is a supervised learning approach where the outcome (target variable) is categorical. It involves training a model on a dataset where the class labels are known, and then using that model to predict the class labels of a new, unseen dataset. Examples of classification problems include email spam detection, customer churn prediction, and disease diagnosis.

The algorithm chosen for the classification task in this project is the Random Forest Classifier. Random Forest is a versatile and widely used machine learning algorithm that can handle both regression and classification tasks. It is an ensemble learning method, meaning it combines the predictions of multiple base estimators to improve generalizability and robustness over a single estimator.

The Random Forest algorithm works by creating a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The primary reason for using Random Forest is its ability to handle large datasets with high dimensionality. It can manage thousands of input variables without variable deletion and provides an effective method for estimating missing data. Another advantage of Random Forest is its robustness to overfitting.

Overfitting occurs when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. Random Forest mitigates this problem by training each tree on a different data sample and then averaging the predictions, which helps to reduce variance and improve the model's accuracy. In this project, the Random Forest model is trained using the SMOTE (Synthetic Minority Over-sampling Technique) dataset and then evaluated on a test set.

The model's performance is measured using various metrics, such as precision, recall, and F1-score. The trained model is then saved using the pickle module in Python for future use. In conclusion, this project demonstrates the application of the Random Forest Classifier in predictive modeling, showcasing its ability to handle complex classification tasks with high accuracy and generalizability. The use of ensemble methods like Random Forest is becoming increasingly popular in the field of data mining due to their robustness and ability to handle large, high-dimensional datasets.

# Scenarios (Exploration of the Effect of k in k-means clustering)

While the primary focus of the project is on using the Random Forest Classifier, an exploration of the effect of the number of clusters (k) in k-means clustering is also conducted. This involves running the k-means algorithm with different values of k and analyzing the resulting clusters' characteristics and quality.

# Results and Analysis

The Random Forest Classifier indeed achieved a high accuracy score, which is a good indication of its effectiveness in predicting customer churn. This model was trained using the SMOTE technique, which helps to balance the dataset by creating synthetic instances of the minority class. This technique can significantly improve the performance of the model on imbalanced datasets, which seems to be the case here.

The model was also evaluated using various metrics such as precision, recall, and f1-score, which provide a more comprehensive understanding of its performance. Precision is the ratio of correctly predicted positive observations to the total predicted positives. High precision relates to the low false positive rate. Recall (Sensitivity) - the ratio of correctly predicted positive observations to the all observations in actual class.

The F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. It is a better measure than accuracy especially for the imbalanced classes. In terms of k-means clustering, the choice of k (the number of clusters) is indeed a critical factor that can significantly impact the characteristics and quality of the resulting clusters.

A common technique for choosing k is the elbow method, which involves plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use. Other methods include the silhouette method and the gap statistic. In this case, PCA was used to reduce the dimensionality of the dataset before applying k-means clustering. This can often improve the results of the clustering by removing noise and redundant features, making the data easier to visualize and interpret. Overall, the combination of these techniques and the careful choice of parameters have led to a robust and effective model for predicting customer churn.

# Conclusion

The project demonstrated the effectiveness of data mining techniques in predicting customer churn in the telecom industry. The results can inform customer retention strategies by identifying customers who are likely to churn

# References

* Gupta, M. K., & Chandra, P. (2020). A comprehensive survey of data mining. International Journal of Information Technology, 12(4), 1243-1257.
* Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. Expert Systems with Applications, 166, 114060.
* Genuer, R., Poggi, J. M., Genuer, R., & Poggi, J. M. (2020). Random forests (pp. 33-55). Springer International Publishing.
* Pradipta, G. A., Wardoyo, R., Musdholifah, A., Sanjaya, I. N. H., & Ismail, M. (2021, November). SMOTE for handling imbalanced data problem: A review. In 2021 sixth international conference on informatics and computing (ICIC) (pp. 1-8). IEEE.
* Peng, J., Wu, W., Lockhart, B., Bian, S., Yan, J. N., Xu, L., ... & Wang, J. (2021, June). Dataprep. eda: Task-centric exploratory data analysis for statistical modeling in python. In Proceedings of the 2021 International Conference on Management of Data (pp. 2271-2280).