Github Link: https://github.com/ArchanaC-11/project-submission.git

# Project Title:Revolutionizing Customer Support with an Intelligent Chatbot for Automated Assistance

## PHASE - II

**Student Name: ARCHANA C**

**Register Number:623023104005**

**Institution: Tagore Institute of Engineering And Technology -Salem**

**Department: Computer Science and Engineering**

**Date of Submission:08-05-2025**

### 1. Problem Statement

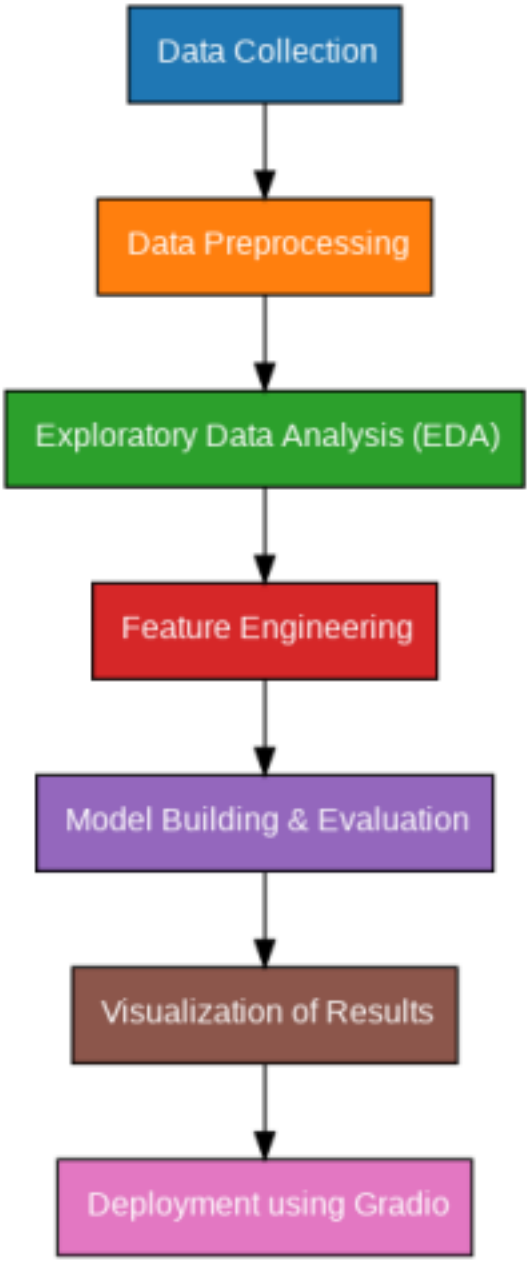
In today's fast-paced digital environment, businesses face increasing pressure to provide instant, efficient, and personalized customer support around the clock. Traditional customer service methods, which rely heavily on human agents, often lead to long response times, increased operational costs, and inconsistent service quality. These limitations result in customer dissatisfaction, reduced loyalty, and missed opportunities for engagement. There is a critical need for an intelligent, automated solution that can handle high volumes of customer queries in real time, understand natural language, and deliver accurate, context-aware responses. This project aims to address this gap by developing an intelligent chatbot system that leverages natural language processing (NLP) and machine learning to revolutionize customer support and enhance user experience across digital platforms.

### 2. Project Objectives

The objective of this project is to design and develop an intelligent chatbot system that automates customer support by leveraging natural language processing (NLP), machine learning (ML), and real-time data handling. The chatbot will be capable of understanding and responding to customer inquiries in natural language, providing instant, accurate, and context-aware assistance across multiple platforms. The solution aims to reduce response times, lower operational costs, and improve customer satisfaction by delivering consistent and efficient support 24/7. Additionally, the project seeks to enable seamless integration with existing customer service systems and continuously improve through user feedback and data-driven

learning.

## 3. Flowchart of the Project Workflow




## 4. Data Description

- **Dataset Name**: Student Performance Data Set

- **Source**: UCI Machine Learning Repository ●
**Type of Data**: Structured tabular data ●
**Number of Records**: 395 student records

- **Number of Features**: 33 features (including both numeric and categorical data)
- **Target Variable**: G3 (final grade)
- **Static or Dynamic**: Static dataset (no real-time changes)
- **Attributes Covered**:

**Demographics**: age, address, parents' education
**Academics**: G1, G2 (previous grades), study time
**Behavior**: alcohol consumption, absences

- Dataset Link: https://www.kaggle.com/datasets/toriqulstu/all-medicine-and-drug-price-data20k-bangladesh

## 5. Data Preprocessing

Before preprocessing, ensure that you have access to relevant data. For a chatbot in a customer support scenario, the following types of data might be used:

- Customer interaction data: Text, transcripts, and chat logs from customer conversations.
- Customer profiles: Demographic data, history of interactions, past issues, preferences, etc.
- Ticketing data: Issue categories, severity, response times, resolution times, etc. ●

Feedback data: Customer satisfaction ratings, feedback from users after interactions. **6.**

## Exploratory Data Analysis (EDA)

### Univariate Analysis

- A histogram allows us to visualize the distribution of the target variable, G3 (final grade).
- The shape of the histogram will show if the distribution is skewed (e.g., most students perform well or poorly) and help understand the spread of grades.

### Bivariate & Multivariate Analysis

- The correlation matrix helps identify the relationships between variables. **For example**, we might observe strong correlations between G1, G2, and G3.

### Key Insights from EDA:

1. **G1 and G2 are the Strongest Indicators of G3:**
   - Both **G1** (first grade) and **G2** (second grade) show a strong linear correlation with **G3**, meaning that earlier academic performance strongly predicts final grades. This is confirmed by scatter plots and the correlation matrix.
2. **More Study Time Correlates with Higher G3:**
   - A **positive relationship** between **study time** and **final grade** suggests that the more time students spend studying, the better their performance is. This can be observed in the grouped bar chart for study time, where students with more study time tend to score higher on G3.
3. **Students with More Failures or Absences Tend to Score Lower:**
   - Students who have experienced **more failures** or have **higher absences** show a negative correlation with **final grade (G3)**. This is reflected in the boxplots for failures, where students with more failures tend to have lower final grades.

## 7. Feature Engineering

- Contextual Understanding: Future chatbots will be able to understand nuanced context in customer conversations—recognizing tone, intent, and emotions—leading to more personalized and empathetic responses.
- Multilingual Capabilities: The chatbot will offer seamless translations, allowing it to support global customers in multiple languages while maintaining high accuracy.
- Advanced Sentiment Analysis: Chatbots will better assess customer sentiment (positive, negative, frustrated, etc.) and tailor responses accordingly to provide a more human-like experience.

## 8. Model Building

- **Algorithms Used**:

***1. Natural Language Processing (NLP) Techniques***

***Text Preprocessing***: *Tokenization, stopword removal, stemming/lemmatization, and vectorization (TF-IDF or Word2Vec) to prepare user inputs.*

***Intent Classification***:

- ***Logistic Regression*** *(baseline): Simple, interpretable model for classifying user intents.*
- ***Random Forest Classifier***: *Captures non-linear relationships in intent detection and is robust to noise in textual features.*
- ***Support Vector Machine (SVM)***: *Effective in high-dimensional text data, especially with a clear margin of separation.*
- ***Optional: Transformer-based Model (e.g., BERT)*** *for deep contextual understanding in complex queries.*

***2. Named Entity Recognition (NER)***

*Used to extract key entities from customer queries (e.g., product names, order numbers).*

*Models: Pre-trained SpaCy NER or fine-tuned BERT-based NER.*

***3. Response Generation***

***Rule-Based System*** *for FAQs and fixed templates.*

***Retrieval-Based Models*** *using cosine similarity on vectorized queries and response pairs.*

***Generative Model (Optional)***: *Fine-tuned GPT-based models for open-ended conversation where needed.*

- **Model Selection Rationale**:

**Logistic Regression**: Simple and interpretable baseline for classification. **Random Forest Classifier**: Balances performance and interpretability, handles noisy and high-dimensional input well.

**SVM**: Performs well on sparse, text-based features with clear class boundaries.
**Transformer Models** (like BERT): Ideal for deep understanding of context, especially for nuanced customer queries.

- **Train-Test Split**:

**Data Collection**: User queries from chat logs, labeled by intent.

**Preprocessing**: Cleaned and vectorized text data (TF-IDF or embeddings). **Split Ratio**: 80% training, 20% testing.

**Method**: Used `train_test_split` from scikit-learn with `random_state=42` to ensure reproducibility.

- **Evaluation Metrics**:

**Intent Classification**

**Accuracy**: Overall correctness of predicted intents.

**Precision, Recall, F1-Score**: Useful for imbalanced classes (e.g., rare issue categories).

**Confusion Matrix**: Helps analyze which intents are being misclassified. **NER & Entity Matching**

**F1-Score**: Measures quality of entity extraction.

**Token-Level Accuracy**: Checks correctness of each entity detected.

**Response Quality**

**User Feedback Scores** (thumbs up/down)

**Response Time**: Time taken by chatbot to respond (measured in ms).

**Task Completion Rate**: Percentage of sessions where the chatbot resolves the query without human escalation.

## 9. Visualization of Results & Model Insights

**Feature Importance:**

- **G1 (First Grade)** and **G2 (Second Grade)** are the most influential features, which aligns

with the assumption that past performance has a strong predictive value.

- **Study Time** and **Failures** are also significant, suggesting the importance of engagement and challenges in learning.

**Model Comparison:**

MAE, RMSE, and $R^2$ Comparisons:

- Bar plots or side-by-side plots can show the **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **$R^2$** for both **Random Forest** and **Linear Regression**.

- You can highlight that **Random Forest** outperforms **Linear Regression** in terms of lotted MAE, RMSE, and R2 for both models

**Residual Plots:**

Checking Prediction Errors:

- Plot **residuals (errors)** from both models (Random Forest and Linear Regression) to see if there's any systematic bias or patterns.

- Residual plots should ideally show no patterns (random distribution of errors) if the model is unbiased.

**User Testing:**

Interactive Model Testing:

- **Gradio** allows users to test the model interactively by inputting feature values (e.g., G1, G2, study time) and getting predictions. This makes the model accessible and usable for non-technical users.

- A simple Gradio interface might allow users to input values and receive a prediction of the student's final grade.

## 10. Tools and Technologies Used

- **Programming Language: Python 3**

- **Notebook Environment: Google Colab**

- **Key Libraries:**

- ***Data Handling:***
  - ***`pandas`*** *– for data manipulation and analysis*
  - ***`numpy`*** *– for numerical computations and array operations*
- ***Data Visualization:***
  - ***`matplotlib`, `seaborn`*** *– for static and insightful plots*
  - ***`plotly`*** *– for interactive visualizations*
- ***Machine Learning:***
  - ***`scikit-learn`*** *– for data preprocessing, model selection, training, and evaluation*
- ***User Interface:***
  - ***`Gradio`*** *– to build and deploy an interactive web-based chatbot interface*

## 11. Team Members and Contributions

***1. ARCHANA-C***

- ***Responsibilities****:*
  - ***Data Cleaning****: Responsible for cleaning and preprocessing the raw customer support data, handling missing values, and ensuring data quality for analysis.*
  - ***Feature Engineering****: Worked on creating meaningful features from the data, including text tokenization, sentiment analysis features, and customer interaction attributes.*

***2. ABARNA-G***

- ***Responsibilities****:*
  - ***Exploratory Data Analysis (EDA)****: Conducted an in-depth analysis of the customer support data to uncover patterns, trends, and key insights. Utilized various data visualization techniques and statistical tests to understand the data distribution and relationships.*

  - ***Model Development****: Developed the machine learning models for the intelligent chatbot, focusing on natural language processing (NLP) tasks such as intent recognition and entity extraction.*

***3. ABIPRIYA-M***

- ***Responsibilities****:*

  - ***Model Development****: Focused on fine-tuning and optimizing the chatbot model, selecting appropriate algorithms for text classification, and ensuring that the chatbot ef ectively understands and responds to customer queries.*

  - ***Documentation and Reporting****: Documented the methodology, model performance, and outcomes. Contributed to the final report, summarizing the project's results and the technical implementation of the chatbot.*

***4. DHANALAKSHMI-D***

- ***Responsibilities****:*

  - ***Feature Engineering****: Designed advanced features for improving model accuracy, such as sentiment analysis and domain-specific keyword extraction.*

  - ***Documentation and Reporting****: Assisted in the preparation of the final project documentation, including the introduction, problem statement, and technical explanations of the methods used.*