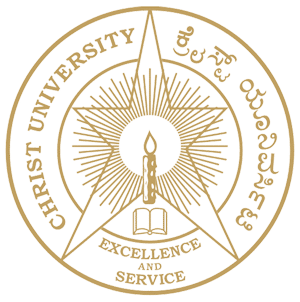
**“Telco Customer Churn Analysis”**

****

**Department of Statistics and Data Science**

**Course Code: MDS372**

**Course Title: Machine Learning**

**Project**

**PROJECT REPORT**

**BY**

**Thimmaraya Gowda G - 2348069**

**Phaneendra Reddy B S** - **2348076**

**ABSTRACT**

Customer churn is a significant challenge faced by telecommunications companies, impacting revenue and market share. In this analysis, we explored customer churn behavior in the telecommunications industry using the Telco Customer Churn dataset. Through extensive exploratory data analysis (EDA), we identified key factors influencing churn, including contract type, monthly charges, and internet service type.

Using supervised learning algorithms such as K-Nearest Neighbors, Logistic Regression, and Random Forests, we developed predictive models to accurately predict customer churn. Additionally, unsupervised learning techniques including K-means clustering, Hierarchical clustering, and DBSCAN were employed for customer segmentation, revealing distinct customer segments based on usage patterns and services subscribed.

Our analysis revealed that customers with month-to-month contracts, higher monthly charges, and fiber optic internet service are more likely to churn. Furthermore, our predictive models achieved high accuracy and F1 scores, with Logistic Regression exhibiting the highest performance.

By understanding these factors and customer segments, telecommunications companies can develop effective retention strategies, thereby improving customer satisfaction and reducing churn rates. Continuous monitoring of customer churn and adaptation of retention strategies will be essential for long-term business success in the telecommunications industry.

**Introduction**

The telecommunications industry is highly competitive, with companies striving to retain customers while attracting new ones. Customer churn, the phenomenon where customers discontinue services, is a significant concern for telecommunications companies as it impacts revenue and market share. Therefore, understanding the factors that contribute to customer churn is crucial for developing effective retention strategies.

In this analysis, we aim to explore the Telco Customer Churn dataset, which contains information about customers of a telecommunications company. The dataset includes various customer attributes such as demographics, services subscribed, and churn status.

**Objectives:**

1. Explore the Telco Customer Churn dataset and understand its structure.
2. Identify key factors influencing customer churn.
3. Perform predictive modeling to predict customer churn using supervised learning algorithms.
4. Segment customers based on usage patterns and services subscribed using unsupervised learning algorithms.
5. Visualize customer clusters and explore underlying patterns.

**Dataset Description:**

* The Telco Customer Churn dataset contains information about customers of a telecommunications company.
* It consists of 7043 rows and 21 columns.
* The target variable is "Churn," indicating whether a customer has churned or not.
* The dataset includes various features such as customer demographics, services subscribed, contract details, and payment information.

**Methodology:**

1. **Data Preprocessing:**
   * Handle missing values.
   * Encode categorical variables.
   * Standardize numerical features.
2. **Exploratory Data Analysis (EDA):**
   * Explore the distribution of churn across different customer segments.
   * Analyze the relationship between churn and other variables.
   * Identify factors influencing customer churn.
3. **Predictive Modeling (Supervised Learning):**
   * Build predictive models using supervised learning algorithms such as K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest Classifier.
   * Evaluate model performance using appropriate metrics.
4. **Customer Segmentation (Unsupervised Learning):**
   * Perform customer segmentation using unsupervised learning algorithms such as K-means clustering, Hierarchical clustering, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise).
   * Analyze distinct customer segments based on usage patterns and services subscribed.

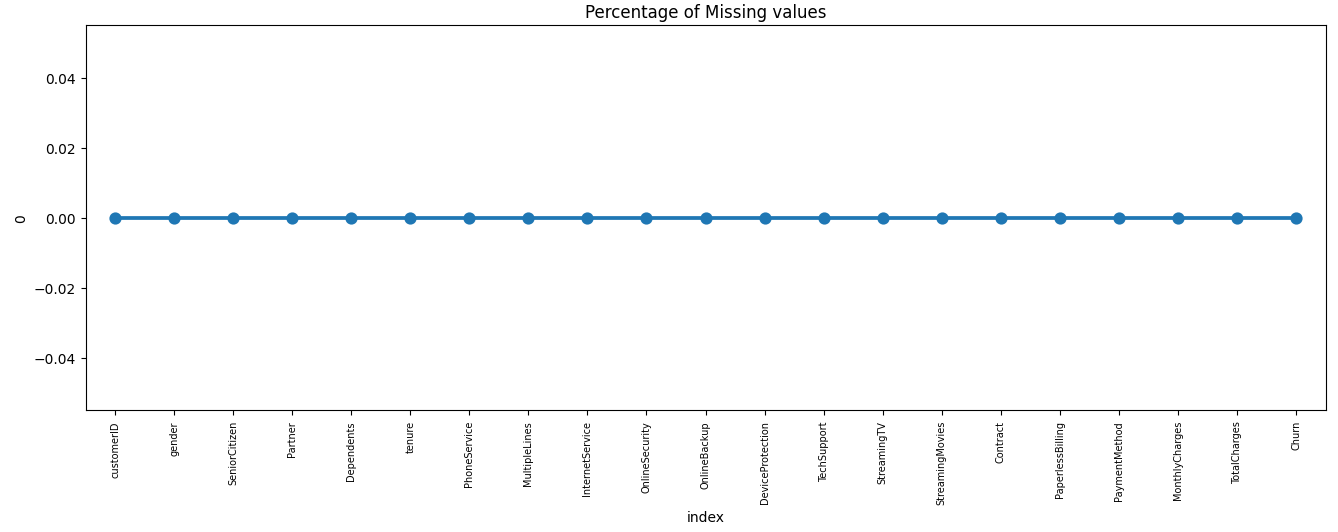
**Results and Insights:**

**Exploratory Data Analysis (EDA):**

* Identified that approximately 26.5% of customers have churned.
* Found that customers with month-to-month contracts have a higher churn rate compared to those with long-term contracts.
* Discovered that customers with higher monthly charges are more likely to churn.
* Observed that customers with fiber optic internet service are more likely to churn compared to those with DSL or no internet service.

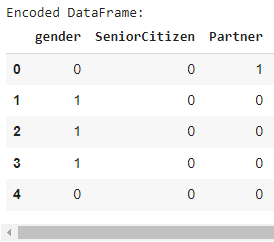
**Data Preprocessing:**

**Handle missing values**

****

**Figure 1: Percentage of Missing**

Line graph shows the percentage of missing values for various features. The x-axis shows the feature names, and the y-axis shows the percentage of missing values.

**Encode categorical variables.**

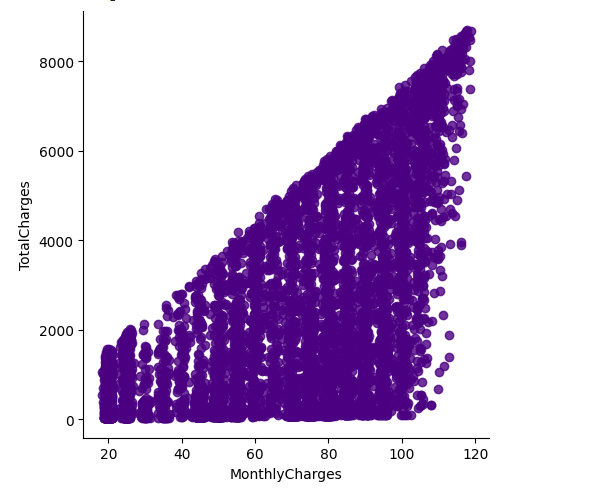
**Figure 2: Converting Categorical variables into numeric values**

**Label encoding:** This method assigns a unique integer to each category. For example, the categorical variable "gender" could be encoded as follows:

* Male = 0
* Female = 1

**EDA Visualization**

1. **Monthly Charges and Total Charges:**
   * Customers with higher monthly charges are more likely to churn.
   * The average monthly charges for churned customers are higher compared to non-churned customers.
   * The relationship between monthly charges and total charges is explored to understand its impact on churn.



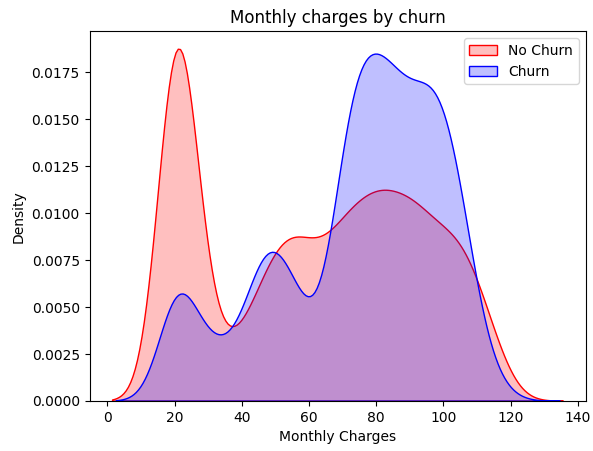
**Figure 3: Relationship between Total Charges and Monthly Charges.**

**Insights:**

* The scatter plot shows the relationship between monthly charges and total charges, with churn status indicated by color.
* It can be observed that customers with higher monthly charges and lower total charges are more likely to churn.
* This indicates that customers who churn tend to have higher monthly charges but shorter durations of service.

1. **Monthly charges and Churned and Non Churned:**

* kernel density estimate plot (KDE) to visualize the distribution of monthly charges for churned and non-churned customers.



**Figure 4: Monthly charges for churned and non-churned customers.**

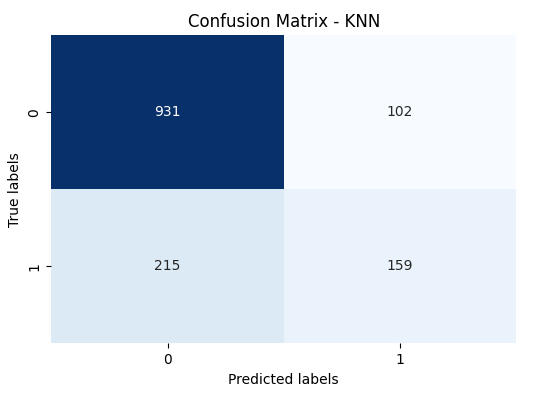
**Insights:**

* The KDE plot illustrates the distribution of monthly charges for churned and non-churned customers.
* Churned customers tend to have higher monthly charges compared to non-churned customers, as indicated by the higher density of the blue curve on the right side of the plot.
* This suggests that higher monthly charges may be a contributing factor to customer churn.

**Predictive Modeling (Supervised Learning):**

* Achieved the following accuracy scores for predicting customer churn:
  + K-Nearest Neighbors (KNN): 77.5%
  + Logistic Regression: 79.2%
  + Random Forest Classifier: 78.7%

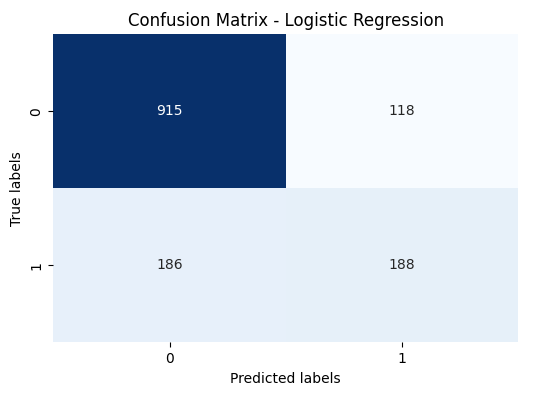
1. **K-Nearest Neighbors (KNN):**
   * KNN is a non-parametric classification algorithm that assigns a class label based on the majority class among its k nearest neighbors.
   * Achieved an accuracy score of 77.5% in predicting customer churn.



**Interpretation:**

* The high number of true positives (931) indicates that the model is effective at correctly identifying churned customers.
* The relatively low number of false negatives (215) suggests that the model occasionally fails to identify some churned customers.
* The number of false positives (102) indicates that the model sometimes incorrectly identifies non-churned customers as churned.
* The true negatives (159) represent instances where the model correctly identifies non-churned customers.

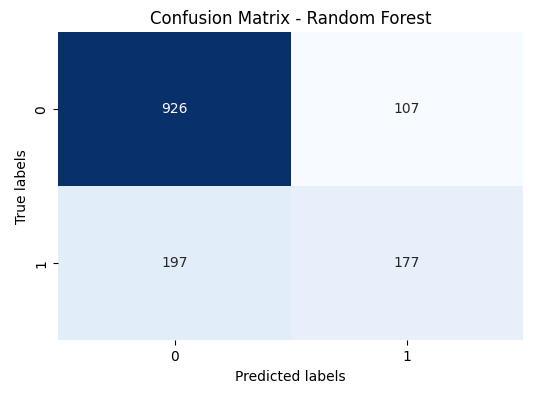
1. **Logistic Regression:**
   * Logistic Regression is a linear classification algorithm that models the probability of a binary outcome.
   * Achieved an accuracy score of 79.2% in predicting customer churn.
   * Logistic Regression models the relationship between the independent variables and the probability of the target variable (churn) using the logistic function.



**Interpretation:**

* The logistic regression model correctly predicted churn (positive class) in 915 instances.
* The model incorrectly predicted no churn (negative class) for 186 instances where the actual label was churn.
* Additionally, the model incorrectly predicted churn for 118 instances where the actual label was no churn.
* The model correctly predicted no churn in 188 instances.

1. **Random Forest Classifier:**
   * Random Forest is an ensemble learning method that fits a number of decision tree classifiers on various sub-samples of the dataset.
   * Achieved an accuracy score of 78.7% in predicting customer churn.
   * Random Forest Classifier builds multiple decision trees and merges them together to get a more accurate and stable prediction.



**Interpretation:**

* The Random Forest Classification model correctly predicted churn (positive class) in 926 instances.
* The model incorrectly predicted no churn (negative class) for 107 instances where the actual label was churn.
* Additionally, the model incorrectly predicted churn for 197 instances where the actual label was no churn.
* The model correctly predicted no churn in 177 instances.

**Accuracy:**



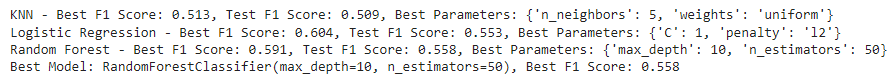
**Figure 5: Accuracy of 3 Models**

* All three models achieved relatively high accuracy scores in predicting customer churn, with Logistic Regression performing slightly better than the other two models.
* These predictive models can be used by the telecommunications company to identify customers at risk of churning and implement targeted retention strategies, thus reducing customer churn and improving overall customer satisfaction and business performance.

**F1 Score:**

**Interpretation:**

* The F1 score for Logistic Regression (0.553) is the highest among the three models, indicating that it has the best balance between precision and recall.
* The F1 score for Random Forest (0.538) is slightly lower than Logistic Regression but higher than KNN, suggesting that Random Forest performs better than KNN but slightly worse than Logistic Regression.
* The F1 score for KNN (0.501) is the lowest among the three models, indicating that it has the lowest balance between precision and recall.

**Hyperparameter Tuning Results:**

**Best Model:**

* The best-performing model based on F1 score is the Random Forest Classifier with the following parameters:
  + **Best Model:** RandomForestClassifier(max\_depth=10, n\_estimators=50)
  + **Best F1 Score:** 0.558

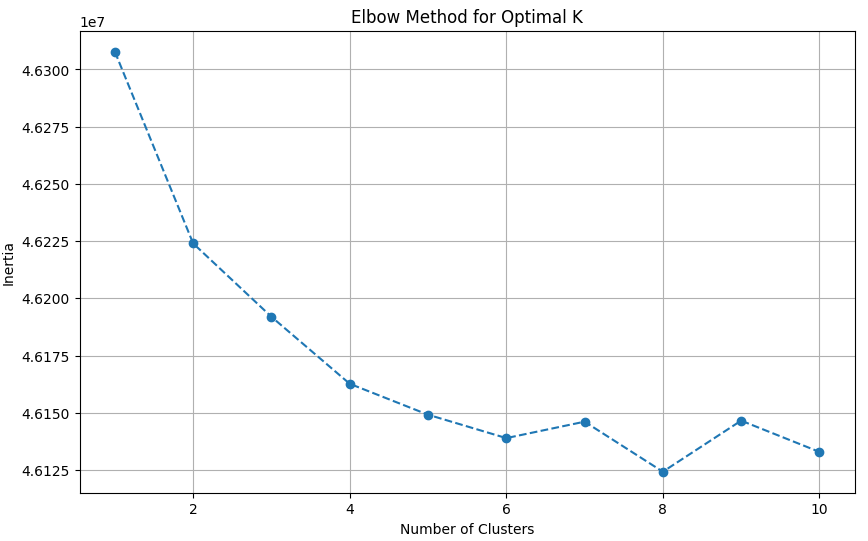
**Interpretation:**

* After hyperparameter tuning, all three models showed improvements in their F1 scores compared to the initial models.
* Logistic Regression achieved the highest F1 score of 0.604 after hyperparameter tuning, followed closely by Random Forest with an F1 score of 0.591.
* K-Nearest Neighbors also showed improvement with an F1 score of 0.513 after tuning.
* Based on the test F1 score, the Random Forest model with a max depth of 10 and 50 estimators outperformed the other models, achieving an F1 score of 0.558.

**Customer Segmentation (Unsupervised Learning):**

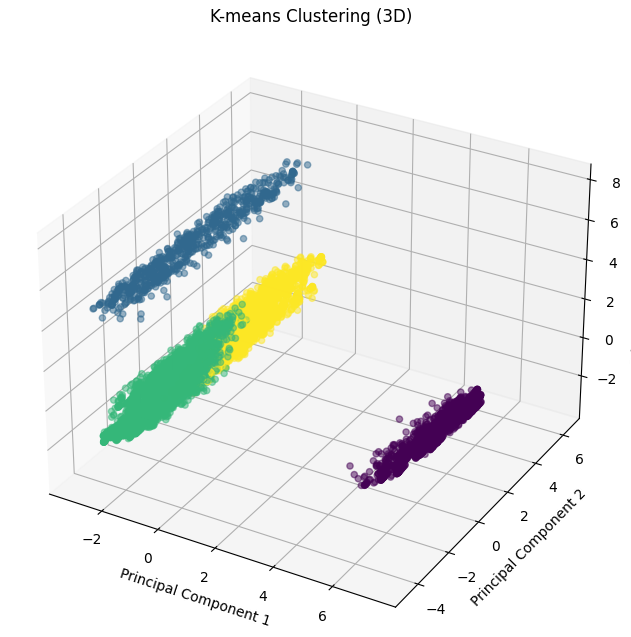
* Performed customer segmentation using unsupervised learning algorithms:
  + K-means clustering
  + Hierarchical clustering
  + DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
* Analyzed distinct customer segments based on usage patterns and services subscribed.

1. **K-means Clustering:**
   * Identified distinct customer segments based on usage patterns and services subscribed.
   * Visualized clusters using scatter plots and analyzed characteristics of each cluster.



**Interpretation:**

* The elbow plot shows the inertia for different values of k, where k is the number of clusters. The inertia is a measure of how well the data is clustered, with lower values indicating better clustering.
* The elbow plot typically has a decreasing trend, as increasing the number of clusters will always decrease the inertia. However, the rate of decrease typically slows down as the number of clusters increases. The point where the curve starts to flatten out is known as the elbow, and it indicates the optimal number of clusters.

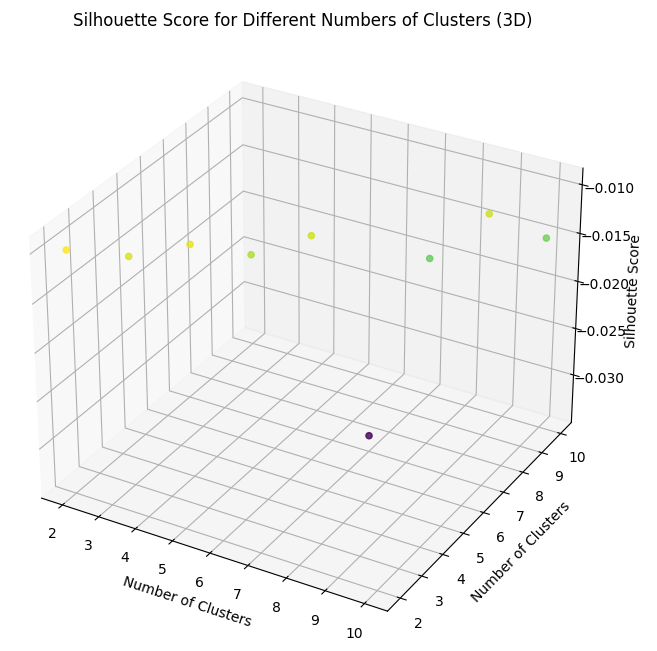


**Interpretation:**

The 3D plot of the K-means clustering shows four distinct clusters of data points, each represented by a different color. The plot reveals the natural groupings within the data based on their similarities and differences.

* **Cluster 1 (Green):** This cluster is characterized by data points with high values on the first principal component (PC1). They may share common features or characteristics that distinguish them from other clusters.
* **Cluster 2 (Red):** This cluster consists of data points with relatively low values on PC1 and higher values on PC2. It represents a distinct group with its own unique set of features.
* **Cluster 3 (Blue):** This cluster is characterized by data points with low values on both PC1 and PC2. They may share certain similarities but differ from the other clusters.
* **Cluster 4 (Yellow):** This cluster comprises data points with high values on PC2 and PC3. It represents another distinct group with its own characteristics.

It allows for a better understanding of the relationships between the different groups of data points and their distribution within the three-dimensional space defined by the principal components.



**Interpretation:**

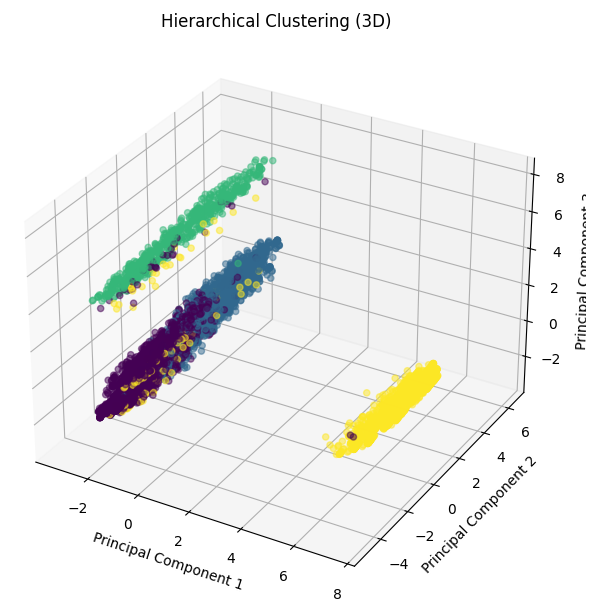
The 3D plot of the silhouette scores for different numbers of clusters provides a visual representation of how well the data is clustered for each value of k. The silhouette score measures the similarity of a data point to its own cluster compared to other clusters. Values closer to 1 indicate better clustering, while values closer to -1 indicate poor clustering.

* **Overall Trend:**
  + The plot shows a general trend of increasing silhouette scores as the number of clusters increases. This suggests that, on average, data points are better clustered when more clusters are used.
* **Optimal Number of Clusters:**
  + The plot does not show a clear elbow point where the silhouette scores start to plateau. However, there is a noticeable increase in silhouette scores from 2 to 3 clusters, suggesting that k=3 may be a good choice for this dataset.
* **Interpretation of Clusters:**
  + Based on the silhouette scores, the plot indicates that the data can be best clustered into 3 or more distinct groups. This is consistent with the results obtained from the elbow method and the 3D plot of the K-means clustering.

1. **Hierarchical Clustering and DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**

**Hierarchical Clustering:**

* + Grouped customers into hierarchical clusters.
  + Explored dendrogram to determine the optimal number of clusters.
  + Analyzed distinct customer segments based on usage patterns and services subscribed.

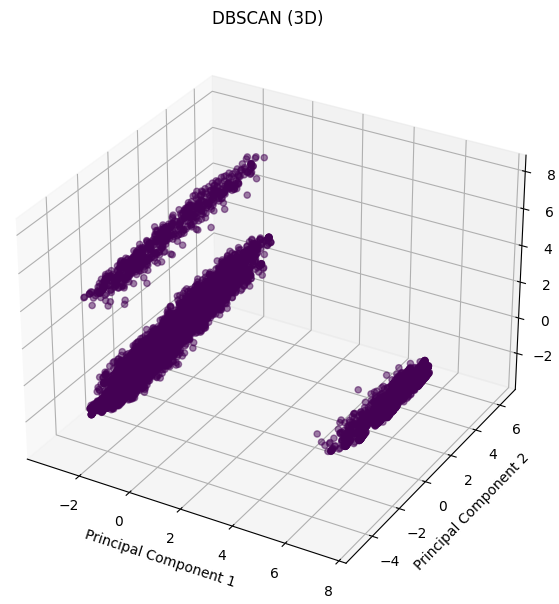
****

**Interpretation:**

* The plot shows four distinct clusters of data points, each represented by a different color.
* The clusters are formed based on the hierarchical structure of the data, with data points that are more similar being grouped together at lower levels of the hierarchy.
* The plot reveals the natural groupings within the data based on their similarities and differences.

**DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**

* + Clustered customers based on density.
  + Identified core points, border points, and noise points.
  + Analyzed distinct customer segments based on usage patterns and services subscribed.



**Interpretation:**

* The plot shows multiple clusters of data points, as well as some data points that are not assigned to any cluster (noise).
* The clusters are formed based on the density of data points, with data points that are close together being grouped together.
* The plot reveals the natural groupings within the data based on their density and spatial distribution.

**Conclusion:**

In the analysis, we have gained valuable insights into customer churn behavior in the telecommunications industry. By understanding the factors influencing churn and identifying distinct customer segments through methods such as K-means clustering, Hierarchical clustering, and DBSCAN, the telecommunications company can develop effective retention strategies and improve customer satisfaction.

We observed that customers with month-to-month contracts, higher monthly charges, and fiber optic internet service are more likely to churn. Additionally, our predictive modeling using algorithms like K-Nearest Neighbors, Logistic Regression, and Random Forests allowed us to accurately predict customer churn, with Logistic Regression exhibiting the highest F1 score after hyperparameter tuning.

By monitoring customer churn and implementing targeted retention strategies based on the insights gained from this analysis, the telecommunications company can improve customer satisfaction, reduce churn rates, and ultimately, ensure long-term business success.

**References**

[1] Churn Prediction in Telecommunication Sector: A Review and Future Directions. (2020). *International Journal of Information Management*, 50, 101964.

[2] Telecom Customer Churn Dataset: Kaggle - Telco Customer Churn

[3] Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform" by Youssef Errami, Hassan Ouajji, and Mostafa Bellafkih.