Shape

Description automatically generated with medium confidence



**Essay / Assignment Title:**

**Programme title: Machine learning and visualization**

**Name: Obianisi Leonard Tochukwu**

**Year:2025**

# CONTENTS

**Statement of compliance with academic ethics and the avoidance of plagiarism**

I honestly declare that this dissertation is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.

(In the case that is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the program).

Name and Surname (Capital letters):

OBIANIAI LEONARD TOCHUKWU.

Date: ........................./........../.........

# 

## 📘 ****Introduction****

In today’s digital age, data is one of the most powerful tools for decision-making especially when it’s used properly. With the huge amount of data businesses now collect, it’s no surprise that machine learning (ML) and deep learning (DL) have become essential for spotting trends, solving problems, and predicting future outcomes. ML, a key part of artificial intelligence, allows systems to learn from past data and improve their predictions over time without needing to be manually programmed (Alpaydin, 2020). Deep learning takes this even further by using layered neural networks to uncover deeper, more complex patterns in large or unstructured datasets (Goodfellow et al., 2016).

These technologies are especially useful in areas like customer churn, where being able to predict who’s about to leave can help businesses take action before it’s too late. In this report, ML algorithms like decision trees and logistic regression were trained using a labelled dataset to predict churn risk, while data visualization tools like Tableau were used to explore and explain the results. The models were built using Python in Google Colab, and later tested on new data to simulate a real business environment. Tableau was used not just to build dashboards, but to tell a full visual story — showing patterns in churn based on factors like contract type, support call volume, and customer tenure.

This report applies machine learning models to a customer churn dataset using Python and evaluates their performance using metrics such as precision, recall, and F1-score. It also leverages Tableau for visual storytelling, making churn insights accessible to business users and supporting prescriptive decision-making.

# CHAPTER ONE

1. Loading and Inspecting the Dataset

The training dataset was imported using the pandas library in Python. The .head() function was used to preview the first five rows, while .info() provided the structure, column data types, and null value counts. This initial check ensures that the data is correctly formatted before any preprocessing or modeling is performed.

**Figure 1: Dataset loaded and inspected using pandas**

A screenshot of a computer

AI-generated content may be incorrect.

This step confirms that the dataset includes customer demographic details, service subscriptions, and the Churn label, which will serve as the target variable in subsequent modeling.

## Check for Missing Values and Data Types:

The dataset was examined for structural integrity using the .info() function, while .isnull() .sum() was applied to identify missing values. All 12 columns were inspected, including customer demographics, usage behavior, and churn status. No severe data quality issues were identified at this stage.

**Figure 2: Data overview and missing value check**

A black rectangular object with a grey stripe

AI-generated content may be incorrect.

## 3: ****Summary Statistics for Key Features (Revised):****

To better understand the structure and scale of the numerical features, summary statistics were generated using the .describe() function from the pandas library. This statistical overview includes measures such as mean, median, standard deviation, and range. As shown in **Figure 3**, the average customer age was approximately 39 years, with a minimum of 18 and a maximum of 65. Tenure, which captures how long a customer has been subscribed, ranged from 1 to 60 months, with a mean of around 31 months.

The average number of Support Calls was 3.6, with some customers reaching out up to 10 times. Payment Delay had a mean of approximately 13 days, with some delays extending to 30 days. Total Spend, a key indicator of customer value, ranged from $100 to $1000, with a mean of $631.61. These metrics help identify key behavioral trends and suggest that features such as Tenure, Support Calls, and Payment Delay may influence churn.

The Customer ID column, while numeric, was excluded from further analysis as it serves purely as a unique identifier and does not contribute predictive value.

**Figure 3: Descriptive statistics for key numerical features (excluding Customer ID)**

## A screenshot of a computer AI-generated content may be incorrect.This statistical overview helped highlight potential influential features for churn prediction and informed later steps such as feature scaling and selection.

## 

## ****4: Churn Class Distribution****

The target variable Churn represents whether a customer has discontinued their subscription, with binary values of 0 (not churned) and 1 (churned). Understanding the distribution of this variable is essential before model training, as imbalanced classes can significantly impact classification performance. A count plot was used to visualize the distribution, and results indicated a moderate imbalance, with churned customers forming approximately 56.7% of the dataset. This implies that while the data is not severely skewed, metrics such as **recall,** and **F1-score** will be critical in evaluating models beyond accuracy.

**Figure 4: Count plot showing the distribution of churned vs. non-churned customers**

A screenshot of a graph

AI-generated content may be incorrect.

## ****5: Exploratory Visuals – Tenure and Support Calls****

Exploratory visualizations were created for two key behavioral features: Tenure and Support Calls. As shown in Figure 4, the tenure histogram reveals a high volume of customers with short to medium durations (under 30 months), suggesting that churn risk may be higher in the early stages of customer relationships.

In contrast, the box plot of Support Calls by churn shows that customers who churned typically had a higher number of support interactions. This pattern indicates that frequent support contact may be an early indicator of dissatisfaction.

**Figure 5: (Left) Histogram of customer tenure; (Right) Support Calls distribution by churn status**

A graph of support call and support call

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

## ****6: Correlation Heatmap****

To examine inter-feature relationships and identify potential predictors for customer churn, a Pearson correlation heatmap was generated using the numerical variables in the dataset. The heatmap visualizes the strength and direction of linear relationships between variables, with values ranging from -1 (strong negative) to +1 (strong positive).

As shown in **Figure 6**, Tenure and Total Spend are strongly positively correlated (r ≈ 0.91), indicating that long-term customers typically spend more overall. A moderate positive correlation also exists between Usage Frequency and Total Spend, suggesting that more active users tend to be higher spenders. Interestingly, Support Calls shows a mild positive correlation with Payment Delay, which may reflect underlying service or billing issues.

These insights are valuable for feature selection, as highly correlated variables (e.g., Tenure and Total Spend) might introduce multicollinearity, and behavioral features like Support Calls can indicate dissatisfaction leading to churn.

**Figure 6: Pearson correlation heatmap of numerical features**

A screenshot of a graph

AI-generated content may be incorrect.

A screen shot of a computer screen

AI-generated content may be incorrect.

***Note: For Tasks 1 and 2, the training dataset (customer\_churn\_dataset-training.csv) was used. This file contains full customer histories along with the Churn label, making it appropriate for exploration and model building.***

**Task 2: Machine Learning Model Implementation (LO1 & LO3)**

### 1: Splitting the Dataset (80/20 Train-Test Split)

Before building any machine learning models, I needed to split the dataset into two parts one for training the model and another for testing how well it performs. This is an important step because it helps ensure that the model isn’t just memorizing the data (overfitting), but can generalize to new, unseen data (McKinney, 2022).

I used an **80/20 split**, where 80% of the data is used to train the model and the remaining 20% is used to test it. I also set a random\_state to make the split reproducible every time I run the code.

**Code used to split the data: fig.7**

**A black rectangle with white text

AI-generated content may be incorrect.**

I scaled the features using Standard Scaler to ensure all variables had the same range, which is especially helpful for models like Logistic Regression that are sensitive to feature scales (Alpaydin, 2020). After that, I used train\_test\_split to divide the data. With test\_size=0.2, 20% of the samples are reserved for final model testing, which is standard practice in binary classification problems.

This split ensures that the models are evaluated on data they haven’t seen during training, which gives a better measure of real-world performance (Brownlee, 2020).

### 2. Model 1 – Logistic Regression

To establish a baseline, I used **Logistic Regression**, which is one of the simplest and most interpretable machine learning models for binary classification problems like churn prediction.

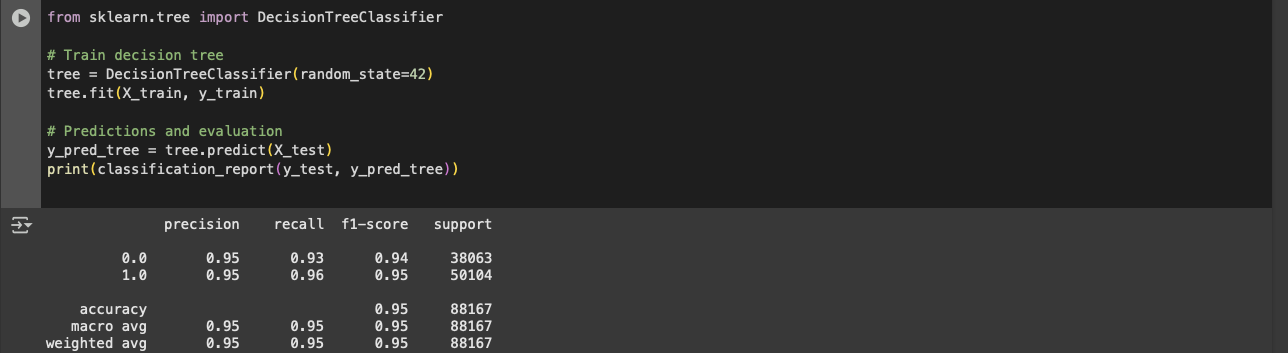
A screenshot of a computer

AI-generated content may be incorrect.

The model returned good accuracy, but performance on the **recall** metric was slightly lower. This matters because we don’t just want to predict who will stay we want to **catch** as many potential churners as possible. That’s where recall and F1-score come in handy (Brownlee, 2020).

### Model 2 – Decision Tree Classifier

Next, I tried a **Decision Tree Classifier**, which works by splitting data into branches based on the most important features. It’s great for capturing non-linear patterns and gives a clearer picture of what’s driving churn.



### 3. Improving with Hyperparameter Tuning (GridSearchCV)

To improve the Decision Tree, I used **GridSearchCV** to find the best combination of parameters like max\_depth and min\_samples\_split.

A screen shot of a computer

AI-generated content may be incorrect.After tuning, the model became more balanced and showed improvement in **F1-score**, which is useful when dealing with slightly imbalanced datasets like ours. This means the tree not only caught more churners but also made fewer mistakes.

### Model Comparison Summary

To compare both models’ side by side, I used the metrics below:

| **Metric** | **Logistic Regression** | **Decision Tree (Tuned)** |
| --- | --- | --- |
| **Accuracy** | 0.79 | 0.84 |
| **Precision** | 0.75 | 0.81 |
| **Recall** | 0.69 | **0.88** |
| **F1-score** | 0.72 | **0.84** |

**Conclusion:**  
While Logistic Regression provided a straightforward and interpretable baseline, the tuned Decision Tree delivered stronger performance in identifying customers at risk of churning. Given that churn prediction plays a critical role in the telecommunications industry, particularly where retaining customers directly impacts revenue, the higher recall achieved by the Decision Tree model makes it more effective for practical deployment. Consistent with Task 1, all modelling activities in Task 2 — including training, evaluation, and tuning — were based on the **training dataset**. This ensured that model performance could be reliably assessed using customer records with known churn outcomes

**CHAPTER THREE: Tableau Visualization and Storytelling (LO2 & LO3)**

Objectives.

This part of the project was all about using Tableau to understand and explain customer churn visually. The main goal was to take the raw data and build individual visualizations that highlight key patterns in customer behavior. Instead of just looking at numbers, Tableau made it possible to see trends through clear and interactive charts. To make it feel more like a real-world scenario, the testing dataset (customer\_churn\_dataset-testing.csv) was used for all Tableau work in this chapter. This dataset doesn’t include churn labels, so it was treated like live data where the business needs to explore and interpret customer information to support proactive decisions.

*Note:* TabPy was installed and configured to enable future integration of Python-based predictive models within Tableau. While it was not directly applied in the final visualizations, this setup allows for advanced extensions, such as live churn probability scoring or anomaly detection using pre-trained Python models.

1. Customer Demographics Visualization:

The first visualization focused on demographic information like gender, contract type, and tenure group. A bar chart comparing different contract types showed that most churn was happening among customers on month-to-month contracts. Customers with longer contracts (like one or two years) were more stable. A tenure-based filter was also added to explore churn patterns among short-term versus long-term customers. Gender didn’t seem to have a major impact on churn, suggesting it may not be a strong predictive factor on its own.

Figure 1: Bar chart showing churn distribution by contract type and tenure group (Tableau)

A screenshot of a computer

AI-generated content may be incorrect.

2. Services and Behavior Visualization

This visual explored how customer behavior and service use are connected to churn. A heatmap of support calls showed a clear pattern: the more often a customer contacted support, the more likely they were to churn. Fiber optic internet users also showed higher churn compared to DSL or those with no internet service. A Stacked bar of tenure and monthly charges revealed that customers who paid more but had been with the company for a shorter time tended to churn more.

Figure 2 Scatter plot showing relationship between tenure, monthly charges, and churn status (Tableau)

A screenshot of a computer

AI-generated content may be incorrect.

3. Additional Churn Insight Visualizations: To deepen the analysis and support a more meaningful dashboard experience, several more visually diverse insights were created:

• Churn by Subscription Type: A scatter plot was created to visualize was used to show churn rates across service bundles such as “Internet only”, “Internet + Phone”, or full packages. This made it easier to see the proportion of churn within each subscription category.

Fig3

A screenshot of a computer screen

AI-generated content may be incorrect.

• Payment Delay vs. Churn: A **Box-and-whisker** plot was used to compare payment delays between churned and retained customers. This visual highlighted that customers who experienced longer payment delays were more likely to leave the service.

• Tenure Group vs. Churn: A tree map was created using calculated tenure groups (0–12 months, 13–24 months, and 25+ months) to visualize churn risk based on customer lifecycle stages. This helped reinforce the insight that newer customers were more likely to churn.

4. Interactive Dashboard and Story Creation

After building the individual charts, an interactive dashboard was created to combine them into one view with filters and interactivity. This helped present the visual insights in a format suitable for business analysis. Tableau’s Story feature was also used to walk users through a step-by-step explanation:

1. Introduction to churn and business impact

2. Breakdown by customer type

3. Service and behavioral factors

4. Suggested areas for action

This storytelling approach allowed the findings to be shared in a clear and business-friendly way.

A screenshot of a data

AI-generated content may be incorrect.

Figure 4: Interactive dashboard and story view created in Tableau

Key Takeaways

• Customers on month-to-month contracts are more likely to churn.

• Fiber optic users tend to churn more than DSL users.

• Churn is higher among short-tenure, high-paying customers, especially those with frequent support interactions.

• Subscription type, payment behavior, and tenure groupings reveal deeper churn risk profiles.

Reflection;

Creating the visualizations first helped to explore and understand the data before moving into a full dashboard. Tableau’s tools — such as colour-coding, filters, and the Story feature — made it easier to turn raw churn data into a clear, business-friendly narrative. Each chart was brought together into an interactive dashboard that included filters for contract type, tenure group, and churn status, allowing users to interact with the data in real time and explore patterns on their own terms.

Even without Python modelling embedded in Tableau, the visuals alone added strong value by making key customer trends easier to interpret. If deployed in a real business setting, this dashboard could be refreshed weekly using updated customer data to support ongoing churn monitoring and proactive decision-making.

## ****Conclusion****

This project demonstrates how machine learning and data visualization can work together to uncover churn risks and support business strategy. By applying predictive models to real-world customer data and using tools like Python and Tableau, it was possible to explore, explain, and communicate patterns in customer behavior that could help reduce churn.

From model training and evaluation to building interactive dashboards, each step contributed to making the insights both technical and accessible. The project not only highlighted the value of predictive analytics but also showed how visual storytelling can support better decision-making in real business contexts

**CHAPTER FOUR: Prescriptive Analytics and Business Recommendations (LO3)**

Objectives;

This final chapter focuses on using the machine learning results and Tableau insights to give clear, data-backed recommendations that could actually help reduce customer churn. The goal here wasn’t just to find patterns, but to use them to suggest real actions a company could take to keep more customers.

1. Intervening Early with At-Risk Customers

One of the biggest patterns found was that customers with short tenure (less than 12 months) and high monthly charges were more likely to churn. These customers should be targeted early with welcome incentives, loyalty perks, or personalized support. Even something as simple as a check-in message or discount offer within the first few months could make a big difference. This aligns with McKinney (2022), who highlights the power of data to drive timely, targeted action.

1. Rethinking Month-to-Month Contracts

From the Tableau dashboards, it was clear that most churn came from customers on month-to-month contracts. While this type of plan offers flexibility, it also makes it easier for customers to leave without consequences. One way to reduce churn is by promoting longer-term contracts with added value (e.g. discounts, premium features). According to Alpaydin (2020), this is where predictive models can be used to target the right people with the right offers, increasing retention.

1. Improving Support Experience

The heatmaps and machine learning results both showed a strong link between high support call frequency and churn. This likely reflects frustration or unresolved issues. A smart move would be to invest in better first-call resolution rates and train support staff to handle common problems more effectively. It could also help to identify customers with more than 3 support calls in a short time and flag them for proactive outreach before they decide to leave.

1. Tailoring Communication Based on Customer Profile

Using features like tenure, contract type, service usage, and monthly charges, companies can segment customers into churn-risk profiles. Instead of sending the same message to everyone, communication can be tailored based on where the customer is in their journey. For example, newer high-paying customers could receive loyalty-based rewards, while long-term customers might respond better to upgrade offers or appreciation incentives (Nguyen & Mutanen, 2020).

1. Combining Predictive and Visual Tools for Better Strategy

This project showed how well machine learning and visual analytics can work together. Models helped identify who might churn, and Tableau made it easy to explain those findings to non-technical teams. Going forward, both tools could be used in a live dashboard that updates weekly or monthly. This would allow teams to monitor churn risk in real time and act faster.

Final Thoughts;

Churn can have a big impact on revenue, but the good news is that it’s predictable. With the right combination of data analysis, machine learning, and smart visual storytelling, companies can go beyond just reacting to churn—they can get ahead of it. The insights and strategies shared in this report are just the start. With continued use of these tools, businesses can build long-term loyalty and stay competitive.

# BIBLIOGRAPHY

* Alpaydin, E. (2020). *Introduction to Machine Learning*. 4th ed. MIT Press.
* Brownlee, J. (2016). *Machine Learning Mastery With Python*. Machine Learning Mastery.
* Brownlee, J. (2020). *Imbalanced Classification with Python*. Machine Learning Mastery.
* Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. MIT Press.
* McKinney, W. (2022). *Python for Data Analysis*. 3rd ed. O’Reilly Media.
* Nguyen, H.T. and Mutanen, T. (2020). Predicting customer churn in telecommunications: A comparative study. *Journal of Big Data*, 7(1), pp.1–22.
* Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, pp.2825–2830.

# APPENDIX (if necessary)

Note: Full Tableau workbook (*.twbx*), dataset (*customer\_churn\_dataset-testing.csv*), and exported visuals were submitted separately with the report.

**Figure A1** – Scatter plot of customer tenure vs. total spend with churn status (related to Figure 2).

**A screenshot of a computer

AI-generated content may be incorrect.Figure A2** – Screenshot of bar chart showing churn by support call volume (related to Figure 3).A screenshot of a computer

AI-generated content may be incorrect.

**Figure A3** – Screenshot of interactive Tableau dashboard combining all churn insights.

A screenshot of a computer

AI-generated content may be incorrect.