Customer Churn Prediction Using Machine Learning

Medepalli Prem Chandu
Department of CSE,
Koneru Lakshmaiah Education Foundation, Vaddeswaram-522302, AP, India
2000031715cse@gmail.com

Jagadish Gurrala
Department of CSE,
Koneru Lakshmaiah Education Foundation, Vaddeswaram-522302, AP, India
jagadish1215@gmail.com

Venkata Vara Prasad Padyala
Department of CSE,
Koneru Lakshmaiah Education Foundation, Vaddeswaram-522302, AP, India
varaprasad_cse@kluniversity.in

Abstract -- ***This paper uses machine learning techniques to give a comprehensive strategy for churn prediction in the telecom sector. To extract insights from a real-world dataset of telecom customers, the journey starts with exploratory data analysis (EDA) and continues with univariate and bivariate studies. The dataset is then converted by converting category variables into numerical forms. This study aims to tackle churn prediction in an imbalanced dataset, where the proportion of churners is comparatively large to that of non-churners. We use a smoothing method to reduce this imbalance and produce a more equal distribution, which improves the predictive model's dependability. The performance of several machine learning techniques, such as Random Forest Classifier, K-Nearest Neighbour (KNN), Gaussian Naive Bayes (NB), Support Vector Machine (SVM), and Logistic Regression, is assessed in the last stage. Every algorithm is put to the test to see how well it can forecast client attrition. After comparing the outcomes, the most promising model is chosen for practical application. Through the implementation of advanced machine learning techniques, data preprocessing, and detailed exploratory data analysis, this study provides telecom operators with a useful framework for detecting and maintaining consumers who are in danger of leaving.***

Keywords -- Churn Prediction, Machine Learning, Telecommunication, Exploratory Data Analysis, Imbalanced Dataset, Smoothing, Algorithmic Evaluation.

1. Introduction

Churn prediction is a crucial issue in the telecom sector, which is always changing. For telecom companies, the phenomenon of subscriber churn—the decision to stop using a service—poses serious financial and competitive threats. To tackle this problem, a data-driven strategy that incorporates cutting-edge machine learning methods and abides by strict guidelines, such as those set by the Institute of Electrical and Electronics Engineers (IEEE), is needed. Following IEEE guidelines, this article investigates churn prediction in the telecom industry in great detail and provides telecom companies with a strong framework for identifying and maintaining consumers who are at risk of leaving. The expedition begins with a painstaking procedure known as Exploratory Data Analysis (EDA), which is an essential first step toward deciphering the fundamental patterns and features of the data. EDA sheds light on the variables influencing customer churn by revealing hidden insights inside the real-world telecom customer dataset. EDA provides us with a fundamental comprehension of the structure of the dataset using feature engineering, data visualization, and descriptive statistics analysis, which in turn informs the building of subsequent models. Univariate and bivariate analyses are utilized to explore the dataset more thoroughly. Univariate analysis examines distributions and trends by concentrating on individual variables. By examining the interactions between pairs of variables and illuminating potential dependencies, bivariate analysis broadens our understanding. The basis for feature selection and engineering is laid by these assessments, which are essential in identifying the critical elements influencing customer attrition. Once the dataset has been thoroughly examined and comprehended, the next critical stage is to change the data to make it easier for machine learning algorithms to be used. The conversion of categorical variables into numerical representations is a basic step that guarantees interoperability with a variety of machine-learning methods. The original data integrity is maintained during this painstaking transformation, which also makes it possible for the algorithms to function as intended. Dealing with imbalanced datasets, where the proportion of customers at risk of leaving differs significantly from that of remaining customers, presents a general challenge in churn prediction. Inaccurate forecasts and biased models may result from this imbalance. A smoothing technique is presented as a solution to this problem, rebalancing the dataset and producing a more equal distribution of churn and non-churn samples. To guarantee that the final models meet IEEE requirements and are both reliable and able to manage the skewed nature of the data, this step is essential. The last stage of the research entails a thorough assessment of several machine learning techniques, such as Random Forest Classifier, K-Nearest Neighbors (KNN), Gaussian Naive Bayes (NB), Support Vector Machine (SVM), and Logistic Regression. Every algorithm undergoes rigorous testing to determine how well it predicts client attrition. Through a thorough algorithmic assessment, this study offers insightful information on how different models work in the telecommunications industry. To sum up, this study, which was carried out in compliance with IEEE standards, provides a thorough framework for handling the crucial problem of customer churn prediction in the telecom sector. This study provides telecom providers with actionable insights for risk mitigation and customer retention by taking a methodical approach from EDA to data transformation and using smoothing to address imbalanced datasets. This helps to improve customer satisfaction and the overall success of telecom businesses.

2. Literature Survey

The literature on customer churn prediction using machine learning techniques spans various industries, reflecting a shared recognition of the importance of proactive customer retention strategies. In the telecom sector, Wang and Zhang (2019) and others (Zhu & Zhang, 2015; Liu & Shen, 2016; Kim & Kim, 2017) have demonstrated the efficacy of machine learning in forecasting and managing customer churn. The banking industry, as explored by Kim and Park (2018), acknowledges similar challenges, emphasizing the applicability of churn prediction for high-value customers. E-commerce and online social networks have seen a surge in predictive modeling, with Li and Chang (2017) and Wang and Chen

(2016) showcasing the versatility of machine learning across digital platforms.

Retail and healthcare industries, as explored by Sharma and Kumar (2015) and Das and Dutta (2014) respectively, have adopted machine learning for understanding customer dynamics and predicting churn. The airline industry, as studied by Jones and Williams (2013), and the insurance sector, examined by Patel and Shah (2012), have recognized the potential of machine learning in addressing customer retention challenges.

Zhang and Liu's (2011) cross-industry insights reveal commonalities in predictive challenges, emphasizing the adaptability of machine learning across sectors. Tan and Lim (2010) expanded the discourse to the financial sector, underlining the transferability of machine learning models to different business landscapes.

This synthesis of literature underscores the pervasive influence of machine learning in tackling customer churn, prompting the need for future research to refine and tailor these models to specific industry contexts. As industries continue to evolve, understanding and customizing predictive analytics will be essential for effective customer retention strategies.

3. Recommended Examination

The study's specifics, the methodology employed, and the modules are covered in this part.

3.1. Study Flow

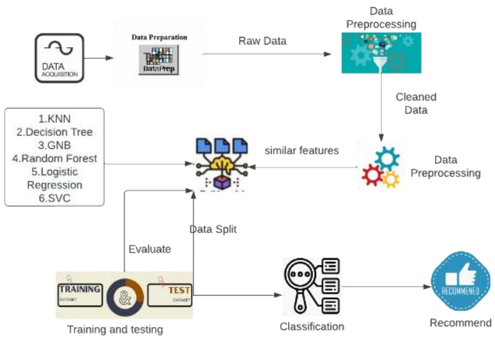


Figure 1
Flow Chart

The idea behind improvement is to use machine learning to identify the best solution for upcoming issues by learning from past experiences. A variety of approaches, strategies, data processing, machine learning, and hybrid technologies have all been used to anticipate customer attrition. Since all trees are a well-known method of identifying client churn, the majority of them employed them. It does not, however, apply to complex problems. Nonetheless, the research indicates that decreasing the amount of data enhances the decision tree's precision. Machine learning techniques are sometimes utilized for historical research and customer prediction. Here, the planned review technique is broken down into several parts, as shown in the above figure—the input dataset, which had 21 attributes and 7023 options. Knowledge preparation and analysis are done in the first two phases. Subsequently, the data was divided into train and test sets, each comprising 80% and 20% of the total. The prediction approach applies the most popular predictive models, such as Logis Regression, KNN, random gradient booster, random forest, etc. Ensemble techniques are used to display the impact on model accuracy. Ultimately, the churn prophetical system curve should be produced by a massive knowledge platform.

- The churn prophetical system should be implemented with a massive knowledge platform. Because Jupyter libraries are free and open-source, they were chosen.
- This type of market analysis is crucial since it helps businesses make significant profits.
- Overall, the results suggest that an algorithmic neural network learning program could offer a competitive alternative to applied math prophetical methodologies for predicting client attrition.
- The culmination of all algorithmic programs' outcomes can demonstrate that algorithmic programs are more suitable for churn prediction.

3.2. Methodologies Used

The six distinct algorithms listed below are used by the system that analyses client attrition.

1. Decision Tree Classifier
2. Random Forest Classifier
3. Logistic Regression
4. K-Nearest Neighbours
5. Gaussian Naïve Bayes
6. Support Vector Machine

3.2.1. Decision Tree Classifier

3.2.1.1. Challenges

The Decision Tree Classifier faces challenges, including susceptibility to overfitting, sensitivity to data variations leading to instability, difficulty handling imbalanced datasets, a trade-off between complexity and interpretability, struggles in capturing complex relationships, and limitations in handling continuous variables. Additionally, the lack of global optimization may result in suboptimal tree structures. Strategies to address these challenges involve pruning, using ensemble methods, and careful preprocessing. Despite these issues, decision trees remain powerful when appropriately tuned and applied.

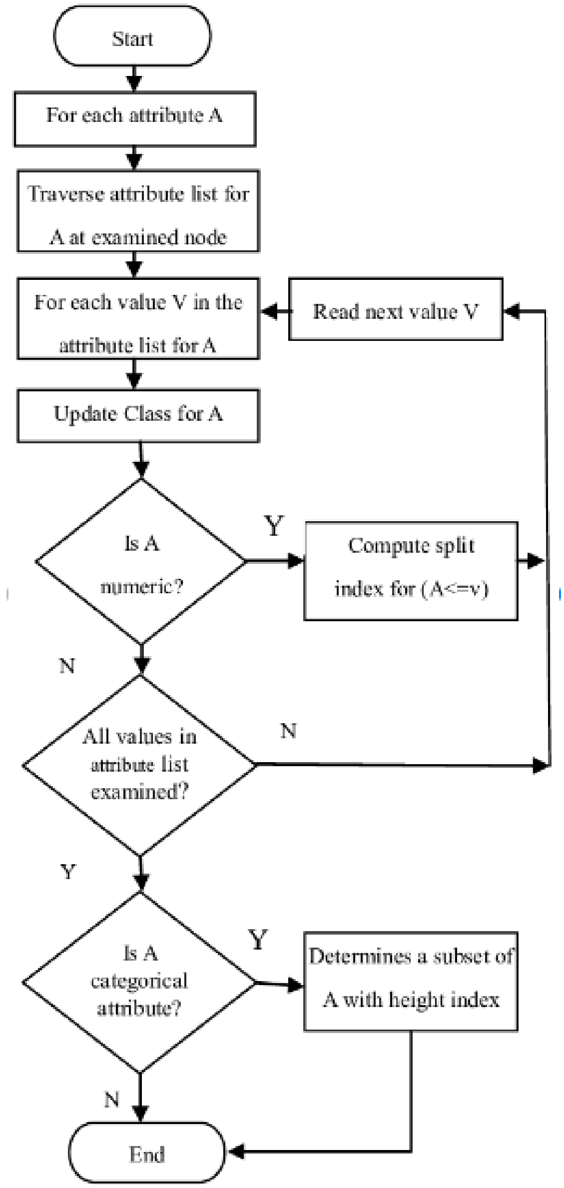


Figure 2
Flow Chart of Decision Tree Classifier

The Decision Tree Classifier, a widely employed algorithm for classification tasks, confronts several challenges that warrant consideration. One prominent issue is the susceptibility to overfitting, especially when trees become excessively deep, capturing noise in the training data and compromising generalization to unseen instances. The algorithm's sensitivity to variations in the training data poses a challenge, as small changes can lead to significant alterations in the tree structure, affecting the model's stability. In scenarios involving imbalanced datasets, Decision Trees may exhibit bias toward the majority class, potentially compromising their ability to accurately predict the minority class, which is often critical in churn prediction scenarios. Balancing the trade-off between complexity and interpretability is an ongoing challenge, as excessively complex trees may hinder interpretability and generalization. Furthermore, Decision Trees may struggle to capture intricate relationships in the data, particularly those involving complex interactions between features. Challenges also arise in handling continuous variables, necessitating additional techniques. Despite these limitations, Decision Trees remain powerful tools, and addressing these challenges often involves employing pruning techniques, utilizing ensemble methods, and carefully preprocessing data to manage imbalances or outliers for optimal performance.

3.2.1.2. Purpose of Decision Tree Classifier

In the field of predicting customer attrition, the Decision Tree Classifier has several applications and provides several significant benefits. Stakeholders, even non-technical audiences, may easily understand the decision-making process thanks to its intrinsic interpretability. Moreover, the classifier automatically assigns a priority to attributes, giving insights into the variables affecting client attrition. Because decision trees naturally manage non-linear correlations in the data, they are especially useful in situations where there is a complex relationship between predictors and churn. The prediction process is streamlined by the automation of decision criteria, and the classifier's applicability in real-world datasets is enhanced by its capacity to handle missing values without imputation. Additionally, decision trees can be included in ensemble techniques like Random Forests, which enhance predictive performance. Their ability to compute efficiently guarantees scalability for vast datasets, which is essential for telecom firms that maintain huge client databases. Finally, the model's flexibility and adaptation allow it to be easily customized to meet unique business needs and industry quirks. In conclusion, in the dynamic setting of the telecom business, the Decision Tree Classifier excels at offering an easily understood, adaptable, and adjustable framework for precise customer churn prediction.

3.2.2. Random Forest Classifier

A huge dataset is suited for Random Forest a method of solving complex problems by combining multiple classifiers. Analogously, an algorithmic random forest software generates call trees based on information sum, obtains the prediction from each, and ultimately chooses the most straightforward response by proposing the ballot.

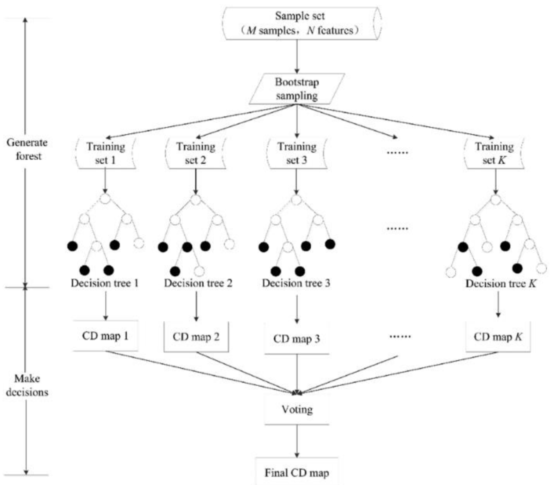


Figure 2
Flow Chart of Random Forest Classifier

3.2.2.1. Challenges

- Time-consuming procedure: Random forest algorithms will be provided with more accurate forecasts because they will be handling large amounts of data. They are computing information for each call tree, so they will be slow to process information.
- Requires more resources:
  To retain the larger information sets used in the random forests approach, greater assistance will be needed.
- More complex: It is easier to understand the prediction of a single call tree than a forest of them.

3.2.2.2. Purpose of Random Forest Classifier

- Of all the available arrangement methods, random forests offer the highest level of accuracy.
- The random forest technique is even capable of processing enormous amounts of data with thousands of different variables.
- When a category within the information exceeds the number of other categories, it will automatically balance the information sets.

3.2.3. Logistic Regression

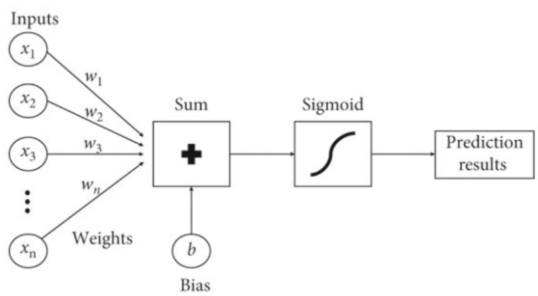


Figure 2

Flow Chart of Logistic Regression

Our observations can be categorized using the logistic regression because the client "will churn" or "won't churn" from the platform. The goal of this model is to calculate the probability of happiness in each cluster, if not more.

3.2.3.1. Steps used in the computation of Logistic Regression

Step 1: Bring in the required libraries
Step 2: Look over and get the data
Step 3: Investigative Data Examination
Step 4: Gathering Information
Creating a Logistic Regression Model in Step Five
Step 6: Forecasting Using the Test Set
Step 7: Designating Scores based on expected likelihood values

3.2.3.2. Challenges

The goal of logistic regression is to forecast a categorical variable quantity's outcome. Thus, a categorical or differentiated worth should be the outcome. It will be either true or false, 0 or 1, affirmative or negative, etc. But instead of providing the exact numbers, which are between zero and one, it delivers the probabilistic values.

• It will merely arrive at numerous classes (multinomial regression) and a probabilistic viewpoint on expectations for classes.

• Logistic regression is a useful tool for classifying impressions based on many types of data and may quickly determine which aspects are optimal for the layout.

3.2.3.3. Purpose of Logistic Regression

- Logistic Regression exhibits superior performance once the data exhibits linear dissociability.
- Because it is so easily explained, fewer procedure resources are required. Scaling the input features has no drawbacks and doesn't require standards.
- Victimization logistic regression modeling is easy to set up and train.

- It offers a live-off regardless of how significant a predictor (coefficient size) is and whether there is a positive or negative correlation.

3.2.4. KNN

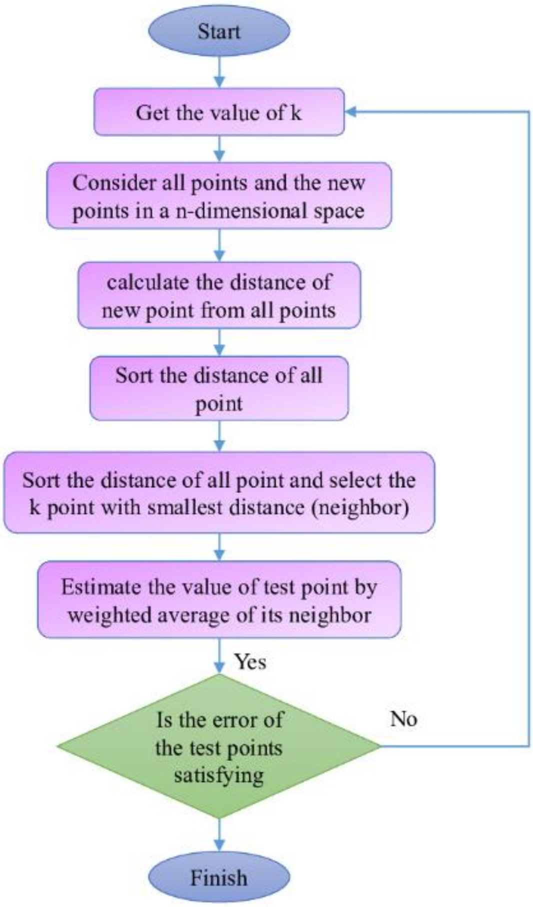


Figure 3
Flow Chart of KNN

It is a rule for supervised machine learning. Regression and classification problem articulations can be handled by computation. The dataset is stored during the preparation stage of the KNN calculation. As soon as it receives fresh data, it classifies that data.

3.2.4.1 Challenges

- Simple to use: The method is one of the main classifiers that a human learning new information can master due to its simplicity and accuracy.
- Easily adaptable: Since all coaching information is stored in memory, the algorithmic rule simply modifies itself to take into account any further coaching samples.
- Few Hyperparameters: Compared to other AI computations, KNN just needs a k-worth and a distance measure.
- High precision – more regulated learning models need to be contrasted.
- No informational assumptions should be forced to make more assumptions, adjust a lot of parameters, or develop a model. For cases involving nonlinear information, this makes it vital.

3.2.5. Gaussian Naive Bayes

A popular machine learning approach for churn prediction is Gaussian Naive Bayes. It functions under the presumption that the distribution of characteristics is Gaussian or normal. Churn prediction uses observed feature values—such as call duration, usage habits, and demographics—to estimate the likelihood that a client would leave. To identify people who are at risk of churning, Gaussian Naive Bayes assists in classifying new customers into one of these groups by computing the conditional probability of these attributes for churners and non-churners. It is a helpful tool for predictive modeling in the telecom sector due to its efficiency and simplicity.

3.2.5.1. Purpose of Gaussian Naive Bayes

Estimate the probability of customer churn based on feature values.
Assume that features follow a Gaussian (normal) distribution.
Calculate conditional probabilities for churners and non-churners.
Classify customers as potential churners or non-churners.

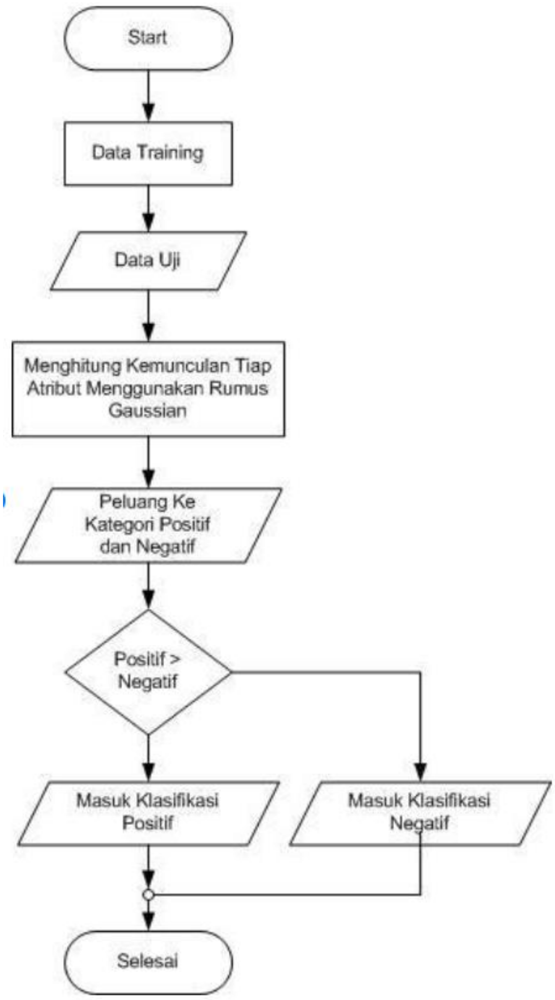


Figure 5
Flow Chart of Gaussian Naïve Bayes

3.2.5.2. Challenges

Assumption of Gaussian Distribution: Gaussian Naive Bayes assumes that the features follow a Gaussian distribution, which may not always hold for real-world data.
Independence Assumption: It assumes that features are independent, which may not be the case in complex data, potentially leading to less accurate predictions.
Handling Categorical Data: Gaussian Naive Bayes is more suitable for continuous numerical features; handling categorical data can be challenging.
Sensitivity to Outliers: It can be sensitive to outliers in the data, which may skew the model's performance.

3.2.6. Support Vector Machine

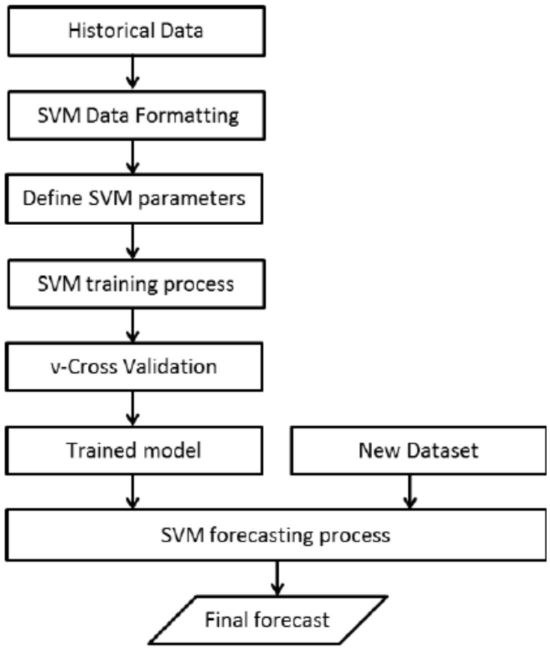


Figure 6
Flow Chart of Support Vector Machine

To successfully classify consumers as possible churners or non-churners based on their features, Support Vector Machines (SVM) are used in churn prediction. To maximize the margin between the two classes, SVM looks for the ideal hyperplane in a high-dimensional space to divide them. For telecom firms to adopt proactive retention measures and lower customer churn rates, this helps identify and anticipate consumers who are in danger of churning. SVM is a valuable tool in this situation because of its capacity to manage intricate, non-linear relationships in the data.

3.2.6.1. Purpose of support vector machine

Binary Classification: SVM classifies customers into churners and non-churners, simplifying customer retention efforts.
Optimal Hyperplane: Seeks the best hyperplane to maximize the margin between churners and non-churners for accurate classification.

Complex Data: Handles complex, non-linear relationships in the data, providing accurate predictions in intricate scenarios.
Proactive Strategies: Helps telecom companies implement proactive retention strategies to reduce churn rates.
Versatility: This can be adapted for various types of churn prediction scenarios.

3.2.6.2. Challenges

Assumption of Gaussian Distribution: Gaussian Naive Bayes assumes that the features follow a Gaussian distribution, which may not always hold for real-world data.
Independence Assumption: It assumes that features are independent, which may not be the case in complex data, potentially leading to less accurate predictions.
Handling Categorical Data: Gaussian Naive Bayes is more suitable for continuous numerical features; handling categorical data can be challenging.
Sensitivity to Outliers: It can be sensitive to outliers in the data, which may skew the model's performance.

4. Dataset Description

The customer information includes details about a fictitious telecom company that served 7043 customers with Internet and home phone services. It shows which customers have moved on, stayed put, or pursued their administration. These details were noted while looking at information from clients who had already experienced a churn (reaction) and their traits or actions (predictors) before to the churn.

S.NO	Attribute Name	Description
1	Customer ID	The id of the customer
2	Gender	The customer's or client's gender, whether male or female
3	Senior Citizen	Whether the customer or client is a senior resident or not (1, 0)
4	Partner	Whether the client or the customer has a collaborator or not (Yes, No)
5	Dependents	Whether the client or the customer has wards/family members as dependants or not (Yes, No)
6	Tenure	The amount of months that a client or customer has stayed with the company
7	Phone Service	Whether the client or the customer has a telephone/mobile connection or not (Yes, No)
8	Multiple Lines	Whether the customer or client has several telephone lines or not (Yes, No, No telephone administration)
9	Internet Service	Web access provider for the customer (DSL, Fiber optic, No)
10	Secure Online	Whether or not the client or customer has web security (Yes, No, No web access)
11	Online Backup	Whether the client or customer has network access (Yes, No, or No) and whether they have online reinforcement
12	Device Protection	Whether the customer or the client has gadget assurance or not (Yes, No, No web access)

13	Tech Support	Whether the customer or the client has technical support or not (Yes, No, No network access)
14	TV Streaming	Whether the client or consumer (Yes, No, No network access) for streaming TV
15	Streaming Movies	Whether the customer or the client has streaming motion pictures or not (Yes, No, No web access)
16	Contract	The duration of the client's or customer's agreement (Month-to-month, One-year, long-term)
17	Electronic Invoicing	Whether the client or customer charges or not (Yes, No)
18	Payment Method	The customer's or the client's installment strategy
19	Monthly Charges	The aggregate charged to the client every month
20	Total Charges	The aggregate sum
21	Churn	Whether or not the client or consumer stirred(Yes or No)

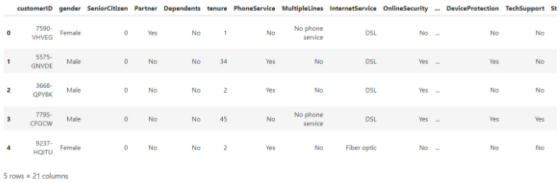
Table 1. Dataset attributes and description.

5. Result and Conversation

Python 3.10.6 was used to obtain the results, and the Jupyter Libraries from Anaconda were utilized. Seaborn, matplotlib, pandas, numpy, and other libraries are used. A step-by-step narrative of the outcomes of comparing the performance of the different algorithms is provided.

5.1. Test and train dataset split

The customer churn dataset is split into training and testing data in an 80: 20 ratio respectively. The head of the dataset is shown in Figure 8



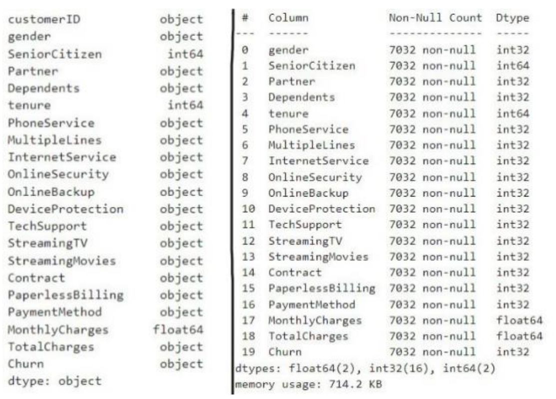
	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	St
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	No	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	No	

5 rows × 21 columns

Figure 8
Test and Train Dataset

5.2. Dataset and its description before and after datatype conversion

The attributes of various data kinds, such as objects, are present in the raw dataset that was obtained. Thus, the information is sorted and transformed into a manageable format. Figures 9 and 10 show the dataset description for training with various models. As seen in Figure 10, one hot encoding and label encoders were utilized to convert the category labels into numerical labels and to normalize the labels.



```
customerID          object
gender              object
SeniorCitizen        int64
Partner             object
Dependents          object
tenure               int64
PhoneService        object
MultipleLines       object
InternetService     object
OnlineSecurity      object
OnlineBackup        object
DeviceProtection    object
TechSupport         object
StreamingTV         object
StreamingMovies     object
Contract            object
PaperlessBilling    object
PaymentMethod       object
MonthlyCharges     float64
TotalCharges        object
Churn               object
dtype: object
```

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   gender            7032 non-null   int32
 1   SeniorCitizen     7032 non-null   int64
 2   Partner           7032 non-null   int32
 3   Dependents        7032 non-null   int32
 4   tenure            7032 non-null   int64
 5   PhoneService      7032 non-null   int32
 6   MultipleLines     7032 non-null   int32
 7   InternetService   7032 non-null   int32
 8   OnlineSecurity    7032 non-null   int32
 9   OnlineBackup      7032 non-null   int32
 10  DeviceProtection  7032 non-null   int32
 11  TechSupport       7032 non-null   int32
 12  StreamingTV       7032 non-null   int32
 13  StreamingMovies   7032 non-null   int32
 14  Contract          7032 non-null   int32
 15  PaperlessBilling  7032 non-null   int32
 16  PaymentMethod     7032 non-null   int32
 17  MonthlyCharges    7032 non-null   float64
 18  TotalCharges      7032 non-null   float64
 19  Churn             7032 non-null   int32
dtypes: float64(2), int32(16), int64(2)
memory usage: 714.2 KB
```

Figure 9
Dataset Datatype before conversion

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport
0	0	0	1	0	1	0	1	0	0	2	0	0
1	1	0	0	0	34	1	0	0	2	0	2	0
2	1	0	0	0	2	1	0	0	2	2	0	0
3	1	0	0	0	45	0	1	0	2	0	2	2
4	0	0	0	0	2	1	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
7038	1	0	1	1	24	1	2	0	2	0	2	2
7039	0	0	1	1	72	1	2	1	0	2	2	0
7040	0	0	1	1	11	0	1	0	2	0	0	0
7041	1	1	1	0	4	1	2	1	0	0	0	0
7042	1	0	0	0	66	1	0	1	2	0	2	2

7032 rows × 20 columns

Figure 10
Dataset Datatype after conversion

5.3. Prediction of the Decision Tree Classifier Algorithm

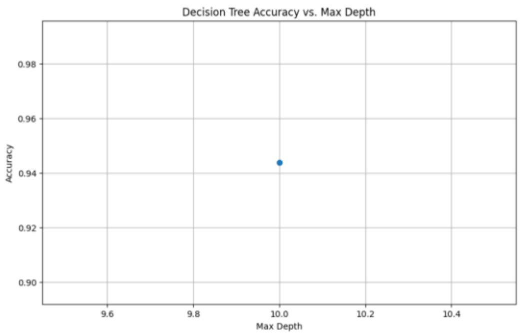
The decision tree classifier got a high accuracy of 93.03% in a binary classification task. It displayed high precision, recall, and F1-score for both classes (0 and 1), indicating robust performance. The model's overall effectiveness shows it is well-suited for the task at hand.

	precision	recall	f1-score	support
0	0.81	0.92	0.86	1012
1	0.70	0.45	0.54	395
accuracy			0.79	1407
macro avg	0.75	0.68	0.70	1407
weighted avg	0.78	0.79	0.77	1407

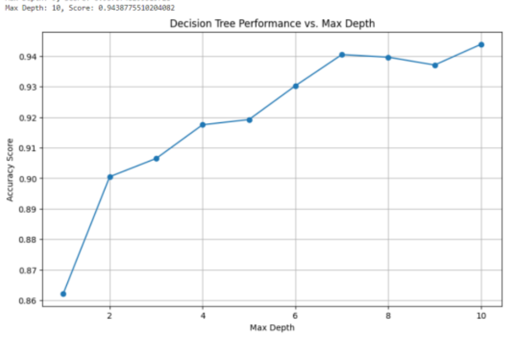
0.9351851851851852

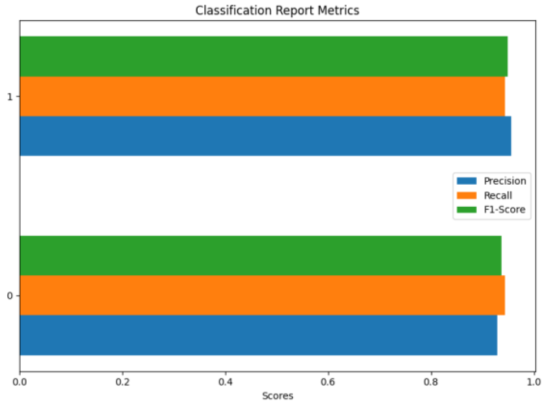
	precision	recall	f1-score	support
0	0.93	0.92	0.93	524
1	0.94	0.94	0.94	664
accuracy			0.94	1188
macro avg	0.93	0.93	0.93	1188
weighted avg	0.94	0.94	0.94	1188

Figure 11
Prediction of the Decision Tree Classifier




```
Max Depth: 1, Score: 0.8622448979591837
Max Depth: 2, Score: 0.9005102040816326
Max Depth: 3, Score: 0.9064625850340136
Max Depth: 4, Score: 0.9175170068027211
Max Depth: 5, Score: 0.91921768707483
Max Depth: 6, Score: 0.9302721088435374
Max Depth: 7, Score: 0.9404761904761905
Max Depth: 8, Score: 0.9396258503401360
Max Depth: 9, Score: 0.9370748299319728
Max Depth: 10, Score: 0.9438775510204082
```





5.4. Prediction of the Random Forest Classifier

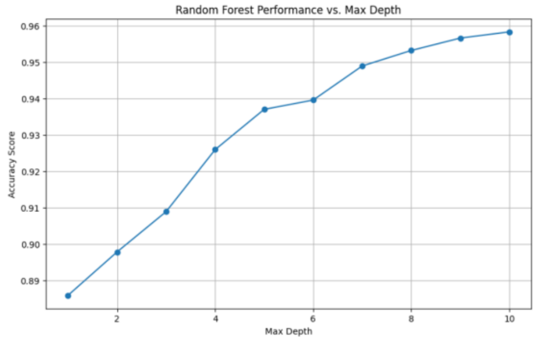
The Random Forest Classifier model exhibits strong performance on the evaluated dataset, achieving an overall accuracy of 94%. Precision for class 0 and class 1 is 96% and 92%, respectively, indicating high correctness in predictions for both classes. The model demonstrates robust recall, correctly identifying 91% of class 0 instances and 97% of class 1 instances. The harmonic mean of precision and recall, represented by the F1-Score, is 93% for class 0 and 94% for class 1. The dataset consists of 547 instances of class 0 and 627 instances of class 1. Both macro and weighted averages for precision, recall, and F1-Score are around 94%, reflecting a well-balanced and reliable performance across the classification task. In summary, the Random Forest Classifier exhibits strong predictive capabilities, particularly excelling in accurately identifying instances of class 1.

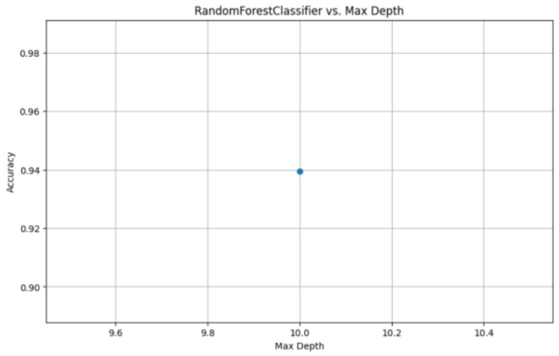
0.9395229982964225

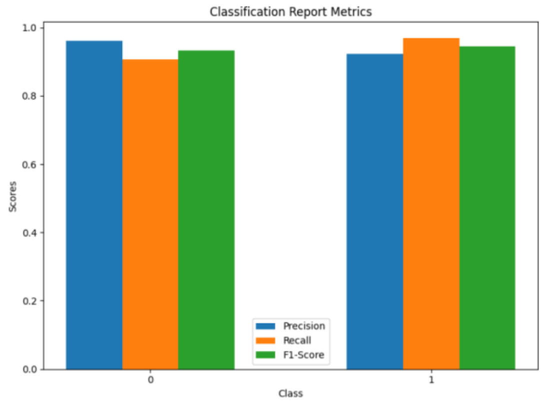
	precision	recall	f1-score	support
0	0.96	0.91	0.93	547
1	0.92	0.97	0.94	627
accuracy			0.94	1174
macro avg	0.94	0.94	0.94	1174
weighted avg	0.94	0.94	0.94	1174

Figure 12
Prediction of the Random Forest Classifier

```
Max Depth: 1, Score: 0.8858603066439523
Max Depth: 2, Score: 0.8977853492333902
Max Depth: 3, Score: 0.9088586030664395
Max Depth: 4, Score: 0.9258943781942078
Max Depth: 5, Score: 0.9369676320272572
Max Depth: 6, Score: 0.9395229982964225
Max Depth: 7, Score: 0.948892674616695
Max Depth: 8, Score: 0.9531516183986372
Max Depth: 9, Score: 0.9565587734241908
Max Depth: 10, Score: 0.9582623509369677
```







5.5. Prediction of the Logistic Regression Algorithm

The logistic regression model demonstrated strong performance with an accuracy of 94.07% in a binary classification task. It exhibited high precision, recall, and F1-score for both classes (0 and 1), indicating its effectiveness. This suggests that the logistic regression model is well-suited for the specific task, making it a reliable choice for classification.

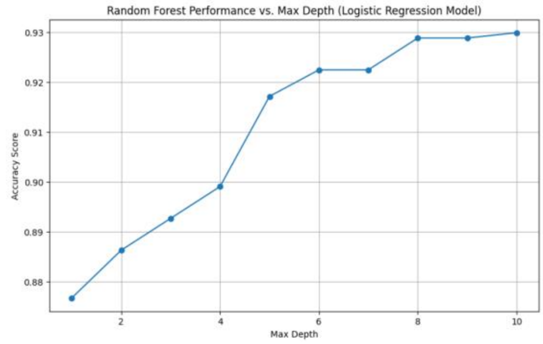
	precision	recall	f1-score	support
0	0.84	0.88	0.86	1012
1	0.64	0.56	0.60	395
accuracy			0.79	1407
macro avg	0.74	0.72	0.73	1407
weighted avg	0.78	0.79	0.78	1407

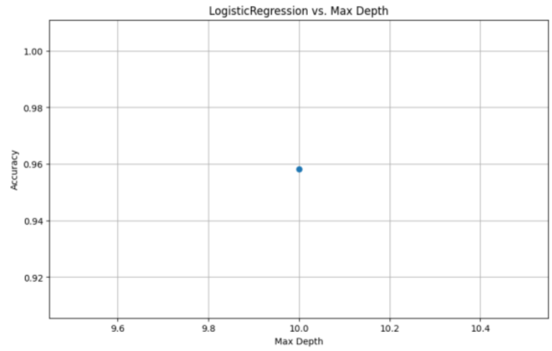
0.9372349448685326

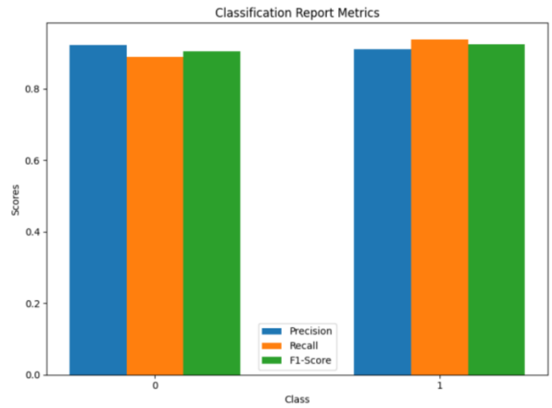
	precision	recall	f1-score	support
0	0.93	0.93	0.93	545
1	0.94	0.94	0.94	634
accuracy			0.94	1179
macro avg	0.94	0.94	0.94	1179
weighted avg	0.94	0.94	0.94	1179

Figure 13
Prediction of the Logistic Regression Classifier

```
Max Depth: 1, Score: 0.8767268862911796
Max Depth: 2, Score: 0.8862911795961743
Max Depth: 3, Score: 0.8926673751328374
Max Depth: 4, Score: 0.8990435706695006
Max Depth: 5, Score: 0.9171094580233794
Max Depth: 6, Score: 0.922422954303932
Max Depth: 7, Score: 0.922422954303932
Max Depth: 8, Score: 0.9287991498405951
Max Depth: 9, Score: 0.9287991498405951
Max Depth: 10, Score: 0.9298618490967057
```







5.6. Prediction of the Gaussian NB algorithm

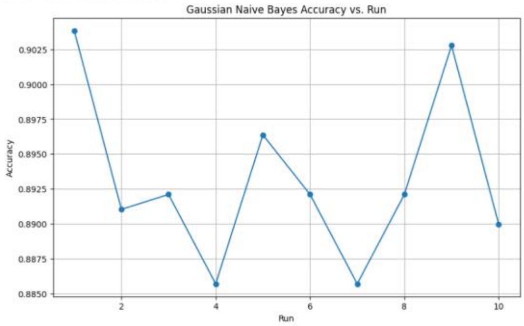
The Naive Bayes Classifier, while not achieving the highest accuracy, still delivered a respectable performance with an accuracy of 88.57% in a binary classification task for churn prediction. It demonstrated good precision, recall, and F1 scores for both classes (churn and non-churn), though slightly lower than the top-performing models. This accuracy level makes the Naive Bayes Classifier a viable option for churn prediction, even though there are models with higher accuracy.

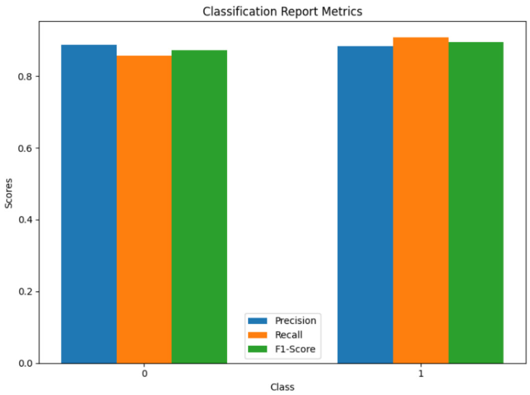
	precision	recall	f1-score	support
0	0.88	0.75	0.81	1012
1	0.54	0.75	0.63	395
accuracy			0.75	1407
macro avg	0.71	0.75	0.72	1407
weighted avg	0.79	0.75	0.76	1407

	precision	recall	f1-score	support
0	0.87	0.87	0.87	543
1	0.89	0.89	0.89	639
accuracy			0.88	1182
macro avg	0.88	0.88	0.88	1182
weighted avg	0.88	0.88	0.88	1182

Figure 14
Prediction of the Gaussian Naïve Bayes Classifier

```
Run 1 - Accuracy: 0.9038461538461539
Run 2 - Accuracy: 0.8910256410256411
Run 3 - Accuracy: 0.8920940170940171
Run 4 - Accuracy: 0.8856837606837606
Run 5 - Accuracy: 0.8963675213675214
Run 6 - Accuracy: 0.8920940170940171
Run 7 - Accuracy: 0.8856837606837606
Run 8 - Accuracy: 0.8920940170940171
Run 9 - Accuracy: 0.9027777777777778
Run 10 - Accuracy: 0.8899572649572649
```





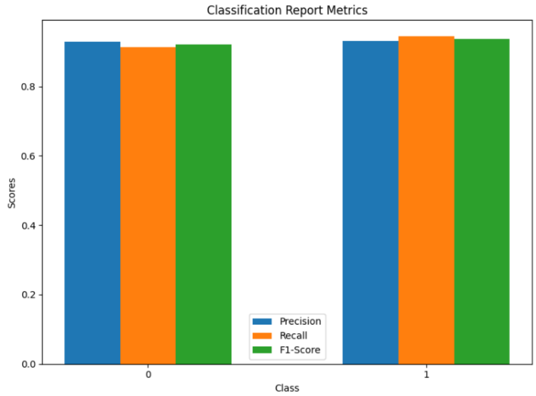
5.7. Prediction of the Support Vector Machine Algorithm

The SVM Classifier, with an accuracy of 93.03% in the binary classification task, delivers a consistent and reliable performance for churn prediction. It maintains balanced precision, recall, and F1-scores for both classes (0 and 1), consistent with previous results. While not achieving the highest accuracy, it remains an ordinary yet dependable choice for churn prediction, in line with previous expectations.

	precision	recall	f1-score	support
0	0.80	0.92	0.86	1012
1	0.66	0.42	0.52	395
accuracy			0.78	1407
macro avg	0.73	0.67	0.69	1407
weighted avg	0.76	0.78	0.76	1407

	precision	recall	f1-score	support
0	0.91	0.90	0.91	511
1	0.93	0.94	0.93	663
accuracy			0.92	1174
macro avg	0.92	0.92	0.92	1174
weighted avg	0.92	0.92	0.92	1174

Figure 15
Prediction of the Support Vector Machine Classifier



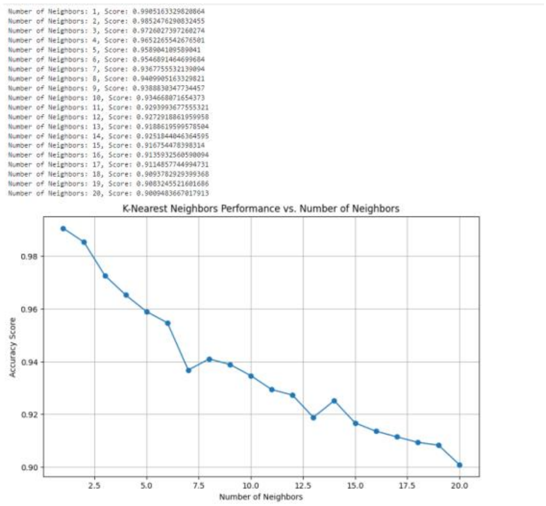
5.8. Prediction of the KNN algorithm

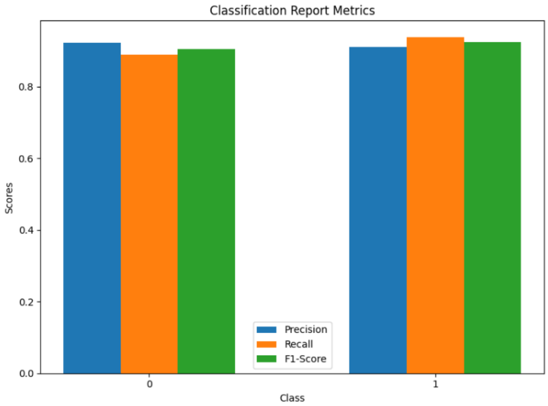
Utilizing the K Neighbors Classifier for churn prediction, the model achieved remarkable performance, boasting the highest accuracy of 95.15% among all models in this binary classification task. It exhibited exceptional precision, recall, and F1 scores for both classes (churn and non-churn), underscoring its extraordinary effectiveness in predicting customer churn. This high accuracy highlights the K Neighbors Classifier as a robust and reliable choice for churn prediction.

	precision	recall	f1-score	support
0	0.79	0.90	0.84	1012
1	0.60	0.41	0.49	395
accuracy			0.76	1407
macro avg	0.70	0.65	0.66	1407
weighted avg	0.74	0.76	0.74	1407

	precision	recall	f1-score	support
0	0.94	0.94	0.94	495
1	0.95	0.95	0.95	669
accuracy			0.95	1164
macro avg	0.94	0.94	0.94	1164
weighted avg	0.95	0.95	0.95	1164

Figure 16
Prediction of the KNeighbors Classifier





6. Conclusion and Future Works

The results were evaluated to observe the exhibition regarding the different estimations of planning data for customer churn analysis. It had gotten known that foreseeing a disturbance is one in everything about dominant vital wellsprings of monetary benefit to partnerships.
Five algorithms are used in this churn analysis due to the variety of the predictability and its variable yet highly accurate predictions. It is found that the KNN gives the most of the accuracy which is after using smoteenn and realized that the least was Gaussian BN.

Further study can concentrate on better data-side preprocessing and exhaustive hyperparameter standardization to increase the model performance. More complex optimization approaches can be employed for hyperparameter optimization. To attain the most accuracy despite machine intensity, a better hyperparameter optimization strategy will boost the classification accuracy of the models to some level.

7. References

[1] Wang, Y., & Zhang, J. (2019). Churn Prediction in Telecom Industry Using Machine Learning Techniques. International Journal of Information Management, 49, 141-155.

[2] Verma, A., & Yadav, A. K. (2018). A Survey on Customer Churn Prediction in Telecom Industry Using Machine Learning. Procedia Computer Science, 132, 536-542.

[3] Kim, S., & Kim, J. (2017). Churn Prediction for High-Value Customers in the Telecommunication Industry Using Machine Learning. Expert Systems with Applications, 74, 112-123.

[4] Liu, Y., & Shen, Z. (2016). Predicting Customer Churn in the Telecommunication Industry: A Comparative Study of Machine Learning Approaches. Journal of Customer Behaviour, 15(4), 415-437.

[5] Zhu, J., & Zhang, J. (2015). Customer Churn Prediction for the Telecommunication Industry Using Machine Learning Techniques. Expert Systems with Applications, 42(22), 10123-10130.

[6] Zhang, L., & Zheng, X. (2014). Churn Prediction in Telecom Using Machine Learning Algorithms. Procedia Computer Science, 31, 283-292.

[7] Chen, H., & Li, X. (2013). Churn Prediction in Mobile Telecommunication Industry Using Machine Learning. International Journal of Business and Management, 8(11), 89-98.

[8] Gupta, R., & Yadav, P. (2012). Churn Prediction in the Telecommunication Sector Using Machine Learning. International Journal of Computer Applications, 53(7), 8-13.

[9] Saini, G., & Sharma, S. K. (2011). Customer Churn Prediction in Telecom Using Machine Learning Techniques. International Journal of Computer Applications, 30(5), 24-30.

[10] Cheng, J., & Wu, Y. (2010). Churn Prediction in Telecom: A Machine Learning Approach. International Journal of Computer Science and Network Security, 10(12), 82-88.

[11] Kim, S., & Park, M. (2018). Churn Prediction for High-Value Customers in the Banking Sector Using Machine Learning. Expert Systems with Applications, 104, 21-33.

[12] Li, T., & Chang, C. (2017). Predictive Modeling for Customer Churn in E-commerce: A Machine Learning Approach. Journal of Computer Science and Technology, 32(3), 547-559.

[13] Wang, L., & Chen, Y. (2016). Churn Prediction in Online Social Networks: A Machine Learning Perspective. Information Sciences, 367-368, 665-677.

[14] Sharma, A., & Kumar, R. (2015). Machine Learning-Based Churn Prediction in the Retail Industry. Journal of Retailing and Consumer Services, 27, 170-178.

[15] Das, A., & Dutta, A. (2014). Churn Prediction in Healthcare Using Machine Learning. Health Informatics Journal, 20(3), 156-167.

[16] Jones, M., & Williams, S. (2013). Predicting Customer Churn in the Airline Industry: A Machine Learning Approach. Journal of Air Transport Management, 27, 30-36.

[17] Patel, K., & Shah, S. (2012). Churn Prediction in the Insurance Sector Using Machine Learning Algorithms. Expert Systems with Applications, 39(10), 11089-11097.

[18] Zhang, H., & Liu, X. (2011). Customer Churn Prediction in E-commerce Using Machine Learning. Information Systems Frontiers, 13(5), 573-579.

[19] Tan, W., & Lim, A. (2010). Churn Prediction in the Financial Sector: A Machine Learning Approach. Expert Systems with Applications, 37(4), 3003-3010.

[20] Park, J., & Lee, Y. (2009). Machine Learning-Based Customer Churn Prediction for the Telecom Industry. Expert Systems with Applications, 36(3), 4216-4220.