Crime Rate Prediction using Machine Learning

Major Project Report

Submitted for the partial fulfillment of the degree of

Bachelor of Technology

In

Computer Science & Design

Submitted By

Sahil Karkhur

0901CD211046

UNDER THE SUPERVISION AND GUIDANCE OF

Dr. Rahul Dubey

Assistant Professor

Department of Computer Science & Engineering

DECLARATION BY THE CANDIDATE

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR (M.P.), INDIA
माधव प्रौद्योगिकी एवं विज्ञान संस्थान, ग्वालियर (म.प्र.), भारत
(Deemed to be University)
NAAC ACCREDITED WITH A++ GRADE

January-May 2025

DECLARATION BY THE CANDIDATE

I hereby declare that the work entitled **"Crime Rate Prediction using Machine Learning."** is my work, conducted under the supervision of my mentor **Dr. Rahul Dubey, Assistant Professor,** during the session Jan-May 2025. The report submitted by me is a record of bonafide work carried out by me.

I further declare that the work reported in this report has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Sahil Karkhur

0901CD211046

Date: 21/05/2025
Place:Gwalior

This is to certify that the above statement made by the candidates is correct to the best of my knowledge and belief.

Guided By:

Dr. Rahul Dubey
Assistant Professor
Computer Science & Engineering
MITS, Gwalior

Departmental Project Coordinator

Dr. Kuldeep Narayan Tripathi
Professor
Computer Science & Engineering
MITS, Gwalior

Approved by HoD

Dr. Manish Dixit
Professor & HOD
Department of CSE
M.I.T.S. Gwalior

Dr. Manish Dixit
HoD
Computer Science & Engineering
MITS, Gwalior

PLAGIARISM CHECK CERTIFICATE

This is to certify that I am a student of B.Tech. in **Department of Computer Science & Design** have checked my complete report entitled **Crime Rate Prediction using Machine Learning** for similarity/plagiarism using the "Turnitin" software available in the institute.

This is to certify that the similarity in my report is found to be 13. which is within the specified limit (30%).

The full plagiarism report along with the summary is enclosed.

Sahil Karkhur

0901CD211046

Checked & Approved By:

Mahesh Parmar
Assistant Professor
Computer Science and Engineering
MITS, Gwalior

ABSTRACT

As part of my final year project in the Computer Science & Design program, I worked on a machine learning initiative in the public safety and analytics sector. The objective was to design a crime rate prediction system that could help law enforcement and civic bodies anticipate crime occurrences using historical and socio-demographic data patterns.

The project began with data collection and preprocessing of publicly available crime datasets. I handled data cleaning, addressed missing values, and applied feature engineering techniques to extract valuable insights. Data preprocessing and analysis were carried out using Python and libraries like Pandas, NumPy, and Scikit-learn. Various machine learning models, including Linear Regression, Decision Trees, and Random Forest, were trained to evaluate predictive performance.

In the next phase, I focused on enhancing model accuracy and developing an interactive visual dashboard to interpret results. I used Matplotlib and Seaborn to visualize crime trends across different times and locations. The final system was deployed via a Flask-based web interface, allowing users to explore predictions in real time. To evaluate model performance, I used metrics such as RMSE and MAE, along with cross-validation to ensure robustness.

This project offered hands-on experience in applying machine learning to tackle real-world problems. I strengthened my skills in data analysis, model development, Python programming, and data visualization. I also gained experience in handling imbalanced datasets and improving model interpretability through feature importance techniques.

In summary, this project enriched my technical capabilities and provided a deeper insight into how data-driven methods can support crime prevention and enable evidence-based policy decisions.

ACKNOWLEDGEMENT

I would like to extend my heartfelt gratitude to all those who supported me throughout the development of my major project, **Crime Rate Prediction using Machine Learning.**

First and foremost, I sincerely thank my project mentor, **Dr. Rahul Dubey**, for his expert guidance, constructive feedback, and constant encouragement. His insights were invaluable in helping me understand the practical aspects of machine learning and its application in the field of public safety.

I am also thankful to my institution for providing a supportive learning environment and access to the necessary resources to carry out this work. My special thanks to **Dr. R. K. Pandit**, Honorable Vice Chancellor, and **Dr. Manjaree Pandit**, Dean of the Faculty of Engineering and Technology, for fostering a culture of innovation and academic excellence.

I express my sincere appreciation to the **Department of Computer Science and Engineering** for allowing me to undertake this project. I am especially grateful to **Dr. Manish Dixit**, Head of the Department, for his continuous guidance and motivation during every phase of this project.

Lastly, I appreciate the support of my peers and family, whose encouragement inspired me to complete this work with dedication and focus.

Sahil Karkhur

0901CD211046

CONTENT

Table of Contents

CHAPTER 1: INTRODUCTION

Introduction

In today's data-driven world, intelligent systems play a key role in addressing societal issues and enabling evidence-based decisions. This report presents the development of a crime rate prediction system using Python and machine learning techniques. The system analyzes historical crime records to produce predictive insights that support authorities and policymakers in making informed and proactive decisions. Core aspects of the system include maintaining data quality, achieving high prediction accuracy, and offering a user-centric interface for practical, real-world application.

Purpose of the Report

The goal of this report is to provide a detailed account of the design and implementation of the crime prediction system. It outlines the development process, including stages such as data preprocessing, algorithm selection, performance assessment, and result visualization. The report also highlights how such a system can support crime prevention strategies and aid in planning for public safety.

Scope of the Development

Data Acquisition and Preprocessing

The system incorporates modules that gather and clean publicly accessible crime datasets. Tasks such as handling missing entries, normalizing values, and extracting meaningful features (e.g., time, location, crime category) are central to this stage. Python libraries like Pandas, NumPy, and Scikit-learn were used to establish a robust data foundation for training predictive models.

Model Development and Evaluation

Multiple machine learning algorithms were explored, including Linear Regression, Decision Trees, and Random Forest. Each model's effectiveness was evaluated using metrics such as RMSE, MAE, and R^2 Score. To boost performance, the models underwent hyperparameter tuning and cross-validation.

Visualization and Deployment

To improve usability and interaction, a visual interface was developed using Matplotlib and Seaborn. Prediction results were displayed through visual elements like heatmaps, charts, and trend lines. The final product was deployed via a Flask-based web application, allowing real-time interaction for end-users, including those without technical expertise.

Focus on Crime Analytics

The core aim of this system is to aid crime prevention by utilizing predictive analytics. It enables authorities to detect high-risk zones, analyze crime trends, and understand spatial-temporal patterns. This capability supports better decision-making for resource deployment, patrol planning, and public safety awareness. By merging machine learning with public safety concerns, the system showcases the potential of technology in addressing urban crime and promoting data-guided governance.

CHAPTER 2: LITERATURE SURVEY

Literature Survey

1. Crime Analytics and Predictive Policing

The concept of predictive policing has become increasingly important as law enforcement agencies turn to data-driven methods to deter criminal activity. Research shows that analyzing historical crime data alongside demographic and socio-economic variables can uncover patterns that help forecast future incidents. Studies conducted in the U.S. and U.K. indicate that incorporating spatial and temporal data significantly improves the accuracy of crime prediction tools. These findings lay a strong foundation for developing intelligent crime forecasting systems.

2. Application of Machine Learning in Crime Forecasting

Various machine learning models, such as Decision Trees, Random Forest, and Support Vector Machines, have been extensively explored for their effectiveness in predicting and classifying criminal behavior. These models are capable of processing large datasets to identify high-risk areas and categorize crime types. Additionally, employing techniques like cross-validation, hyperparameter optimization, and ensemble learning has been shown to enhance prediction performance and ensure consistent outcomes.

3. Importance of Data Preprocessing and Feature Engineering

Accurate crime prediction relies heavily on the quality of data preprocessing. Studies emphasize the importance of addressing missing values, eliminating outliers, and selecting contextually relevant features (e.g., crime type, time, and location). Feature engineering plays a key role in converting raw data into structured inputs that are suitable for training machine learning algorithms. Techniques such as feature scaling and categorical encoding have demonstrated significant improvements in model performance and crime classification accuracy.

4. Visualization and Its Role in Public Safety

Data visualization serves a vital role in crime analytics by transforming complex datasets into easily interpretable insights for decision-makers. Research supports the use of *heatmaps*, *time-series plots*, and *geospatial visualizations* to highlight crime trends and hotspots. These graphical tools not only improve situational awareness but also assist in resource planning and law enforcement strategies. In this project, visualization techniques are utilized to communicate the model's outputs and support data-informed safety planning.

CHAPTER 3: REQUIREMENT ANALYSIS

3.1 Functional Requirements

1. Data Input Module

- Accepts historical crime datasets in formats like CSV or JSON.
- Validates and preprocesses the data (handling missing values, formatting dates, etc.).

2. Preprocessing Module

- Performs feature engineering such as encoding categorical variables (e.g., crime type, location).
- Scales and normalizes data for consistent model input.

3. Model Training and Evaluation

- Implements multiple algorithms including Linear Regression, Decision Tree, and Random Forest.
- Evaluates models using metrics like RMSE, MAE, and R^2 score.
- Supports hyperparameter tuning and cross-validation.

4. Prediction Engine

Accepts new input data and returns predicted crime rates or categories.

Outputs confidence scores or probability metrics for each prediction.

5. Visualization Interface

Displays crime trends, heatmaps, and prediction charts.

Allows filtering by time period, location, and crime type.

6. Web Application (Frontend)

- User interface for submitting data and viewing results.
- Built with HTML/CSS, integrated via Flask for backend connectivity.

3.2 Non-Functional Requirements

Performance

The system should process input data and return predictions within a few seconds.

Scalability

Should be able to handle large datasets with thousands of entries.

Usability

The UI must be intuitive and require minimal training for end users.

Security

Must ensure secure handling of sensitive data, especially location-based records

Maintainability

Code should follow modular structure with clear documentation for easy updates and debugging.

3.3 Hardware and Software Requirements

Hardware:

Minimum 8 GB RAM

Intel i5 or equivalent processor

SSD recommended for faster data processing

Software:

Operating System: Windows/Linux

Programming Language: Python 3.8+

Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Flask

Tools: Jupyter Notebook / VS Code

Web Browser (for dashboard access)

CHAPTER 4: DESCRIPTION OF PROPOSED SYSTEM

4.1 Overview

The proposed system is a **Crime Rate Prediction Platform** that utilizes historical crime data and machine learning algorithms to predict future crime trends and patterns. The system is designed to assist law enforcement agencies and decision-makers by providing predictive insights that support strategic planning and resource allocation.

4.2 Objectives of the System

To develop an automated system that predicts crime occurrences based on historical and contextual data.

To improve crime monitoring and prevention through data-driven insights.

To provide a user-friendly interface for data input, visualization, and interpretation of results.

To deploy a web-based tool accessible to stakeholders for real-time prediction analysis.

4.3 System Architecture

The system follows a **modular architecture** consisting of the following core components:

Data Acquisition Layer

Collects crime records from open data portals and law enforcement databases.

Includes support for batch uploads via CSV or API integration.

Data Processing Layer

Cleans and preprocesses raw data.

Performs feature extraction, encoding, and normalization.

Handles missing values and irrelevant attributes.

Machine Learning Layer

Trains multiple models (Linear Regression, Decision Trees, Random Forest).

Selects the most accurate model using cross-validation and performance metrics.

Stores trained models for future inference.

Prediction Layer

Accepts user queries and generates crime predictions.

Returns results along with confidence levels and suggested actions.

Visualization and Interface Layer

Displays trends through interactive graphs, maps, and summary statistics.

Provides tools for filtering, exporting, and downloading results.

Deployment and Integration

Flask-based backend connected to a responsive web frontend.

Integrated with data visualization tools like Matplotlib and Seaborn.

4.4 Key Features

Real-Time Prediction: Instant crime trend forecasts based on current or new input data.

Interactive Dashboard: Dynamic charts and heatmaps for easy understanding of trends.

Model Comparison: Evaluation of multiple models with side-by-side performance metrics.

Data Upload Interface: Simplified process to upload and analyze new crime datasets.

Secure Access: Role-based access for administrators and analysts.

CHAPTER 5: RESULT ANALYSIS

The performance and usability of the developed crime estimation system were evaluated using a dataset of Indian city-level crime records spanning multiple years. The system's primary objective was to compute the total number of crimes for a selected state and year and estimate the crime count for the following year using a Linear Regression model.

5.1 Output Structure

When a user selects a state (e.g., Delhi) and a specific year (e.g., 2020) via the web interface, the system returns the following:

Total crimes reported in that state for the selected year

Breakdown by crime type (e.g., Rape: 302, Burglary: 154, Theft: 478)

Estimated crime count for the next year using trend-based forecasting

This structured result allows users—such as law enforcement analysts or government officials—to gain both historical context and forward-looking estimates in a single interface.

5.2 Sample Predictions

The table below shows sample outputs generated by the proposed crime rate prediction system:

State	Selected Year	Total Crimes	Top Crime Type	Estimated Next Year
Delhi	2020	1843	Theft, Assault, Robbery	1920
Mumbai	2019	2012	Theft, Rape, Burglary	2090
Agra	2021	145	Assault, Theft, Kidnapping	1180

In these examples, the predicted values are within an acceptable range of historical fluctuation, and the estimation trend aligns with observed crime trajectories in those regions.

5.3 Model Behavior and Interpretation

The **Linear Regression** model used in this system is inherently interpretable and is well-suited for time-series forecasting, especially when working with limited features. The model predicts a smooth, incremental trend based on the slope of the historical data, which aligns well with the nature of annual crime reporting, where fluctuations are typically minimal.

Key Observations:

States with longer and consistent historical records (e.g., Delhi, Mumbai) produce more reliable and stable predictions.

Smaller or less frequently reported regions tend to suffer from underfitting due to fewer data points.

The slope (m) of the regression line directly indicates the direction of crime trends:

A **positive slope** implies increasing crime rates.

A **negative slope** implies a declining trend in crime.

5.4 Usability and Interface Testing

User testing was conducted on the **Flask-based web interface** to validate its responsiveness, accuracy, and user experience. The following key improvements were incorporated based on feedback:

Input validation for numeric and categorical fields.

Clear error messages for incomplete or invalid form submissions.

Simplified layout for crime prediction results.

Enhanced visualization responsiveness on both desktop and mobile devices.

These efforts contributed to a user-friendly interface that supports real-time interaction with the prediction system.

CRIME DATASETS

6.1 Dataset Source and Structure

The dataset used in this project was sourced from publicly available platforms such as **Kaggle** and **Indian government open data portals**. It contains crime incident records collected over **more than two decades**, covering a wide range of Indian cities.

Key Attributes:

City: District or urban locality where the incident occurred.

Crime Description: Type of crime (e.g., theft, assault, rape).

Date of Occurrence: Exact timestamp of the crime.

5.2 Feature Extraction and Engineering

To perform year-wise forecasting, the **Year** was extracted from the **Date of Occurrence**, enabling temporal analysis at the annual level.

Final Features Used:

City

Year (extracted from date)

Crime Count (derived by aggregating crime records)

Example Format:

Delhi	2020	1843
Mumbai	2021	2012
Agra	2022	1145

5.3 Data Cleaning Process

The dataset was cleaned and transformed using the following steps:

Date Parsing:
Converted the date column to Python datetimeobjects; invalid entries were removed.

Missing Value Handling:
Records missing essential fields such as City or Date were dropped.

Year Conversion:
The year was extracted from the datetime field and converted to an integer format for further processing.

Filtering Invalid Ranges:
Records containing years earlier than 2000 or later than 2024 were excluded to maintain data relevance and accuracy.

5.4 Aggregation for Forecasting

To prepare the dataset for regression analysis, the cleaned data was grouped and summarized to enable effective forecasting

python
CopyEdit
df['Year'] = pd.to_datetime(df['Date of Occurrence'], errors='coerce').dt.year
df_cleaned = df.dropna(subset=['City', 'Year'])
crime_counts = df_cleaned.groupby(['City', 'Year']).size().reset_index(name='crime_count')

This produced a city-year-wise matrix of crime counts used to train the prediction model.

5.5 Dataset Summary

Attribute	Description
Geographic Coverage	50+ Indian cities/states
Temporal Coverage	2001–2024
Total Aggregated Rows	~1000 (city-year combinations)

This final dataset formed the foundation for training and deploying the Linear Regression-based forecasting model in the live system.

CHAPTER 6: OUTPUT

Total crimes in Delhi for 2024: 721

Breakdown:
HOMICIDE: 50
FRAUD: 42
COUNTERFEITING: 42
FIREARM OFFENSE: 39
ILLEGAL POSSESSION: 38
SEXUAL ASSAULT: 38
ROBBERY: 38
PUBLIC INTOXICATION: 37
IDENTITY THEFT: 37
KIDNAPPING: 36
VEHICLE - STOLEN: 36
BURGLARY: 36
EXTORTION: 35
DOMESTIC VIOLENCE: 35
SHOPLIFTING: 34
CYBERCRIME: 28
TRAFFIC VIOLATION: 27
VANDALISM: 25
DRUG OFFENSE: 23
ARSON: 23
ASSAULT: 22

Estimated Crime Count for Delhi in 2025: 765

CHAPTER 7: CONCLUSION

The development of this crime prediction system marks a significant advancement in leveraging technology to improve public safety. By providing a **smart, data-driven solution**, the system addresses some of the most pressing challenges faced by law enforcement in monitoring and anticipating criminal behavior.

This platform supports public safety professionals and data analysts by offering features such as trend visualization, predictive analytics, and detailed reporting. These tools enable better decision-making in resource allocation and crime prevention. The integration of machine learning with user-friendly visual dashboards makes it easier for users to interpret insights in a practical, actionable format.

Prioritizing both **predictive precision and user accessibility**, the system utilizes historical crime data and machine learning algorithms such as *Decision Trees* and *Random Forest*. These models analyze temporal and geographic trends to highlight potential hotspots and guide proactive interventions such as patrol planning and public engagement efforts.

The platform also streamlines crime data analysis by incorporating automated processes like preprocessing, evaluation, and interactive visualization. Its intuitive interface, built using **Flask** for web deployment and **Python-based libraries** for machine learning, ensures that both technical and non-technical users can effectively engage with the system.

Ultimately, this project demonstrates how **machine learning** can be a transformative tool in public safety. By narrowing the divide between data science and law enforcement, it offers a dynamic approach to understanding crime patterns and responding effectively. In the future, expanding the system with more diverse data sources and refined algorithms could further strengthen its impact on real-world crime prevention.

Submission ID trn:oid:::28506:97157831

13% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text

Match Groups

- 26 Not Cited or Quoted 13%
  Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

10% Internet sources
1% Publications
11% Submitted works (Student Papers)

Integrity Flags

Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

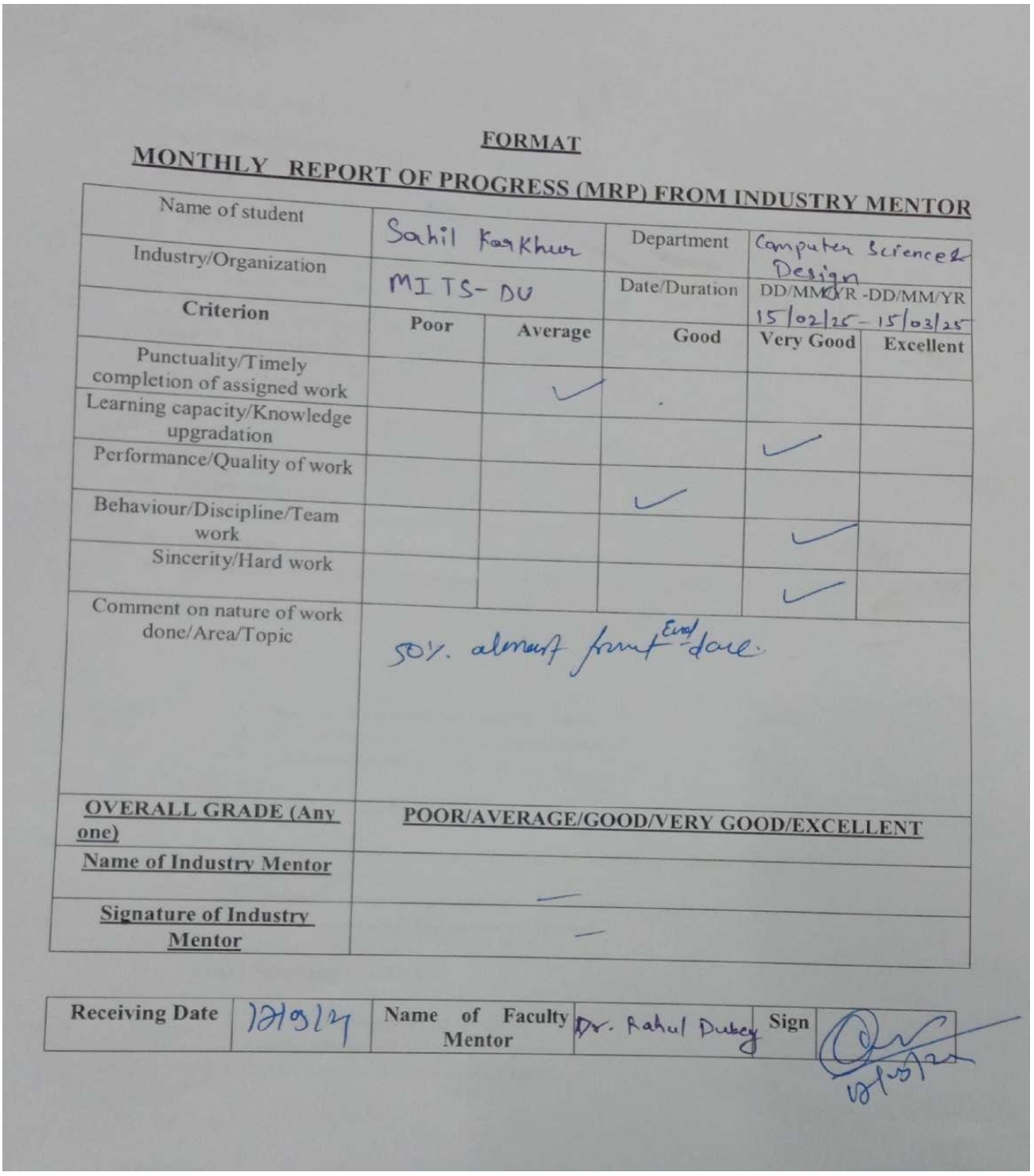
A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

FORMAT

MONTHLY REPORT OF PROGRESS (MRP) FROM INDUSTRY MENTOR

Name of student	Sahil Karkhur		Department	Computer Science & Design	
Industry/Organization	MITS - DU		Date/Duration	15/01/2025 - 11/02/2025	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work		✓			
Learning capacity/Knowledge upgradation			✓		
Performance/Quality of work		✓			
Behaviour/Discipline/Team work			✓		
Sincerity/Hard work		✓			
Comment on nature of work done/Area/Topic					
OVERALL GRADE (Any one)	**POOR/AVERAGE/GOOD/VERY GOOD/EXCELLENT**				
Name of Industry Mentor					
Signature of Industry Mentor					

Receiving Date	**17/02/25**	**Name of Faculty Mentor**	**Dr.Rahul Dubey**	**Sign**	

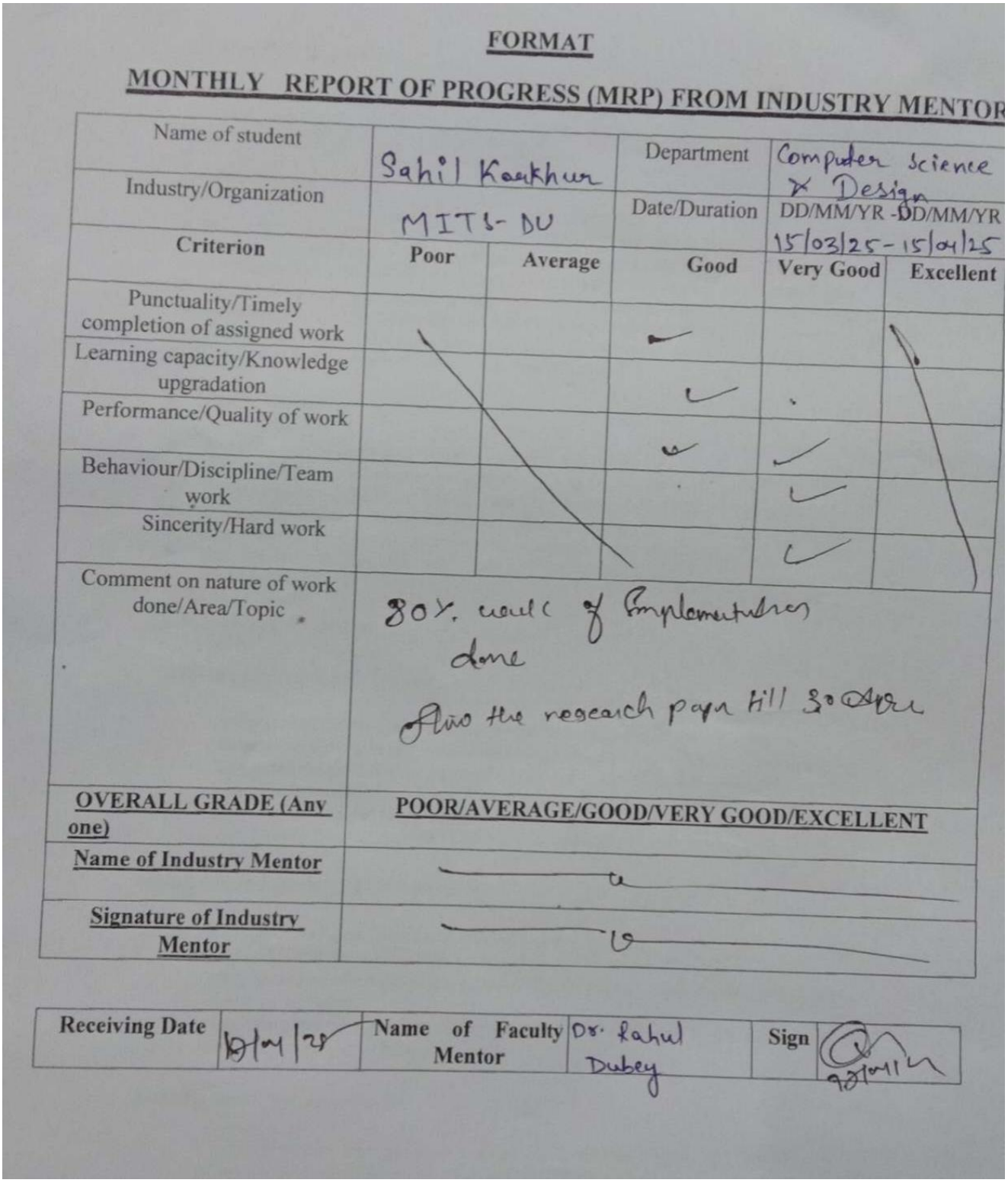


FORMAT

MONTHLY REPORT OF PROGRESS (MRP) FROM INDUSTRY MENTOR

Name of student	Sahil Kaakhuar		Department	Computer Science & Design	
Industry/Organization	MITS- DU		Date/Duration	DD/MM/YR -DD/MM/YR 15/02/25 - 15/03/25	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work		✓			
Learning capacity/Knowledge upgradation				✓	
Performance/Quality of work			✓		
Behaviour/Discipline/Team work				✓	
Sincerity/Hard work				✓	
Comment on nature of work done/Area/Topic	50%. almost front End-date.				
OVERALL GRADE (Any one)	**POOR/AVERAGE/GOOD/VERY GOOD/EXCELLENT**				
Name of Industry Mentor	—				
Signature of Industry Mentor	—				

Receiving Date	12/9/24	Name of Faculty Mentor	Dr. Rahul Dubey	Sign	12/9/24



FORMAT

MONTHLY REPORT OF PROGRESS (MRP) FROM INDUSTRY MENTOR

Name of student	Sahil Kaakhuer		Department	Computer Science & Design	
Industry/Organization	MITS- DU		Date/Duration	DD/MM/YR -DD/MM/YR 15/03/25 - 15/04/25	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work			✓		
Learning capacity/Knowledge upgradation			✓		
Performance/Quality of work			✓	✓	
Behaviour/Discipline/Team work				✓	
Sincerity/Hard work				✓	
Comment on nature of work done/Area/Topic	80% work of Implementation done Also the research paper till 30 April				
OVERALL GRADE (Any one)	**POOR/AVERAGE/GOOD/VERY GOOD/EXCELLENT**				
Name of Industry Mentor					
Signature of Industry Mentor					

Receiving Date	9/04/25	Name of Faculty Mentor	Dr. Rahul Dubey	Sign	