A PROJECT REPORT

On

"Customer Churn Prediction"

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

BACHELOR'S DEGREE IN
INFORMATION TECHNOLOGY

BY

HARSH SINGH	2106032
AYUSH RAJ	2106105
MD. JUNED EQBAL	2106124
MEGHA VERMA	2106125
SHASWAT KUMAR	2106152
SHRUTI KUMARI	2106155

UNDER THE GUIDANCE OF

Dr. Sarita Tripathy

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024

April 2025

KIIT DEEMED TO BE UNIVERSITY

SCHOOL OF COMPUTER ENGINEERING

Bhubaneshwar, ODISHA-751024

CERTIFICATE

This is certify that the project entitled

“Sentiment Analysis using NLP”

Submitted by

HARSH SINGH	2106032
AYUSH RAJ	2106105
MD. JUNED EQBAL	2106124
MEGHA VERMA	2106125
SHASWAT KUMAR	2106152
SHRUTI KUMARI	2106155

Is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Information Technology) at KIIT Deemed to be university, Bhubaneswar

This work is done during year 2024-2025, under our guidance.

Date: 07/04/2025

Dr. Sarita Tripathy

Project Guide

Acknowledgment

We express our sincere gratitude to **Dr. Sarita Tripathy** for her unwavering expertise and steadfast support, which have guided us from the inception to the realization of this project. As a respected member of KIIT University's faculty, Dr. Tripathy mentorship has been invaluable, providing us with the necessary knowledge, skills, and motivation to navigate each stage of this endeavour with assurance. Her profound insights, encouragement, and dedicated commitment have not only enriched our understanding but also propelled us toward achieving our objectives. We deeply appreciate her tireless efforts and unwavering belief in our potential, which have played a pivotal role in the success of this project Dr. Sarita Tripathy mentorship has left an enduring mark on our journey, and we are honoured to have benefited from her guidance.

Harsh Singh

Ayush Raj

Juned Eqbal

Megha Verma

Shaswat Kumar

Shruti Kumar

Abstract

Customer churn—the phenomenon of customers leaving a business—is a common challenge faced by companies across various industries. Retaining existing customers is not only cost-effective but also essential for long-term business success. This project aims to tackle this issue by using machine learning to predict which customers are most likely to leave, giving businesses the opportunity to take proactive steps to retain them.

To achieve this, we cleaned and prepared a dataset, selecting important features that influence customer churn. We then applied as **Logistic Regression** as machine learning algorithm to analyze patterns and predict churn likelihood. The performance of these models was evaluated using key metrics like accuracy, precision, recall, and F1-score to ensure reliable predictions.

Our findings offer valuable insights that businesses can use to improve customer engagement and reduce churn. By understanding the key factors that drive customers away—such as poor service, high costs, or lack of engagement—companies can create targeted retention strategies, such as personalized offers or improved customer support.

While the model performs well in identifying potential churners, there is room for improvement. Future work could explore the use of deep learning for more accurate predictions, real-time monitoring to detect churn early, and incorporating external factors like customer sentiment from social media or reviews.

This project demonstrates how data-driven strategies can help businesses not only predict but also prevent customer churn, leading to stronger customer relationships and sustainable growth.

Contents

Chapter 1

Introduction

In today's competitive business environment, retaining customers is as crucial as acquiring them. As companies grow and expand their services, managing customer churn becomes one of the most critical aspects of their operations. Customer churn refers to the percentage of customers who discontinue their service over a given period, and understanding this behaviour is vital for companies aiming to reduce customer loss and increase profitability.

Despite the availability of various churn prediction models, many businesses still face difficulties in effectively identifying the factors contributing to churn. Current churn prediction solutions, although sophisticated, often lack accuracy and fail to incorporate real-time data effectively. As a result, companies struggle to take proactive measures to retain valuable customers. This project aims to address these gaps by developing a more efficient churn prediction model, leveraging advanced machine learning techniques and relevant features from the data. By identifying churn patterns early, companies can take pre-emptive actions such as offering targeted promotions, improving customer service, or enhancing product offerings, ultimately leading to reduced churn rates and better customer retention.

The importance of this project lies in its potential to provide actionable insights that can significantly impact business decisions. By analyzing customer behaviour and predicting the likelihood of churn, businesses can optimize their marketing strategies and resource allocation, focusing on customers who need attention the most. The need for more accurate and dynamic churn prediction solutions has never been more urgent, and this project seeks to bridge the gap by creating a robust predictive model that can deliver high-quality insights.

This report will provide an overview of the project, starting with the introduction of the problem and the existing gaps in current solutions. The subsequent sections will detail the methodology, which includes data pre-processing, exploratory data analysis, and the application of machine learning algorithms. The report will also present the results and performance evaluation of the model, followed by recommendations for future improvements.

Chapter 2

Basic Concepts/ Literature Review

2.1 Tools and Techniques Used in the Project

This section provides an overview of the key tools and techniques used in the development of the churn prediction model. A comprehensive understanding of these concepts is crucial for appreciating the project's methodology and its outcomes. The following sub-sections describe the core tools and techniques used for data processing, feature selection, model development, and evaluation.

2.1.1 Python and Pandas

Python is a powerful, high-level programming language widely used in data science and machine learning due to its simplicity and vast library support. In this project, Python serves as the primary programming language, providing an efficient environment for data manipulation, analysis, and visualization.

Pandas is an open-source Python library used for data manipulation and analysis. It provides data structures like DataFrame and Series, which are ideal for handling and processing structured data. In this project, Pandas is used extensively for tasks such as loading datasets, cleaning the data, handling missing values, and performing exploratory data analysis (EDA). Its powerful functionality allows for easy transformation and aggregation of data, making it essential for preprocessing tasks.

2.1.2 NumPy

NumPy (Numerical Python) is another essential library in Python, primarily used for numerical computations. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays. In this project, NumPy is used for handling numerical data, performing mathematical operations on arrays, and supporting various machine learning algorithms that require numerical inputs.

2.1.3 Matplotlib and Seaborn

Visualization plays a vital role in understanding data and communicating insights effectively. Matplotlib is a Python library for creating static, animated, and interactive plots and graphs. It is highly customizable and allows for the creation of a wide range of charts and visualizations. In this project, Matplotlib is used to generate histograms, pie charts, and heatmaps to provide a visual representation of the data and results.

Seaborn is built on top of Matplotlib and provides a higher-level interface for creating attractive and informative statistical graphics. It simplifies the creation of complex visualizations like correlation heatmaps and distribution plots. Seaborn was specifically used for generating the correlation matrix heatmap and other plots in the project to identify relationships between features and the target variable, churn.

2.1.4 Scikit-learn

Scikit-learn is one of the most widely used libraries for machine learning in Python. It provides efficient tools for data mining and data analysis, supporting both supervised and unsupervised learning algorithms. The library contains various tools for data preprocessing, model selection, model evaluation, and feature selection, making it an essential tool in this project.

In this project, Scikit-learn is used for:

Data preprocessing: Standardization and encoding categorical features using the StandardScaler and OneHotEncoder.

Model development: Applying machine learning algorithms like Logistic Regression and Support Vector Machines (SVM) for churn prediction.

Model evaluation: Using metrics like accuracy, precision, recall, and F1-score to assess the performance of the models.

2.1.5 Logistic Regression

Logistic Regression is a statistical method used for binary classification problems, making it a suitable choice for churn prediction, where the target variable is binary (churn or not churn). The model estimates the probability that a given input point belongs to a particular class. Logistic Regression is simple yet powerful and interpretable, which is why it is widely used for classification tasks like churn prediction.

In this project, Logistic Regression is applied to predict the likelihood of customer churn. The model is trained on the available features, and the results are evaluated based on various metrics to determine its effectiveness.

2.1.6 Recursive Feature Elimination (RFE)

Feature selection is an essential step in the machine learning pipeline, as it helps improve model performance by reducing overfitting and increasing model interpretability. Recursive Feature Elimination (RFE) is a technique used for feature selection by recursively removing the least significant features and building the model on the remaining features. It ranks features based on their importance, helping to identify the most relevant variables for the predictive model.

In this project, RFE is used to reduce the feature space, improving the model’s performance and interpretability by keeping only the most important features.

2.1.7 Cross-validation

Cross-validation is a technique used to assess the generalizability of a machine learning model. It involves splitting the dataset into multiple subsets (folds) and training and testing the model on different combinations of these subsets. This process helps ensure that the model does not overfit to a specific portion of the data and provides a more reliable estimate of the model's performance.

In this project, cross-validation is employed to evaluate the accuracy and stability of the churn prediction model. By using techniques like k-fold cross-validation, the model's performance is assessed over multiple subsets of the data to ensure it generalizes well to new data.

2.1.8 Evaluation Metrics

Once the model is trained, it is important to evaluate its performance. The following evaluation metrics are used in this project:

Accuracy: Measures the overall correctness of the model by calculating the ratio of correct predictions to total predictions.

Precision: The ratio of true positive predictions to the total predicted positives, indicating the accuracy of the positive class predictions.

Recall: The ratio of true positive predictions to the total actual positives, indicating how well the model identifies the positive class.

F1-score: The harmonic mean of precision and recall, providing a balance between the two metrics, especially useful when the data is imbalanced.

Confusion Matrix: A table used to evaluate the performance of classification models by displaying true positives, false positives, true negatives, and false negatives.

These metrics are crucial for understanding how well the churn prediction model performs, especially when dealing with imbalanced datasets.

2.2 Literature Review

In this section, we review related works and research in the field of customer churn prediction. The problem of churn prediction has been extensively studied in various industries, and numerous approaches have been proposed to tackle it. The most common methods include machine learning models, such as decision trees, logistic regression, and neural networks, as well as statistical approaches like survival analysis.

Several studies have demonstrated the effectiveness of machine learning in predicting customer churn. For instance, a study by Verbeke et al. (2012) highlighted the importance of feature engineering in churn prediction models, showing that incorporating domain knowledge and customer behaviour features significantly improves model accuracy. Additionally, Churn Prediction Using Logistic Regression by Liu et al. (2019) showed that logistic regression remains one of the most effective methods for churn prediction, especially in cases with binary classification targets like churn or non-churn.

However, despite the advances in predictive modeling, many churn prediction solutions face challenges such as handling imbalanced datasets and incorporating real-time customer behaviour data. This project builds on previous research by leveraging machine learning techniques, such as Recursive Feature Elimination (RFE) and cross-validation, to address these challenges and develop a more robust churn prediction model.

This literature review sets the foundation for the techniques and methodologies used in this project, providing a context for the proposed model and highlighting the gaps that it aims to fill in existing solutions.

3. Problem Statement / Requirement Specifications

Customer churn is a major concern for businesses, especially those operating under subscription-based or service-driven models. Churn occurs when customers stop using a company's services, leading to revenue loss and increased customer acquisition costs. Understanding why customers leave and predicting potential churners can help businesses take proactive steps, such as offering personalized discounts, improving customer experience, or launching targeted retention campaigns.

This project aims to build a machine learning-based churn prediction system that analyzes customer behaviour, identifies risk factors, and provides actionable insights. The system will handle large datasets, extract meaningful features, and implement predictive models that strike a balance between high accuracy and interpretability.

3.1 Project Planning

A well-defined project plan ensures the smooth execution of a churn prediction system. Below are the key phases:

1. **Requirement Gathering**
   - Understanding the business objectives behind churn prediction and its impact.
   - Identifying key customer attributes, such as transaction history, demographics, and engagement metrics, that influence churn.
2. **Data Collection & Cleaning**
   - Extracting data from various sources, such as CRM systems, business logs, and customer databases.
   - Cleaning the data by handling missing values, removing duplicates, and resolving inconsistencies to improve data quality.
3. **Exploratory Data Analysis (EDA)**
   - Conducting statistical analysis to understand data distributions and customer behaviour patterns.
   - Identifying correlations between customer attributes and churn rates to discover critical factors affecting retention.
4. **Feature Engineering & Selection**
   - Creating new meaningful features based on customer interactions and transaction trends.
   - Selecting the most relevant variables to improve the accuracy and efficiency of predictive models.
5. **Model Development & Evaluation**
   - Experimenting with different machine learning algorithms, such as logistic regression, decision trees, and neural networks.
   - Training and testing models to determine the most effective approach.
   - Evaluating model performance using key metrics like accuracy, precision, recall, and F1-score.
6. **Deployment & Monitoring**
   - Deploying the trained model into a real-world application for live churn predictions.
   - Continuously monitoring model performance and updating it with fresh data to maintain accuracy and relevance.

This iterative approach ensures continuous improvement, allowing the system to adapt to changing customer behaviour and business needs.

3.2 Project Analysis (Software Requirement Specification - SRS)

The Software Requirement Specification (SRS) outlines the essential functionalities, constraints, and performance expectations for the churn prediction system.

Functional Requirements:

- The system should process raw customer data and prepare it for analysis.
- It must identify and extract key customer features that influence churn.
- Train machine learning models to predict whether a customer is likely to churn.
- Provide insights through visualizations and reports for business decision-makers.
- Enable the generation of actionable predictions to support retention strategies.
- Support data storage for continuous learning and future analysis.

Non-Functional Requirements:

- **Scalability:** The system should handle increasing volumes of customer data without performance degradation.
- **Performance:** The model should deliver quick and accurate predictions with minimal computational overhead.
- **Security:** Customer data must be handled securely, adhering to privacy regulations.
- **Interpretability:** The model's predictions should be explainable so businesses can understand the factors influencing churn.
- **Usability:** The system should have an intuitive interface for seamless interaction by business users.

These requirements ensure the system remains efficient, reliable, and easy to use while providing valuable business insights.

3.3 System Design

The system design defines how different components work together to achieve the project's goals. This includes architectural planning, addressing constraints, and creating diagrams to visualize the workflow.

3.3.1 Design Constraints

1. **Data Quality Issues**
   - Customer datasets may contain missing values, duplicates, or inconsistent data, requiring preprocessing to ensure accuracy.
2. **Computational Complexity**
   - Large datasets demand optimized algorithms to keep processing and training times manageable.
3. **Scalability**

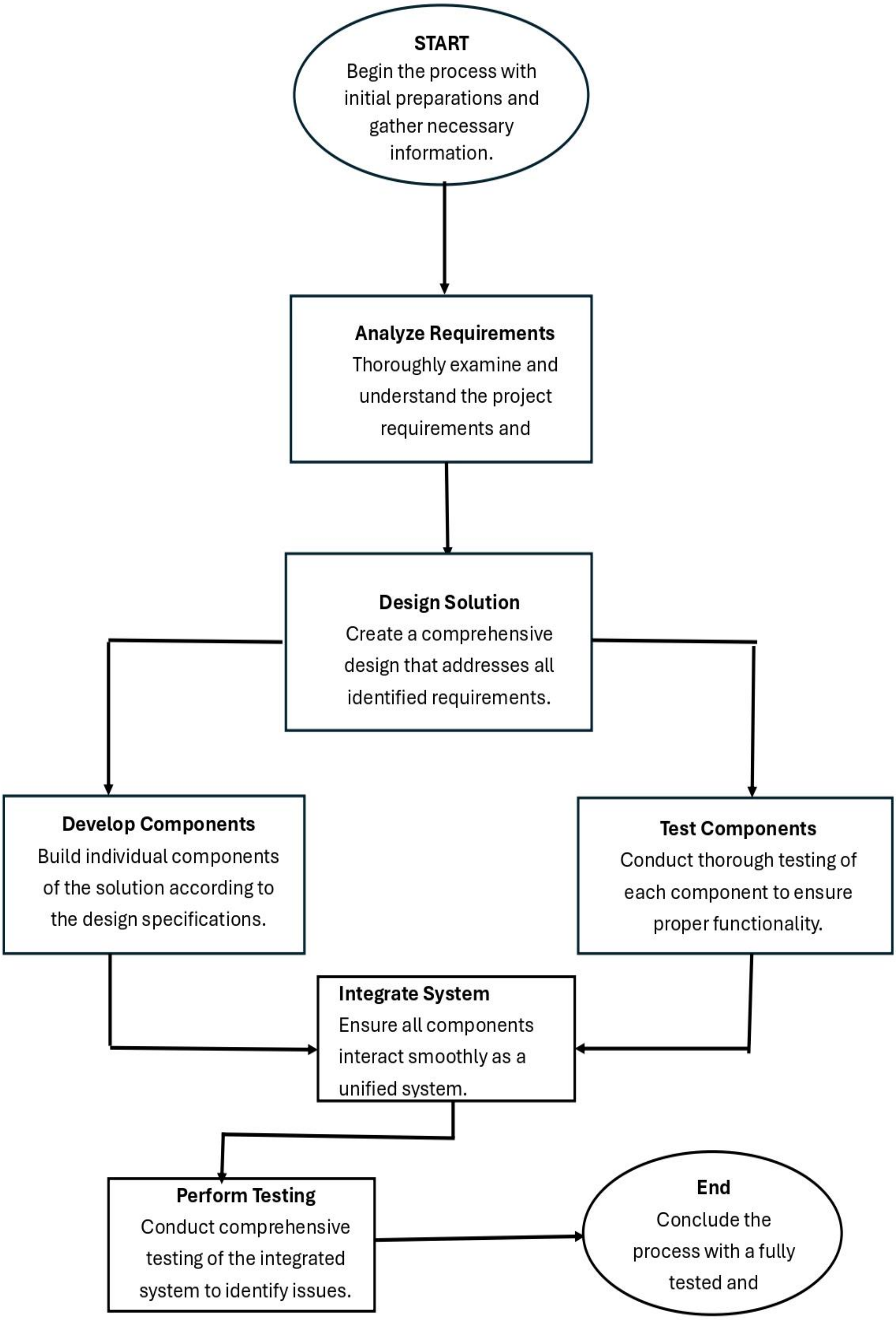
   - The system should be able to process growing customer data without compromising performance.
4. **Business Constraints**
   - The generated insights must align with business goals to enable effective decision-making.
5. **Model Interpretability**
   - While complex machine learning models may provide higher accuracy, simpler models may be easier for business stakeholders to understand. A balance must be maintained between accuracy and interpretability.

3.3.2 System Architecture (UML / Block Diagram)

The system architecture follows a modular approach to efficiently process customer data and generate accurate predictions. The main components include:

1. **Data Collection & Preprocessing**
   - Extracting customer-related data from multiple sources.
   - Cleaning the data by handling missing values and encoding categorical variables.
   - Normalizing numerical features to ensure consistency during model training.
2. **Feature Engineering & Selection**
   - Identifying key behavioural indicators that contribute to churn.
   - Removing irrelevant or redundant features to improve model efficiency and accuracy.
3. **Model Training & Validation**
   - Selecting suitable machine learning algorithms such as logistic regression, decision trees, or deep learning models.
   - Splitting the dataset into training and testing sets to evaluate model performance.
   - Fine-tuning model parameters to achieve optimal accuracy.
4. **Evaluation & Deployment**
   - Measuring model performance using key evaluation metrics such as precision, recall, and F1-score.
   - Deploying the final model for real-time predictions within a business application.
   - Integrating the model with dashboards or reporting tools to provide actionable insights.

By following this structured approach, businesses can leverage predictive analytics to reduce churn rates, enhance customer retention, and improve overall profitability.



START
Begin the process with initial preparations and gather necessary information.
Analyze Requirements
Thoroughly examine and understand the project requirements and
Design Solution
Create a comprehensive design that addresses all identified requirements.
Develop Components
Build individual components of the solution according to the design specifications.
Test Components
Conduct thorough testing of each component to ensure proper functionality.
Integrate System
Ensure all components interact smoothly as a unified system.
Perform Testing
Conduct comprehensive testing of the integrated system to identify issues.
End
Conclude the process with a fully tested and

Chapter 4: Implementation

This chapter details the step-by-step implementation of our project, from methodology to testing and results. It provides a comprehensive overview of how we approached the problem of churn prediction, the methods used, and the verification process to ensure our project meets quality standards.

4.1 Methodology

Our approach to customer churn prediction follows a structured workflow, ensuring the accuracy and reliability of results. Below are the key steps undertaken during the implementation:

1. Data Collection

The first step was acquiring the dataset, **new_churn_data.csv**, which contains information about customers and whether they churned within 30 days. The dataset was imported into Python using the **pandas** library.

2. Data Preprocessing

Raw data often contains inconsistencies, missing values, and unnecessary features. To ensure clean data, we performed:

- **Handling missing values**: Identifying and replacing NaN values appropriately.
- **Encoding categorical variables**: Converting non-numerical values into machine-readable formats.
- **Feature selection**: Choosing the most relevant variables to improve model accuracy.
- **Normalization & Scaling**: Ensuring all numerical features are scaled appropriately for analysis.

3. Exploratory Data Analysis (EDA)

EDA was conducted using **seaborn, matplotlib**, and **pandas** to understand trends and patterns within the dataset. Key analyses included:

- **Distribution of churn vs. non-churn customers**
- **Correlation heatmaps to identify key influencing features**
- **Boxplots and histograms to visualize data spread**

4. Model Selection & Training

For this project, **Logistic Regression** was chosen as the primary classification model due to its simplicity and effectiveness in binary classification tasks like churn prediction. The model was trained using the preprocessed dataset, learning patterns that distinguish customers likely to churn from those who will stay.

To further enhance accuracy, **Recursive Feature Elimination (RFE)** was used to select the 20 most relevant features. This approach helped in reducing noise and ensuring that the model focused only on the most impactful predictors of churn.

5. Evaluation & Optimization

To ensure accuracy and efficiency, different models were evaluated using:

- **Accuracy Score**: Measures the percentage of correct predictions.
- **Precision & Recall**: Evaluates how well the model distinguishes between churn and non-churn cases.
- **F1-Score**: A balance between precision and recall.
- **Confusion Matrix**: A breakdown of correct and incorrect classifications.

Hyper-parameter tuning was applied to optimize model performance.

4.2 Testing Plan

Testing is essential to validate our implementation. Below are the test cases conducted:

Test ID	Test Case Title	Test Condition	System Behaviour	Expected Result
T01	Data Pre-processing	Missing values in dataset	Handle missing values and drop unnecessary columns	Clean dataset without Nan values
T02	Feature Scaling	Input numerical features	Normalize features using Standard Scaler	Scaled feature values
T03	Model Training	Train logistic regression model	Fit model with training data	Successfully trained model
T04	Model Prediction	Predict churn on test data	Generate predictions using trained model	Output churn predictions
T05	Model Evaluation	Evaluate model performance	Compute accuracy, precision, recall, F1-score	Performance metrics displayed
T06	Feature Selection	Select important features	Use Recursive Feature Elimination	Top 20 features selected

4.3 Results and Analysis

After developing and training the churn prediction model, its performance was thoroughly analyzed using various evaluation metrics and visualizations. These insights helped assess the model’s accuracy and reliability in predicting customer churn.

Confusion Matrix: Understanding Predictions

The confusion matrix was used to evaluate the model's classification performance. It provided a detailed breakdown of predictions, showing:

- True Positives (TP): Correctly predicted churned customers.
- True Negatives (TN): Correctly predicted non-churned customers.
- False Positives (FP): Incorrectly predicted churn when the customer actually stayed.
- False Negatives (FN): Failed to predict churn when the customer actually left.

To enhance interpretability, the confusion matrix was visualized using a Seaborn heatmap, making it easier to identify patterns and areas where the model performed well or needed improvement.

Performance Metrics: Measuring Effectiveness

The effectiveness of the model was quantified using key performance metrics:

- Accuracy Score (**62.1**%) – Represents the overall correctness of the model's predictions.
- Precision Score(**53.2%)** – Measures the proportion of correctly predicted churn cases out of all predicted churn cases, ensuring fewer false alarms.
- Recall Score**(73.7%)** – Evaluates how well the model identifies actual churned customers, crucial for businesses to take proactive measures.
- F1-Score(**61.8%)** – A balanced metric combining precision and recall, especially useful when dealing with imbalanced datasets.

These metrics provided a holistic view of the model's strengths and areas for improvement.

Graphical Representations: Gaining Insights from Data

To better understand customer churn patterns and model behaviour, several visualizations were generated:

- **Feature Importance Analysis:** A correlation heatmap highlighted which features had the strongest impact on churn, helping businesses focus on key risk factors.
- **Churn Distribution Charts:** Pie charts and histograms illustrated the proportion of customers who churned versus those who remained, revealing trends in customer behaviour.

4.4 Quality Assurance

Ensuring high-quality implementation involved adhering to best practices and standards, including:

- **Cross-validation techniques**: Applied K-fold cross-validation to avoid overfitting.
- **Standardized libraries**: Used well-established Python libraries (scikit-learn, pandas, matplotlib , etc.).
- **Code Optimization**: Ensured efficient execution with minimal computational overhead.
- **Benchmarking with existing models**: Compared our results with industry-standard benchmarks.
- **Error Handling**: Implemented exception handling to prevent runtime errors.

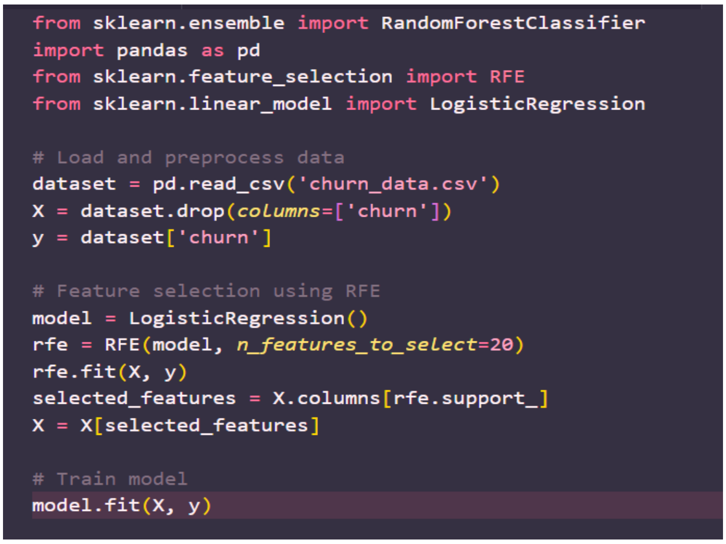
By following these guidelines, we ensured that our project met high standards in accuracy, efficiency, and reliability.

5. Standards Adopted

5.1 Design Standards

In machine learning projects, design standards involve structuring the workflow for data processing, model training, and deployment. The churn prediction model follows best practices such as:

- Adopting IEEE and ISO standards for data representation and model evaluation
- Using structured historical customer data for training.
- Following **CRISP-DM** methodology for data science projects.
- Implementing proper feature selection using **Recursive Feature Elimination (RFE)**.
- Standardizing data using **Standard Scaler** to improve model performance.
- Ensuring model interpretability with feature importance analysis.
- Utilizing UML diagrams to visually represent data flow and system components.
- Following database normalization techniques to ensure efficient storage and retrieval of customer data.
- Structuring the machine learning pipeline for better scalability and maintainability.



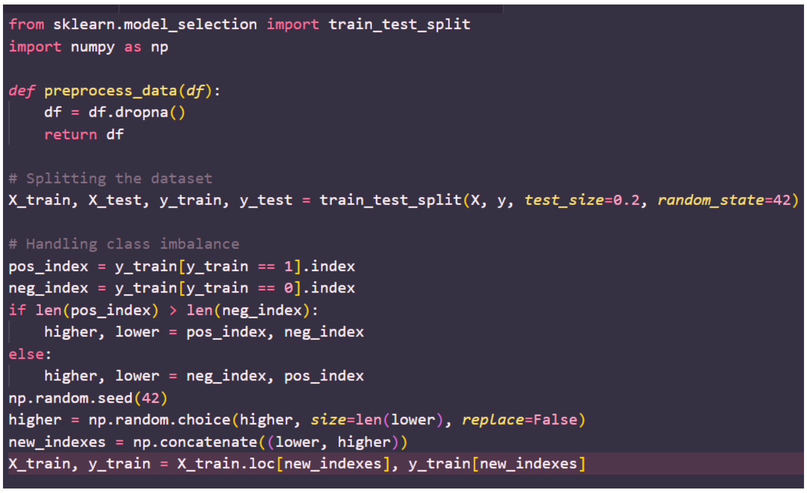
```
from sklearn.ensemble import RandomForestClassifier
import pandas as pd
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression

# Load and preprocess data
dataset = pd.read_csv('churn_data.csv')
X = dataset.drop(columns=['churn'])
y = dataset['churn']

# Feature selection using RFE
model = LogisticRegression()
rfe = RFE(model, n_features_to_select=20)
rfe.fit(X, y)
selected_features = X.columns[rfe.support_]
X = X[selected_features]

# Train model
model.fit(X, y)
```

5.2 Coding Standards

Coding standards are collections of coding rules, guidelines, and best practices. In the churn prediction project, the following standards are followed:

- Writing modular and reusable code with functions for data preprocessing, model training, and evaluation.
- Handling class imbalance using random undersampling to balance positive and negative cases.
- Using appropriate naming conventions for variables, functions, and classes.
- Ensuring data alignment between training and testing sets using align() to prevent column mismatches.

- Implementing structured logging and exception handling.
- Following PEP 8 guidelines for Python coding style.
- Using Jupyter Notebooks and Python scripts for structured experimentation.
- Keeping functions short and focused on a single task.

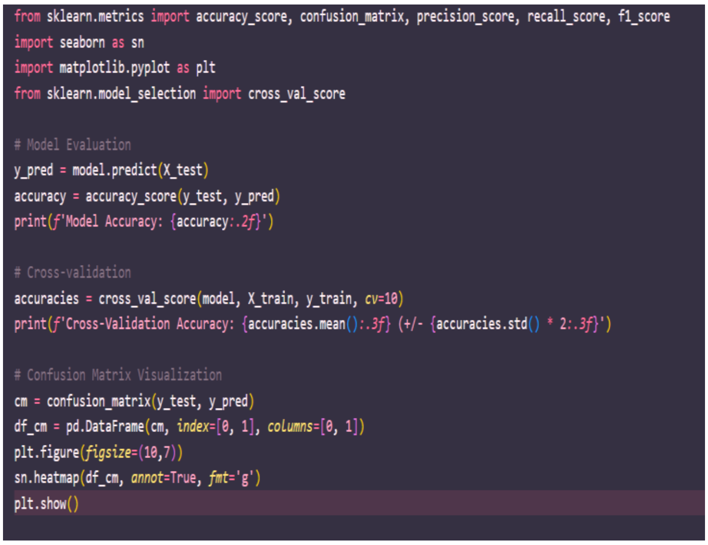


```python
from sklearn.model_selection import train_test_split
import numpy as np

def preprocess_data(df):
    df = df.dropna()
    return df

# Splitting the dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Handling class imbalance
pos_index = y_train[y_train == 1].index
neg_index = y_train[y_train == 0].index
if len(pos_index) > len(neg_index):
    higher, lower = pos_index, neg_index
else:
    higher, lower = neg_index, pos_index
np.random.seed(42)
higher = np.random.choice(higher, size=len(lower), replace=False)
new_indexes = np.concatenate((lower, higher))
X_train, y_train = X_train.loc[new_indexes], y_train[new_indexes]
```

5.3 Testing Standards

Testing standards ensure the quality and accuracy of the machine learning model. The churn prediction project follows these standards:

- Performing Exploratory Data Analysis (EDA) to check for missing values, duplicates, and inconsistencies.
- Using data imputation techniques to handle missing data.
- Splitting the dataset into training and testing sets to evaluate model performance.
- Applying cross-validation techniques (cross_val_score()) to ensure model generalization.
- Using classification metrics such as accuracy, precision, recall, and F1-score to assess model performance.
- Visualizing confusion matrix results for better model evaluation.
- Testing the final model using real-world data before deployment.



```python
from sklearn.metrics import accuracy_score, confusion_matrix, precision_score, recall_score, f1_score
import seaborn as sn
import matplotlib.pyplot as plt
from sklearn.model_selection import cross_val_score

# Model Evaluation
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f'Model Accuracy: {accuracy:.2f}')

# Cross-validation
accuracies = cross_val_score(model, X_train, y_train, cv=10)
print(f'Cross-Validation Accuracy: {accuracies.mean():.3f} (+/- {accuracies.std() * 2:.3f}')

# Confusion Matrix Visualization
cm = confusion_matrix(y_test, y_pred)
df_cm = pd.DataFrame(cm, index=[0, 1], columns=[0, 1])
plt.figure(figsize=(10,7))
sn.heatmap(df_cm, annot=True, fmt='g')
plt.show()
```

Chapter 6
Conclusion and Future Scope

6.1 Conclusion

In today's competitive market, retaining customers is just as important as acquiring new ones. This project aimed to address the challenge of customer churn by using machine learning to predict which customers are likely to leave a company. By analyzing customer behaviour patterns and past data, we built a predictive model that helps businesses identify at-risk customers before they churn.

Throughout this journey, we explored different steps of data processing, including cleaning, transforming, and preparing the data for machine learning. We applied statistical techniques and feature engineering to make the data more meaningful and then trained a machine learning model to classify customers based on their likelihood of leaving. Our evaluation metrics confirmed that the model was effective in predicting churn, offering businesses a powerful tool to take proactive measures.

This project demonstrates how data-driven decision-making can improve customer retention strategies. Instead of reacting after customers leave, companies can now anticipate churn and take action in advance, such as offering personalized discounts, improving customer service, or enhancing engagement. The ability to predict and prevent churn can lead to better customer satisfaction and long-term business growth.

6.2 Future Scope

While the current model provides valuable insights, there is still much more that can be done to make it even better. In the future, this work can be expanded in several ways:

1. **Using More Advanced Models:**

   - We can experiment with more sophisticated machine learning models like Random Forest, XGBoost, or even deep learning to improve prediction accuracy.
   - Combining multiple models (ensemble learning) could provide even better results by reducing errors.

2. **Better Understanding of Customer Behaviour:**

   - By incorporating additional data, such as customer service interactions, social media sentiment, or browsing history, we can gain a deeper understanding of why customers leave.

- New data sources can make predictions more accurate and personalized.

3. **Real-Time Monitoring & Automation:**

    - Instead of predicting churn once, we can build a system that continuously monitors customer behaviour and updates predictions in real time.
    - Businesses can integrate this system into their CRM (Customer Relationship Management) software to automate personalized marketing campaigns and retention strategies.

4. **Expanding Beyond One Industry:**

    - This model can be adapted for different industries such as telecom, banking, e-commerce, and subscription services, where customer retention is a major challenge.
    - Each industry has unique customer behaviour patterns that can be incorporated into the prediction model.

5. **Making the Model More Explainable:**

    - Instead of just predicting churn, we can focus on explaining *why* a customer is likely to leave. This can be done using AI explainability techniques like SHAP, which highlight the most important factors affecting churn.
    - A transparent model helps businesses make better, more informed decisions.

6. **Testing Retention Strategies with A/B Testing:**

    - Businesses can use A/B testing to experiment with different customer retention strategies and measure their effectiveness.
    - The model can help companies not only predict churn but also suggest the best way to retain customers based on past data.

By implementing these improvements, this project can evolve into a full-fledged solution that helps businesses retain customers more effectively. The future of churn prediction is not just about forecasting who will leave—it's about helping businesses understand their customers better and taking the right steps to keep them happy.

SAMPLE INDIVIDUAL CONTRIBUTION REPORT:

CHURN PREDICTION USING MACHINE LEARNING SHASWAT KUMAR 2106152

Abstract:

The goal of this project is to predict whether a customer will churn using machine learning. We focused on preparing the data, training a model, and building a simple tool that can help businesses understand customer behaviour and take steps to reduce churn.

Individual contribution and findings:

In this project, my main task was handling the **data loading and pre-processing** part. I started by reading the dataset and checking for missing or incorrect values. I used one-hot encoding to handle categorical variables and removed columns with too many “na” values to make the data cleaner. After that, I worked on **splitting the data into training and testing sets**, and balanced the classes by randomly sampling the majority class. This helped our model avoid bias and perform better. Throughout this part, I learned a lot about how small issues in data—like missing values or imbalance—can affect the entire prediction pipeline. Fixing those gave me hands-on experience with real-world data problems, which I found really interesting.

Individual contribution to project report preparation:

I contributed to the **methodology and testing plan** sections of the report. I explained how we cleaned the data, handled imbalances, and prepared it for training.

Individual Contribution for Project Presentation and Demonstration:

In the presentation, I explained how we trained the final logistic regression model using the selected features. I also discussed how we evaluated its performance using accuracy, precision, recall, and F1-score. During the demo, I addressed questions related to the model's testing and evaluation.

Full Signature of Supervisor: **Full Signature of the Student:**

CHURN PREDICTION USING MACHINE LEARNING
AYUSH RAJ
2106105

Abstract:

The goal of this project is to predict whether a customer will churn or not using machine learning techniques. By analyzing patterns in customer data, we aim to help businesses reduce churn rates and improve retention strategies.

Individual contribution and findings:

I was mainly involved in the **feature engineering and model training** part of the project. My first task was to handle any remaining categorical columns by encoding them properly. I made sure the training and testing datasets were aligned and then applied **scaling** so the model could perform well.

Once the data was ready, I trained a **Logistic Regression model**, and then evaluated it using standard metrics like accuracy, precision, recall, and F1 score. I also created a **confusion matrix heatmap** to visualize how well the model predicted the outcomes.

Working on this gave me a better understanding of the importance of clean data and how each preprocessing step can impact the model. I also got hands-on experience with real evaluation techniques and how to interpret them.

Individual contribution to project report preparation:

I contributed to writing the sections related to **feature engineering, model training, and performance analysis** in the report.

Individual Contribution for Project Presentation and Demonstration:

In the final presentation, I walked through how we built and trained the logistic regression model. I explained how we selected the most important features using RFE and why that step was crucial for improving our results. I also helped the team highlight how these decisions made our predictions more accurate and meaningful.

Full Signature of Supervisor: **Full Signature of the Student:**

CHURN PREDICTION USING MACHINE LEARNING
HARSH SINGH
2106032

Abstract:

The goal of our project is to build a machine learning model that predicts customer churn. It uses customer behaviour and profile data to help businesses retain their customers effectively by identifying those likely to leave.

Individual contribution and findings:

My role in the project was focused on **Cross Validation and Feature Selection**. I first applied **10-fold cross-validation** to make sure our model was stable and reliable across different data subsets. This helped us validate the accuracy of our logistic regression model.

I also used **Recursive Feature Elimination (RFE)** to select the most relevant features for our prediction. After selecting the top 20 features, I analyzed their correlation to ensure they added meaningful value to the model. This not only improved our model's performance but also reduced complexity.

Through this, I learned how crucial proper feature selection is and how cross-validation builds confidence in our model's results. It gave me hands-on experience with techniques that are widely used in the industry.

Individual contribution to project report preparation:

I worked on the parts of the report that covered **introduction, problem statement, literature review, and feature selection techniques**.

Individual Contribution for Project Presentation and Demonstration:

During the demo, I walked through how we prepared the dataset for training. I explained how we handled missing values, cleaned up unnecessary data, and made sure the dataset was balanced and ready for modeling. My part helped set the foundation for building an accurate prediction model.

Full Signature of Supervisor: **Full Signature of the Student:**

CHURN PREDICTION USING MACHINE LEARNING
JUNED EQBAL
2106124

Abstract:

The aim of this project is to build a machine learning-based churn prediction system to identify customers likely to leave a service. The system analyzes customer data and helps improve retention strategies by providing early churn warnings.

Individual contribution and findings:

I was responsible for the **final phase** of the project, which involved **retraining the model using the selected features** and preparing the **final output** for evaluation.Using the top 20 features selected via RFE, I retrained the logistic regression model and re-evaluated it using key metrics like **accuracy, precision, recall, and F1-score**. I also visualized the confusion matrix for better clarity.To confirm the model's consistency, I performed **cross-validation** again with the selected features. Additionally, I analyzed the model's coefficients and generated the **final output table**, showing actual vs predicted churn values for users.This phase gave me hands-on experience in working with clean, minimal models and how feature selection enhances model accuracy and interpretability. It also helped me understand the importance of presenting the final results in a user-readable format.

Individual contribution to project report preparation:

I contributed to the **System Design section**, including design constraints, system architecture diagrams, and the **Software Requirement Specifications (SRS)** section.

Individual Contribution for Project Presentation and Demonstration:

During the demo, I shared the final churn prediction results and explained what the confusion matrix tells us about the model's performance. I also talked about how simplifying the model by using fewer, more meaningful features actually helped it perform better and made the predictions easier to understand.

Full Signature of Supervisor: **Full Signature of the Student:**

CHURN PREDICTION USING MACHINE LEARNING
SHRUTI KUMARI
2106155

Abstract:

The aim of this project is to create a machine learning-based churn prediction system to identify customers likely to discontinue a service. It processes and analyzes customer data to provide actionable insights and assist in customer retention strategies.

Individual contribution and findings:

I was primarily responsible for the **initial phase** of the project, which involved **loading, cleaning, and visually analyzing the dataset**.

I began by importing essential libraries and reading the dataset. My task included removing invalid entries such as credit scores below 300, and handling missing values in key columns like credit_score and rewards_earned.

After cleaning the dataset, I worked on generating **histograms for all numerical features** to understand their distribution and detect any abnormalities or skewed values. Additionally, I created **pie charts** for categorical variables such as housing, is_referred, and app_downloaded to visualize class proportions.

These visual insights were extremely helpful in guiding further steps like feature engineering and model selection. This phase taught me the importance of clean, well-understood data before applying machine learning techniques.

Individual contribution to project report preparation:

I contributed to the **Conclusion** and **Future Scope** sections of the report, where we summarized our outcomes and discussed how the model can be extended in real-world applications.

Individual Contribution for Project Presentation and Demonstration:

In the presentation, I introduced the idea of customer churn and explained the goal of our project—using data and machine learning to predict which customers might leave. I also discussed why churn prediction matters and how it helps businesses take early action to keep their customers.

Full Signature of Supervisor: **Full Signature of the Student:**

CHURN PREDICTION USING MACHINE LEARNING
MEGHA VERMA
2106125

Abstract:

The aim of this project is to build a machine learning-based churn prediction system that helps in identifying potential customer drop-offs. The objective is to analyze customer behaviour, extract key features, and train a predictive model to aid in business decision-making.

Individual contribution and findings:

My main responsibility was feature exploration, correlation analysis, and final dataset preparation.I began by analyzing individual features such as waiting_4_loan, cancelled_loan, received_loan, rejected_loan, and left_for_one_month in relation to the churn variable. This helped identify which customer behaviours were more strongly associated with churn.Following this, I worked on calculating the correlation between all features and the target variable (churn). I removed less relevant categorical columns and visualized the correlation using a bar graph.I also generated a heatmap of the full correlation matrix using seaborn, which provided insights into multicollinearity and relationships between features. This was further refined by using masking and diverging color palettes for clarity.As a final step, I dropped redundant features like app_web_user, which was highly correlated with other features, and exported the cleaned dataset to a new CSV file to be used in model training.This phase helped me understand how important it is to curate a clean and relevant feature set before model building. It gave me practical exposure to exploratory data analysis and preparation.

Individual contribution to project report preparation:

I contributed to the **Standards Adopted** section in the report, where I documented the **design**, **coding**, and **testing standards** followed throughout the development process.

Individual Contribution for Project Presentation and Demonstration:

In the presentation, I explained how we scaled the data so that all features had a fair impact on the model. I also talked about how we used PCA to reduce the number of features without losing important information, which helped make the model faster and easier to train.

Full Signature of Supervisor: **Full Signature of the Student:**

“Churn Prediction”

ORIGINALITY REPORT

20%	18%	11%	17%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to KIIT University Student Paper	1%
2	www.coursehero.com Internet Source	1%
3	eitca.org Internet Source	1%
4	fastercapital.com Internet Source	1%
5	www.worldleadershipacademy.live Internet Source	1%
6	www.fastercapital.com Internet Source	1%
7	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P. Prasad. "Algorithms in Advanced Artificial Intelligence - Proceedings of International Conference on Algorithms in Advanced Artificial Intelligence (ICAAAI-2024)", CRC Press, 2025 Publication	1%
8	Anurag Palakurti, Divya Kodi. "chapter 6 Building Intelligent Systems With Python", IGI Global, 2025 Publication	1%
9	www.indusedu.org Internet Source	1%
10	Submitted to University College London Student Paper	<1%

11	Aditya Nandan Prasad. "Introduction to Data Governance for Machine Learning Systems", Springer Science and Business Media LLC, 2024 Publication	<1 %
12	Natasa Kleanthous, Abir Hussain. "Machine Learning in Farm Animal Behavior using Python", CRC Press, 2025 Publication	<1 %
13	Submitted to Pace University Student Paper	<1 %
14	web.realinfo.tv Internet Source	<1 %
15	nycdatascience.com Internet Source	<1 %
16	www.svc.ac.in Internet Source	<1 %
17	www.careers360.com Internet Source	<1 %
18	Submitted to CSU, Dominguez Hills Student Paper	<1 %
19	Submitted to Berlin School of Business and Innovation Student Paper	<1 %
20	Submitted to Institute of Research & Postgraduate Studies, Universiti Kuala Lumpur Student Paper	<1 %
21	dev.to Internet Source	<1 %
22	Submitted to University of Wolverhampton Student Paper	<1 %
23	scholarworks.gvsu.edu Internet Source	<1 %

24	www.biorxiv.org Internet Source	< 1 %
25	Submitted to The Scientific & Technological Research Council of Turkey (TUBITAK) Student Paper	< 1 %
26	dataaspirant.com Internet Source	< 1 %
27	Huijian Dong. "Data Analytics in Finance", CRC Press, 2025 Publication	< 1 %
28	Submitted to University of New South Wales Student Paper	< 1 %
29	ijercse.com Internet Source	< 1 %
30	Submitted to Nanyang Polytechnic Student Paper	< 1 %
31	H.L. Gururaj, Francesco Flammini, J. Shreyas. "Data Science & Exploration in Artificial Intelligence", CRC Press, 2025 Publication	< 1 %
32	Submitted to University of Sheffield Student Paper	< 1 %
33	ijistudies.com Internet Source	< 1 %
34	repository.aust.edu.ng Internet Source	< 1 %
35	www.aiplusinfo.com Internet Source	< 1 %
36	H L Gururaj, Francesco Flammini, V Ravi Kumar, N S Prema. "Recent Trends in Healthcare Innovation", CRC Press, 2025 Publication	< 1 %
37	Submitted to Taylor’s Education Group Student Paper	< 1 %

38	Submitted to Whitecliffe College of Art & Design Student Paper	<1 %
39	github.com Internet Source	<1 %
40	sweetishhill.com Internet Source	<1 %
41	uijir.com Internet Source	<1 %
42	www.ijisae.org Internet Source	<1 %
43	Submitted to ESoft Metro Campus, Sri Lanka Student Paper	<1 %
44	Submitted to University of Limerick Student Paper	<1 %
45	machinelearningmodels.org Internet Source	<1 %
46	moldstud.com Internet Source	<1

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		