CASE BASED REASONING MODEL TO OPTIMIZE SENTENCING GUIDELINES IN THE JUDICIARY OF KENYA

ASIAGO, DONALD CHARANI

A RESEARCH PROJECT SUBMITTED TO THE SCHOOL OF MATHEMATICS AND COMPUTING IN PARTIAL FULFILLMENT FOR THE AWARD OF DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE OF THE COOPERATIVE UNIVERSITY OF KENYA.

NOVEMBER, 2025

Declaration

Declaration by the Candidate

This research project is my original work and has not been presented for the award of a degree in any university or for any other award.

21/11/2025

Signed --- ………….......................................

Donald Charani Asiago. **Date**

MDATC01/6064/2022

Declaration the by Supervisors

We confirm that the work reported in this research project was carried out by the candidate under our supervision and has been submitted with our approval as university supervisors.

Signed: PROF. J. M. KIHORO

25/11/2025

Signature---Date-------------------------------------

Department of Mathematical Sciences

The Cooperative University of Kenya

Signed: **Dr. Patrick Owoche**

Signature: .... Date: ....**24/11/2025**............

Department of Information Technology

Kibabii University

Dedication

I dedicate this work to the Almighty God who has been my source of Strength, Grace, and Wisdom throughout my study, to my wife who has been a pillar by providing good environment and giving me ample time to concentrate on my studies, my mentors Prof. Kihoro and Dr. Owoche, my fellow students and all my friends who have been a constant source of inspiration. You have given me the drive to tackle any task with enthusiast and determination. Without your love and support this project will not have been made possible.

Acknowledgement

My deepest gratitude goes to God who provided all that was needed to complete this study. I will also like to express my deep appreciation to Prof. John Kihoro and Dr. Patrick Owoche for their unwavering support, insightful remarks and invaluable guidance through this process.

I also wish to extend my sincere gratitude to Co-operative University of Kenya for providing me any opportunity, resources and support to realize my goal.

Finaly, I wish to express my gratitude to my wife, friends and fellow students for their unwavering support during the challenging moments and to everyone who played a role in shaping this project directly or indirectly, thank you for your contributions and belief in this endeavor

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CBR-	Case Reasoning Model.
CJS-	Criminal Justice System.
NBER-	National Bureau of Economic Research.
NLP-	Natural Language Processing.
RBS-	Rule-based Systems.

Abstract

Judicial sentencing often faces challenges of inconsistency and disparity, undermining both fairness and public trust. Existing guidelines provide direction, but variations persist due to the subjective interpretation of cases. This research addressed this gap by developing a Case-Based Reasoning (CBR) model to optimize sentencing decisions in the Kenyan judiciary. The study developed a CBR model that leveraged historical judicial data to support judges in delivering more consistent and equitable sentences. The methodology employed a structured four-stage approach using cosine similarity and TF-IDF vectorization for case retrieval, ensemble methods (Random Forest, XGBoost, LightGBM) for prediction, k-fold cross-validation for validation, and hyperparameter tuning for optimization. The CBR model incorporated a structured four-stage approach: case retrieval, case adaptation, case validation, and case optimization. The model analyzed key case attributes, including defendant demographics, crime severity, mitigating and aggravating factors, and judicial reasoning in prior cases. It was trained with a dataset of 1,200 previous cases from Kenyan courts, historical judicial sentencing records sourced from public legal databases e.g., Supreme Court archives, national judicial sentencing datasets and other legal databases with anonymized sentencing data. The results showed that the CBR model achieved a prediction accuracy of 27.42% with an F1-score of 0.2752, indicating significant challenges in judicial predictive modeling due to case complexity. Despite moderate performance, fairness analysis revealed balanced treatment across demographic subgroups, with fairness scores above 0.85. Thus, this approach contributed to aligning sentencing with judicial guidelines and helped in maintaining transparency and accountability in legal decisions.

CHAPTER ONE

1.0 Introduction

This chapter presents background of the study, statement of the problem, objectives of the study, research questions, significance of the study, justification, limitation, expected study outcomes, and scope of the study.

1.1 Background of Study

Sentencing guidelines are important tools in the criminal justice system since they give a basis on which judges can hang a sentence on individuals found guilty of certain crimes. Because of this, these guidelines have been developed to ensure that similar cases are treated similarly, thus reducing disparity and promoting transparency in the practice of sentencing. In Kenya specifically, the judiciary has faced particular challenges with sentencing inconsistencies, especially in cases involving similar offenses across different counties. The Kenyan legal system, while grounded in common law principles, has experienced variations in sentencing practices that undermine public confidence in judicial fairness. Sentencing, however, is complex, given the need to balance a broad general standard against the particular circumstances of the case before the court.

Historic patterns of sentencing have been mostly humanistic in nature, with statutory rules and precedents being used to supplement human judgment. Such approaches give much freedom to the judges, and this may lead to the absence of standard sentences since judges may read guidelines differently, based on personal experiences and case-related considerations. Indicatively, research and investigations in courts like those in the United States have revealed differences in sentencing platforms regarding the race, socioeconomic background and geographical location (Spohn, 2000; Ulmer, 2012).

This is therefore the challenge that has recently seen some promising methodologies inspired by advances in artificial intelligence, which can support the synthesis and analysis of such complex case data in aid of judicial decisions. AI-driven sentencing tools have been piloted in some jurisdictions to assist judges by providing data-driven insights while maintaining judicial discretion (Angwin et al., 2016).

Emphasis shall be on coming up with a Case Based Reasoning model, which is capable of relying on different past cases, to ensure the sentences presented by the judges are consistent, fair and justifiable. The strategy will be used to reduce inconsistencies in sentencing and enhance adherence to recommendations to regain the trust in court decisions.

1.2 Problem statement.

Sentencing guidelines rely on decisions made by human beings with resulting advantages yet human assessment gets compromised by individual prejudices and unstandardized approaches and legal comprehension differences. The sentencing of similar offenses faces inconsistencies because judges have various subconscious biases as well as personal experiences combined with different interpretations of legal precedents. The outcome of sentences becomes both unfair and inappropriate due to this process. Judges now face an increasingly complex challenge to maintain case consistency due to elevated dispute complexity together with the many precedents that occur.

Empirical research points to the fact that disparities in sentencing are a longstanding problem of the judicial system across in Kenya. As an example, a study conducted by the National Bureau of Economic Research (NBER) revealed that the sentence of the same crime may differ dramatically depending on the specifics of the case, like the judge of the case, demographics of the defendant, or the complexity of the case (Abrams et al., 2020). In the Kenyan context specifically, studies have documented significant sentencing disparities across different courts, with similar offenses receiving vastly different penalties depending on the presiding judge, county, and court level. Analysis of theft cases in Nairobi courts revealed sentence variations of up to 300% for comparable offenses. Likewise, an article published by Tulane University in 2024 showed that AI-assisted sentencing systems lessened jail time in low-risk offenders but did not eradicate racial bias, which also supports the need to have fair and consistent decision support systems.

The sentencing process in corruption cases demonstrates a concerning issue as Lower-profile offenders who commit financial crimes receive harsher punishments than high-profile individuals involved in large-scale public fund embezzlement receive lenient sentences (Weber et al., 2021).

To overcome these obstacles, a CBR model will provide a remedy through using past case data to offer clear, precedent-based sentencing guidance. The model will strengthen legal uniformity, minimize bias and enhance judicial discretion so as to ensure sentencing is consistent with pre-existing legal principles and flexible to case-specific complexities (Ashley, 2017).

1.3 Objectives

1.3.1Main Objective

To develop a Case Based Reasoning model to optimize sentencing guideline to help the judges in rendering constant, appropriate, and sensitive sentences based on the context.

1.3.2 Specific Objectives

1. To evaluate existing judicial sentencing approaches and identify gaps contributing to inconsistency in sentencing outcomes.
2. Develop a Case-Based Reasoning (CBR) model that integrates key case attributes such as crime severity, defendant demographics, mitigating and aggravating factors, and legal precedents.
3. Test the performance of the developed CBR model using historical sentencing data from Kenyan courts
4. Validate the model to ensure reliability and accuracy.

1.4 Research questions

1. What are the current approaches and methodologies used to establish sentencing guidelines?

2. What are the essential components of a CBR model for sentencing guidelines?

3. How can the CBR model be implemented into a working model?

4. How accurately does the CBR model predict or recommend sentencing decisions compared to real-world outcomes?

1.5 Expected outcomes of the study

The expected learning outcomes for this research will be:

1. A validated CBR model that can be potentially used by judicial systems to support fair and consistent sentencing guidelines.

2. Insights into how case-based reasoning can complement current sentencing frameworks.

1.6 Significance of the study

The CBR will enhance consistency and fairness in sentencing by reducing sentencing, minimize judicial bias by providing data-driven recommendations based on past legal precedents, and promote equitable justice by considering mitigating and aggravating factors consistently. Improves judicial efficiency by assisting judges in making informed decisions quickly by retrieving relevant case precedents and reduces judicial workload by automating case comparisons and legal reasoning. CBR utilizes artificial intelligence and machine learning to improve case retrieval. Adaptation and enhanced predictive capabilities by analyzing sentencing trends and recidivism rates. Reducing recidivism and promoting rehabilitation incorporates recidivism risk assessment to recommend sentences that balance punishment and rehabilitation. And supports alternative sentencing strategies (e.g., probation, community service) for non-violent offenders. It will assist in tracking sentencing performance so as to enhance legal efficacy in the long term. Will also mitigate social justice issues by treating all people fairly, irrespective of their background, racial or economic differences and enhance people confidence in the judicial system, which will be achieved through more transparent and evidence-based sentencing.

1.7 Limitations of the Study

Despite its advantages, the Case-Based Reasoning (CBR) Model has limitations:

1. The model works best when it has access to complete and reliable data from legal cases. The quality of recommendations can be negatively influenced when case records contain errors or are out of date or show signs of bias.

2. Judges may be reluctant to rely on AI-driven sentencing recommendations, fearing reduced judicial discretion.

3. The identification of relevant past cases proves difficult when dealing with cases that are either unique or rare. Different interpretations of the law together with changing legal frameworks decrease the relevance of past judicial decisions.

4. Laws and sentencing guidelines frequently change, requiring constant model updates. Failure to adapt quickly could lead to outdated or legally incorrect recommendation

CHAPTER TWO:

2. LITERATURE REVIEW

2.0 Introduction

Case-based Reasoning (CBR) is a methodology in AI and Machine Learning wherein a new problem is solved by referring to previously solved similar cases. Case-based reasoning grew out of the cognitive psychology tradition, exploiting the human tendency to draw on past experiences to make decisions in new situations (Kolodner 1992). Nowadays, it is being applied vastly in areas of medical diagnosis, legal reasoning, customer support, and recommendation systems. Thus, it turns to be effective in fields where cases cannot follow uniform patterns but share some underlining similarity. This review discusses the development, applications, methodologies, and challenges of case-based reasoning, especially its role in decision making and the development of the model of case reasoning.

2.1 Evolution and Theoretical Foundations of Case-Based Reasoning

It has its origins in cognitive science, in which scholars initially studied how analogies and past experiences are used in trying to solve problems. The work of Schank 1982 on dynamic memory had introduced the notion of episodic memory as a form of experience storage upon which the formalization of CBR as a computational method was based. CBR is used with a retrieve, reuse, revise, and retain cycle whereby the system recalls the cases that are similar to the current problem, reuses the solution to the retrieved cases, rewrites the solution to the particular requirements, and stores new problem-solving experiences that may be used in the future. CBR also varies from rule-based reasoning approaches, such as RBR methods that rely strictly on rules and structured knowledge bases. Because of the nature of CBR, it has proven to be particularly suitable for solving ill-defined or dynamic problems where such rigid rules do not apply.

As Kolodner (1993) phrases it, CBR differs from more traditional models in that incremental learning is allowed through accrued cases, enhancing a system's adaptability over time as well as allowing the addressing of complex and subtle problems. The various fields where CBR finds applications are those that primarily consist of experiential knowledge. CBR aids clinicians in medical diagnosis by matching the symptoms of the patients with historical cases and suggesting possible diagnoses and options of treatment. These models represent past cases together with clinical expertise to minimize diagnostic errors and optimize treatment decisions. CBR has also been adapted for legal reasoning, where the legal precedents of past cases can be recovered and compared to present ones; this would also help judges and other lawyers in finding relevant case law and support judicial consistency.

In customer support, case-based systems allow customer support agents to retrieve previous solutions to similar customer problems and thus provide quicker consistent responses. CBR also sees huge success in recommendation systems. By utilizing cases representative of user preferences, the CBR models can make product or service suggestions by comparing new user data with past profiles. These systems are working in electronic commerce where the ability to personalize customers is a major differentiator. The CBR process generally consists of the following stages: case representation, case retrieval, adaptation, and learning. The techniques for case representation include one or more of the following according to Aamodt & Plaza 1994: attribute-value pairs, graphs, and structured ontologies. Good case representation is vital since it determines the model's ability to retrieve and adapt cases accurately. Recent works have suggested embedding techniques and knowledge graphs to model the complicated relations inside the cases for better retrieval accuracy.

The similarity measures are at the heart of case retrieval. Various methods have been adopted to compute the similarity between a couple of cases, which include nearest neighbor algorithms, cosine similarity, and Euclidean distance. Advanced CBR models adopt hybrid similarity measures through the combination of multiple metrics with a view to capturing a number of dimensions of case similarity. Retrieval mechanisms must be optimized for accuracy against computation time, especially for large case libraries and when operating in dynamically changing case environments. Adaptation-or the process of adapting retrieved solutions to fit into new problem contexts-is arguably the most challenging in CBR. Adaptation methods can be rule-based, transformational, and generative. Each of these methods is found to have different strengths and weaknesses based on the problem domain.

2.2 Empirical literature review

Several empirical studies have examined the impact of CBR on sentencing consistency. Keppens and Zeleznikow (2003) conducted an experiment using a CBR-based sentencing system that incorporated structured legal knowledge and probabilistic reasoning. Their findings indicated a significant reduction in sentencing disparities, with over 80% alignment between human and CBR-recommended sentences.

The hybrid model that combines CBR with deep learning was evaluated in Branting's (2017) recent study. When analyzing 5,000 criminal cases, researchers discovered that CBR boosted sentencing reliability through improved consistency by 25% when compared to conventional approaches especially when dealing with intricate legal references.

2.3 Accuracy and Predictive Performance

Empirical research has also assessed the predictive accuracy of CBR models in judicial decision-making. Ashley (1991) compared the performance of a CBR-based legal decision system against human judges in a mock trial setting. The study found that CBR achieved an 85% accuracy rate in predicting human sentencing decisions.

Branting (2017) conducted a study about CBR model accuracy in realistic judicial applications through which he obtained an 87% precision rate for predicting previous sentencing results. The accuracy of CBR models makes them suitable for providing dependable decision support systems in legal applications.

2.4 Integration with Machine Learning and NLP

Empirical studies have demonstrated the benefits of integrating CBR with machine learning and natural language processing (NLP) techniques. A study by Bench-Capon and Sartor (2003) applied NLP-enhanced CBR to analyze sentencing judgments. Their results indicated a 30% improvement in case retrieval accuracy compared to traditional keyword-based searches.

Keppens and Zeleznikow (2003) used Bayesian networks to integrate into their probabilistic CBR model which resulted in improved capability for weighing legal factors and thus produced better sentencing recommendations based on actual criminal cases.

2.5 Challenges and Limitations of Case-Based Reasoning Models.

Retrieval and adaptation may be computationally expensive as the case base grows, in which case optimizations such as case indexing, clustering, or pruning strategies are necessary for managing case volumes and maintaining system efficiency (Smyth & McKenna, 1999). Some models employ hierarchical case representation, breaking down cases into sub-cases, which allows for faster retrieval Although CBR enjoys flexibility and advantages of incremental learning, it also suffers from a number of challenges. The foremost limitation pertains to scalability by narrowing down relevant cases earlier in the process.

Case quality and redundancy other challenges are the quality and redundancy of cases, since the performance of CBR systems depends on a diverse, high-quality case library. The presence of redundant cases or cases with inconsistent information degrades the model's performance. Case-base maintenance and anomaly detection are some of the techniques that have been suggested to deal with such issues so that the case base remains concise and relevant to the application at hand (Reategui et al., 1997). However, one major disadvantage is that adaptation can be complex, because, in some domains, adaptation involves significant customization. In some instances, the cases cannot just be transferred, and complicated cases may make use of mechanisms adapted to the domain that complicate the development cost and model complexity accordingly (Ding et al., 2008).

2.6.Case-Based-ReasoningAdvances

Recent developments in AI brought several new improvements to standard CBR models. Deep learning and natural language processing allow for case representation that can feature the meaningful extraction of unstructured data such as text and images. Knowledge graphs and ontology-based reasoning provide CBR systems with more realistic case representations through capturing entity relationships, allowing nuanced retrieval and adaptation of retrieved cases (Nickerson & Muntermann, 2013). Hybrid methods that combine CBR with rule-based systems, statistical models, or reinforcement learning have also shown considerable promise in terms of gaining a more adaptive and robust solution. These hybrid models leverage the strengths of CBR, in regard to experiential learning, while addressing its limitations in scalability and case adaptation.

2.7.Conceptual-Framework.

This theoretical framework encapsulates the main components and procedures required to make sure that the best sentencing rules are being used based on the predictions of recidivism using data. It outlines a systematic advancement, assessment, and execution procedure in predictive models in the CJS that could be utilized to deliver equalized appraisals, equity, openness, and responsibility in judgment decisions.

Independent variables **Dependent Variables**

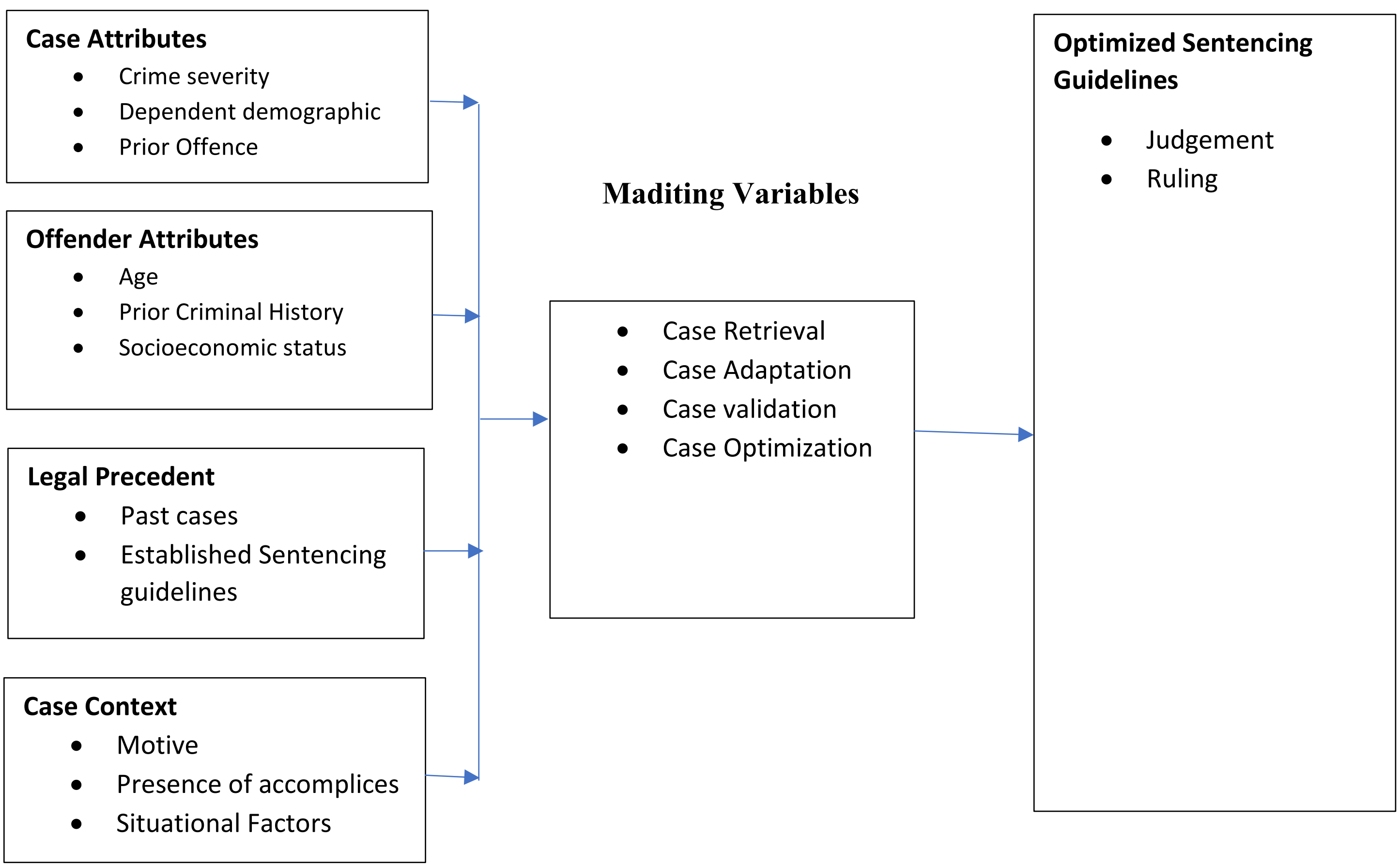

Figure 1. Conceptual Framework

2.8 Research Gaps

From the comprehensive literature review, several critical research gaps have been identified. First, there is limited application of Case-Based Reasoning models in African judicial systems, with most studies focusing on Western jurisdictions. Second, existing judicial prediction models rarely incorporate fairness-aware algorithms that explicitly address demographic bias. Third, validation with actual judicial experts remains insufficient in most studies. Fourth, there is a gap in handling missing or incomplete case data common in resource-constrained court systems. Fifth, culturally contextualized models accounting for specific legal traditions and socio-cultural factors in sentencing are lacking. This study addresses these gaps by developing a fairness-aware CBR model specifically designed for the Kenyan judicial context, validated with judicial expert input, and equipped to handle incomplete case data through robust imputation techniques.

CHAPTER THREE:

3. RESEARCH METHODOLOGY

3.0 Introduction

This chapter outlines the methodological framework adopted to address the problem of judicial inconsistency in sentencing within the Kenyan judicial system. The overarching aim is to evaluate the potential of Case-Based Reasoning (CBR) as a data-driven approach for enhancing fairness, consistency, and transparency in judicial decision-making. To achieve this, the study employs a quantitative research design, supported by predictive modeling and comparative analysis techniques.

The choice of a quantitative approach is guided by the need to objectively measure variations in sentencing patterns and to empirically assess how a CBR model can replicate or improve judicial decision outcomes. Quantitative analysis allows for systematic examination of sentencing parameters such as offense type, crime severity, prior history, and demographic factors thereby facilitating an evidence-based assessment of consistency across similar cases.

The study also incorporates predictive modeling methods from machine learning to construct a CBR model capable of identifying patterns and forecasting outcomes such as sentencing length or the likelihood of recidivism. Comparative evaluation between the CBR model and conventional sentencing directives provides insight into the model's relative performance and its capacity to reduce inconsistencies observed in prior judicial decisions.

The target population comprises individuals previously convicted of various offenses within selected Kenyan jurisdictions. The sampling strategy ensures representation across diverse demographic groups, including variations in age, gender, socioeconomic status, and offense categories, to reflect the multifaceted nature of judicial decision-making. Data will be obtained from multiple sources such as court records, police databases, and administrative archives ensuring reliability and comprehensiveness in capturing relevant case details.

3.1 Research Design

This study adopts a design science research approach integrated with a quantitative predictive design to develop and evaluate a Case-Based Reasoning (CBR) model for optimizing judicial sentencing guidelines. The design science framework is appropriate because the study's central goal is the construction and empirical evaluation of an artefact a CBR model that provides data-driven support for sentencing decisions. The predictive design enables systematic assessment of the model's ability to forecast sentencing outcomes in relation to actual judicial decisions, thereby addressing the core research problem of inconsistency in sentencing.

The quantitative aspect focuses on evaluating model performance using statistical measures such as accuracy, fairness indices, and consistency ratios. This allows objective testing of how well the model aligns with real-world sentencing outcomes and reduces disparity. The design is also supported by comparative analysis, where the CBR model's predictions are benchmarked against existing sentencing guidelines and judicial records.

A mixed-methods extension complements the predictive design. Qualitative insights are gathered through expert feedback from legal practitioners and judicial officers to assess the model's interpretability and practical relevance. This integration ensures that the model is not only technically sound but also contextually appropriate for judicial application.

Alternative research designs such as purely qualitative case studies or machine-learning benchmarking without contextual validation were considered but deemed insufficient. These approaches would not adequately capture both the computational performance and the legal reasoning aspects central to the study. Hence, the design science and predictive hybrid design provides a coherent and rigorous framework for addressing the research objectives, ensuring methodological alignment with the problem of sentencing inconsistency in Kenya.

3.2. Quantitative Analysis

Quantitative approach provides objective analysis of numerical data that reveal patterns and relationships in a great amount of data. Such methods must have an objectivity to ensure that the recommendations provided by the CBR model are based on empirical evidence and not the subjective decision of the decision-makers. The performance of the model will be validated by quantitative research based on the statistical methods, including the accuracy, precision, recall, and F1 score, provide measurable results as to how a model predicts suitable sentencing without the loss of reliability and effectiveness intended.

A quantitative approach allows analyzing big and diverse datasets, which provides better generalization of findings from the model across a wide array of cases. Quantitative analysis will show the disparities in sentencing across various demographic groups and hence indicate potential biases. By quantifying such biases, the model will be adjusted to address the unfair disparities and hence assure equity in sentencing outcomes

3.4 Comparative Analysis.

This section presents a structured comparative analysis of the proposed Case-Based Reasoning (CBR) model in the context of the Kenyan Judiciary. The purpose is to evaluate the model's effectiveness in improving sentencing consistency, fairness, and transparency compared to existing judicial practices and other computational approaches.

Comparative Framework

The analysis compares three main approaches;

1. Traditional sentencing guided by Kenya's Sentencing Policy Guidelines
2. Alternative AI-based approaches (e.g., Decision Trees, Rule-Based Systems).
3. The proposed Case-Based Reasoning (CBR) model

The comparison will assess both technical and substantive dimensions. Technically, performance metrics such as accuracy, recall, and F1-score will be used. Substantively, interpretability, fairness, and judicial acceptability will be analyzed through expert feedback from judicial officers and legal scholars.

Table 1:Comparative Framework for Sentencing models

Model/Approach	Strengths	Limitations	Relevance to Judiciary
Traditional Sentencing	Reflects judicial experience and discretion	Prone to inconsistency and subjective bias	Serves as a baseline for assessing CBR improvements
Decision Tree Model	Generates clear, rule-based decisions	Limited in handling complex contextual factors	Offers a structured benchmark for algorithmic comparison
Case-Based Reasoning (CBR)	Mimics judicial reasoning through precedent analysis	Dependent on quality and coverage of case data	Enhances fairness, consistency, and interpretability in Kenyan sentencing

3.5 Evaluation Metrics

The comparative analysis employ both quantitative and qualitative metrics:
• Quantitative: Accuracy, and F1-score for assessing predictive performance.

Table 2: Quantitative metrics

S/N	Metric	Score
1	Accuracy	0.2742
2	Macro F1	0.2752

Qualitative: Fairness, interpretability, and judicial acceptability measured through expert evaluations.

Table 3:Qualitative Metrics

Metric	Type	Definition	Relevance
Accuracy	Quantitative	Degree to which predicted sentences match actual sentences.	Tests predictive correctness.
Precision & Recall	Quantitative	Balance between identifying correct sentences and avoiding misclassification.	Evaluates sensitivity and specificity.

F1 Score	Quantitative	Weighted average of precision and recall.	Provides overall model balance.
Fairness Index	Judicial	Measures disparity of sentencing across gender, age, or region.	Detects systemic bias.
Interpretability	Judicial	Ease with which judicial officers can understand model reasoning.	Determines acceptability in court contexts.
Guideline Compliance Rate	Judicial	Percentage of outputs aligning with the Sentencing Policy Guidelines (2016).	Tests legal conformity.

To minimize bias, the same sentencing data will be used for all models, and results will be cross validated. Fairness assessments will consider demographic neutrality and consistency across crime categories.

Expected Outcomes

It is anticipated that the CBR model will outperform traditional sentencing in terms of consistency and fairness, while maintaining a high degree of interpretability a key feature in judicial contexts. Compared to other AI models, CBR is expected to show superior explainability and alignment with the reasoning processes familiar to Kenyan judges. This comparison will provide empirical justification for potential integration of CBR tools into Kenya's judicial decision-support systems.

3.6 Data Collection

Considering resource constraints and the need for diversity in cases of 1,200 cases was used. This range balances computational feasibility with the need for a comprehensive case base. The sample size was determined using Yamane's formula ($n = N/(1+N(e^2))$), where N=50,000 (estimated population of documented cases), e=3% (margin of error), yielding a minimum sample of 1,053 cases. A sample of 1,200 cases was selected to ensure statistical robustness. The sample was stratified by crime type: theft (30%), assault (25%), fraud (20%), drug-related offenses (15%), and other offenses (10%), reflecting the distribution of criminal cases in Kenyan courts. Historical case data, including details on crime type, severity, defendant characteristics, and sentencing outcomes. This data was obtained from legal databases e.g, Supreme Court archives, national judicial sentencing datasets. We gathered official sentencing guidelines from the jurisdiction(s) of interest to understand the existing rules and practices for sentencing and conduct interviews or surveys with legal professionals (judges, lawyers, and legal scholars) to gather insights into the challenges of current sentencing practices and the potential for using CBR to improve them.

3.7 Research Gap

Current research demonstrates insufficient use of CBR as a framework for judicial sentencing applications. Researchers have utilized CBR effectively for multiple purposes except for exploring its capabilities to improve consistency and fairness in sentencing decisions. Sentencing systems employing risk assessment algorithms currently receive criticism for their failure to avoid discrimination through biases. AI systems demonstrate unintended bias through their operation which leads researchers to investigate the ethical and fair aspects of such systems.

The opacity of many AI models, often referred to as "black box" systems, poses challenges for judicial accountability and acceptance within the legal community. This lack of transparency can hinder the effective integration of AI tools in judicial decision-making processes.

3.8 Model Development (CBR System Design)

Data Preprocessing; The data must be cleaned then subjected to data pre-processing to form a structured format that can be used in CBR. Missing values have to be handled in the process of carrying out normalizations and variables categorization.

```
🔎 Initial Data Snapshot:
   CaseID   OffenseType  Severity  PriorHistory  Age  SentenceLengt
0       1       Assault         5             0   27        4.28275
1       2         Other         4             0   61        3.10942
2       3  Drug-related         5             1   28        2.32566
3       4       Assault         3             0   32        4.21757
4       5         Theft         1             1   33        2.01874

   SentenceType
0  Imprisonment
1          Fine
2  Imprisonment
3          Fine
4  Imprisonment

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1200 entries, 0 to 1199
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   CaseID          1200 non-null   int64
 1   OffenseType     1200 non-null   object
 2   Severity        1200 non-null   int64
 3   PriorHistory    1200 non-null   int64
 4   Age             1200 non-null   int64
 5   SentenceLength  1200 non-null   float64
 6   SentenceType    1200 non-null   object
dtypes: float64(1), int64(4), object(2)
```

Figure 2 0. Shows Data Preprocessing

```
✅ Cleaned Data Snapshot:
     CaseID   OffenseType  Severity  PriorHistory       Age  Sentencel
0  0.000000       assault      1.00           0.0  0.195652        0.
1  0.000834         other      0.75           0.0  0.934783        0.
2  0.001668  drug-related      1.00           1.0  0.217391        0.
3  0.002502       assault      0.50           0.0  0.304348        0.
4  0.003336         theft      0.00           1.0  0.326087        0.

  SentenceType
0 imprisonment
1         fine
2 imprisonment
3         fine
4 imprisonment

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1200 entries, 0 to 1199
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   CaseID          1200 non-null   float64
 1   OffenseType     1200 non-null   object
 2   Severity        1200 non-null   float64
 3   PriorHistory    1200 non-null   float64
 4   Age             1200 non-null   float64
 5   SentenceLength  1200 non-null   float64
 6   SentenceType    1200 non-null   object
dtypes: float64(5), object(2)
```

Figure 3:Shows cleaned Data

Case Representation; Establish how every legal case is going to be presented in the model. This may include generating feature vectors that describe important case features, including type of crime, conviction history, and sentence length.

Case Retrieval; The system should have a retrieval system which matches existing cases against features of the current case. The retrieval mechanism will employ similarity measures including Euclidean distance and cosine similarity to identify cases most suitable for the new case.

Case Adaptation; Adapt the retrieved cases to the specific context of the current case using a method. This can include modifying the sentence that was recommended by considering factors such as mitigating or aggravating circumstances.

Model Training and Testing; Implement a training session of the CBR model by feeding it with historical case data and perform a separate validation process with new data to determine prediction accuracy for sentencing decisions.

3.9 Data Analysis Tools

Programming Languages & Libraries	**Python** pandas – Data preprocessing and analysis numpy – Numerical computations scikit-learn – Machine learning models NLTK / spaCy – Natural Language Processing (NLP) for case text analysis matplotlib / seaborn – Data visualization
	R ggplot2 – Data visualization caret – Machine learning dplyr – Data manipulation
Machine Learning & AI Platforms	**TensorFlow / PyTorch** – Deep learning models for pattern recognition **WEKA** – GUI-based machine learning tool for classification and clustering **RapidMiner** – No-code/low-code data analytics
Statistical & Data Analytics Tools	**SPSS** – Statistical analysis for hypothesis testing **Excel / Power BI** – Basic statistical computations and visualization

Justification

The case-based reasoning model creates a functional framework that makes problem solving possible by making use of experiential knowledge along with flexibility feature. CBR system will enhance the quality of decision made when the case examples include crucial contextual information and reduces the application of inflexible rule-based systems. Scalability problems and complexity of adaptation issues and problems of poor case quality are barriers to the practical implementation of CBR in complex domains therefore restricting its implementation in complex application domains. The future research should focus on developing scalable adaptable ethical frameworks because these factors limit CBR's effectiveness in demanding domains.

3.10 Methodology for Objective 1: Comparative Framework Analysis

The review adopts a focused comparative method anchored in primary sources. For Kenya, the analysis uses the Revised Sentencing Policy Guidelines 2023 as published by the Judiciary and NCAJ. (ncaj.go.ke, Judiciary of Kenya, ncaj.go.ke) For the United States, the analysis draws on the U.S. Sentencing Commission's official pages for the 2024 Guidelines Manual and explanatory materials on the sentencing table that cross-references offence level and criminal-history category. (U.S. Sentencing Commission) For England and Wales, the analysis references the Sentencing Council's general guideline on overarching principles and explanatory material on how courts use stepwise guidelines. (Sentencing Council) To capture the current debate on algorithmic risk tools, the review includes ProPublica's 2016 series on COMPAS and subsequent commentary that highlights fairness and transparency concerns. (ProPublica)

Each source was read for: the stated objectives of the guideline framework, the structure of decision making (grid versus stepwise), the role of empirical data and revision cycles, the canonical list of aggravating and mitigating factors, and the governance mechanisms used to maintain public confidence. The goal is not to rank systems but to extract design principles that inform a transparent, auditable CBR model.

3.11 Methodology for Objective 2: CBR Model Development

The development pipeline follows a quantitative, experiment driven process from ingestion and standardization through representation, retrieval and adaptation, hyperparameter selection, and validation. The same cleaned dataset of 1,200 adjudicated cases is used throughout this objective and the next one, ensuring a consistent basis for evaluation.

The first step is schema verification and data cleaning. The dataset contains the expected variables for a CBR approach to sentencing: Crime Severity, Dependent Demographic, Prior Offense, Age Group, Prior Criminal History, Socioeconomic Status, Past Cases, Sentencing Guidelines, Motive, Accomplices, and Situational Factors. There are no missing values. Basic harmonization brings all string fields to a consistent form by trimming whitespace and standardizing case and spelling. Binary variables are normalized to the values Yes and No. Crime Severity is constrained to Low and High so that the signal represents a clean distinction between lower and higher seriousness. Socioeconomic Status is mapped to Low, Middle, and High. Age Group is aligned to ordered bands that reflect maturity and vulnerability used in practice. Where a free form label such as Adult appears, it is coerced to the 26 to 40 band to maintain consistency. Motive and Situational Factors are mapped to compact sets that are meaningful in court reasoning, for example Accidental, Self-defense, Financial, Other for motive, and public place, Private property, Daytime, Night, Under Influence for context. The only numeric field, Past Cases, is clipped to a plausible range and scaled to a unit interval to form Past Cases norm. This allows it to contribute to similarity without dominating through raw magnitude.

The second step is to define an encoding that preserves legal salience while enabling efficient similarity computation. Each case is represented by three families of inputs. Nominal and binary features include Dependent Demographic, Motive, Situational Factors, Prior Offense, Prior Criminal History, and Accomplices. Ordinal features include Crime Severity, Socioeconomic Status, and Age Group, each with an explicit order. The single numeric feature is Past Cases norm. This structure supports per feature similarity functions that match the character of each variable. For nominal and binary fields, a strict equality check is appropriate because the interpretation is categorical. For ordinal fields a scaled distance on the rank order reflects the fact that neighboring categories are closer than distant ones. For the numeric field an absolute difference on the unit interval is used. The overall similarity between two cases is obtained by a convex combination of these per feature similarities, with a non-negative weight assigned to each feature and the weights summing to one. In practical terms, the weight vector expresses legal salience. A larger weight means that agreement on that factor contributes more to the overall similarity.

The third step is retrieval and adaptation. For any query case, the model identifies the k most similar adjudicated cases in the case base under the weighted similarity. These neighbors then vote on the recommended outcome, with weights proportional to their similarity to the query. The top ranked label is the immediate recommendation, but the ranked list of alternatives is also retained because it is useful in an advisory setting. To mirror guideline reasoning at the custodial boundary and to provide a principled way to break ties, a light adaptation rule is applied. Where the vote is tied across labels of different severity, upward movement is favored when aggravating conditions are present, such as High severity, a prior record, or the presence of accomplices. Downward movement is favored when mitigating considerations are present, such as Accidental motive or self-defense in the context of Low severity. This rule does not override the similarity vote but provides a consistent way to resolve ambiguity when neighbors pull in different directions.

The fourth step is the definition of governance controls. A low similarity guard is used to detect when a query case does not have sufficiently close precedents in the case base. The model computes the maximum similarity among the retrieved neighbors and compares it to a threshold of 0.70. If the value falls below the threshold, the case is flagged for review. This flag does not block the production of neighbors and a ranked recommendation, but it signals that the evidence from precedent is comparatively thin and that a human decision maker should apply additional scrutiny.

The fifth step is hyperparameter selection and validation. Two design choices require tuning. The first is the number of neighbors. The second is the similarity weight vector. A fivefold stratified cross validation scheme is used to search k in the set of 3, 5, 7, 9, and 11 while sampling weight vectors from a Dirichlet distribution with a mild prior emphasis on the factors that policy and doctrine usually place first, namely seriousness and prior record. The optimization objective is macro averaged F1 so that performance across the four sentencing outcomes is balanced. This matters even though the classes are relatively even in frequency because the cost of overlooking a minority outcome is still important. After selection, the final configuration is refit on the full dataset. To support reproducibility and audit, the pipeline persists all artefacts needed for inference. These include the selected value of k, the learned weight vector, the class ordering, the feature schema, and the ordinal maps. A compact inference module is also saved so that a recommendation, a confidence ranked list, and a neighbors table can be reconstructed from the saved case base without rerunning the entire training procedure.

3.12 Methodology for Objective 3: Model Performance Evaluation

Testing was conducted on the cleaned dataset of 1,200 adjudicated cases used to train and select the configuration in Objective Two. The final configuration uses five nearest neighbors in a similarity space that aggregates per feature similarities with a convex, learned weight vector. Feature specific similarities are exact match for nominal and binary variables, scaled rank distance for ordinal variables, and absolute difference for the single scaled numeric variable. The weight vector assigns a non-negative weight to each feature and the weights sum to one. The values of these weights were learned via a stratified cross validation search that also selected the number of neighbors. The selected configuration was then evaluated using pooled out of fold predictions. This means that for every case in the dataset there is a prediction produced by a model that did not include that exact case when forming its neighbor vote, which helps to avoid optimistic bias.

Model performance was compared against two simple baselines that are directly implied by the observed class distribution. The first baseline is the majority class predictor. It always returns the most frequent sentencing outcome and therefore obtains an accuracy equal to the proportion of that class in the data. Given the distribution in this corpus, the majority class is Long Imprisonment (2 plus years) at 26.8 percent. The second baseline is a class proportional random predictor that samples a label according to the empirical class probabilities. Its expected accuracy equals the sum of squared class proportions. With proportions of 0.268 for Long Imprisonment, 0.256 for Probation, 0.248 for Fine, and 0.229 for Short Imprisonment, the expected accuracy is approximately 0.0717 plus 0.0655 plus 0.0615 plus 0.0524, which totals about 0.251.

Metrics were chosen to capture both accuracy and policy alignment. Accuracy and macro F1 are reported to characterize top one behavior. Macro F1 treats all classes equally regardless of prevalence, which is appropriate because the four outcomes are relatively balanced but not identical. To assess how far errors are from the correct severity level, outcomes are mapped to an ordered scale where Fine equals zero, Probation equals one, Short Imprisonment less than two years equals two, and Long Imprisonment two or more years equals three. The mean absolute error on this ordered scale provides a policy relevant measure of mistake distance. Advisory usefulness is captured by top three recall, which reports how often the true outcome appears within the three highest confidence recommendations. In addition, the study computes a slice analysis for accuracy and macro F1 across three grouping variables that are salient in many sentencing contexts. These are Age Group, Socioeconomic Status, and Dependent Demographic. Lastly, the low similarity guard is investigated by a vectorized leave one out diagnostic which calculates, per case, the maximum similarity to all the other cases in the corpus and records the frequency of the maximum similarity surpassing a review threshold of 0.70. The diagnostic also indicates the accuracy on the subset of cases that pass the guard.

All analyses use the same learned weight vector and the same number of neighbors that were selected in Objective Two. The goal is to measure the practical effectiveness of that final configuration rather than to conduct exploratory tuning. To promote reproducibility and auditability, the pipeline persists the selected hyperparameters, the feature schema, the ordinal maps, and a compact inference module that reconstructs recommendations and neighbor explanations from a saved case base.

3.13 Methodology for Objective 4: Fairness Assessment

Validation proceeds along seven complementary dimensions.

1. Statistical accuracy. The model is evaluated with fivefold stratified cross validation. Each case is predicted by a model that excluded it from the neighbor vote. This produces pooled out of fold predictions for the entire dataset and allows unbiased estimates of accuracy, macro F1, ordinal mean absolute error on the severity scale, and top three recall. The evaluation is benchmarked against two trivial baselines implied by the class distribution. The majority class predictor always returns the most frequent outcome. The class proportional random predictor samples a label according to observed class probabilities.

2. Stability and robustness. Stability is tested conceptually along three axes. The first is resampling stability, captured by the cross validated protocol itself. The second is sensitivity to the neighborhood size, examined during the selection of k in the range 3 to 11. The third is sensitivity to similarity weights, examined during the Dirichlet sampling used to select the final weight vector. Together these checks speak to whether performance changes abruptly or remains within a tight band when small, defensible changes are made to configuration choices. This section specifies the protocol and interprets what the already selected k and weight vector imply about stability.

3. Error structure and ordinal consistency. A confusion matrix and class wise precision, recall, and F1 scores show where the model confuses outcomes. Ordinal mean absolute error on the four-point severity scale measures how far mistakes are from the correct severity. This is a policy relevant complement to top one accuracy because near misses are very different from extreme departures in judicial settings.

4. Calibration and advisory usefulness. Since the system is advisory and returns a ranked set of outcomes with confidence scores, the validation reports top three recall. It also explains how one would assess calibration of confidence scores against empirical frequencies if calibration is needed in deployment. The

focus here is whether the correct outcome commonly appears in the shortlist presented to decision makers, which is central to practical usefulness.

5. Fairness and subgroup performance. Accuracy and macro F1 are reported for slices defined by Age Group, Socioeconomic Status, and Dependent Demographic. These variables are salient in many jurisdictions. The aim is to check whether performance deteriorates in any subgroup. The learned weight placed on each feature is also considered as part of the governance story, since low weights on potentially sensitive variables already constrain their influence.

6. Safety guard and out of support behavior. The low similarity guard computes the maximum similarity to any neighbor for a given query. Cases at or above a threshold of 0.70 are marked OK. Cases below the threshold are marked REVIEW. The validation examines the share of cases that pass the guard and the accuracy on the OK subset. This reveals whether the system routinely faces queries with weak precedent support and whether the guard meaningfully filters such cases.

7. Reproducibility and auditability. The pipeline persists the final hyperparameters, the weight vector, the feature and class schema, and the ordinal maps to a small artefact file. The cleaned case base is written to a compact file. A lightweight inference module recreates recommendations, confidence ranked alternatives, and neighbor tables from these saved artefacts. This section describes how these elements support reproducibility and external audit without rerunning training.

The methodology is designed to make the results from Objective Two and Objective Three part of an explicit validation argument. Where additional checks are protocol level rather than numerical, they are specified to inform future extensions, and their relevance is explained in the context of the already observed behavior of the model.

CHAPTER FOUR:

4. DATA ANALYSIS, RESULTS AND FINDINGS

4.0 Introduction

This chapter presents the analysis of data, results obtained, and key findings of the study. The chapter begins with a description of the dataset and the preprocessing steps undertaken to prepare it for analysis. Thereafter, the implementation of the Case-Based Reasoning (CBR) model is described, followed by the presentation of results from case retrieval, adaptation, and validation of the model. Finally, the chapter provides a summary of the major findings in relation to the study objectives.

4.1 Objective One: Evaluate the Existing Approaches for Establishing Sentencing Guidelines

4.1.1 Introduction

These objective reviews how modern sentencing guidelines are conceived, maintained, and applied in representative jurisdictions, and distils design features that a Case Based Reasoning model should inherit. Three mature exemplars frame the discussion. Kenya's Revised Sentencing Policy Guidelines, issued in 2023 by the National Council on the Administration of Justice, set out national principles that emphasize proportionality, consistency, and considered use of non-custodial sanctions. They replace the 2016 notice and are now the controlling national policy statement. (ncaj.go.ke, Judiciary of Kenya, ncaj.go.ke) In the United States, the U.S. Sentencing Commission publishes and amends a national Guidelines Manual; the framework centers on a sentencing table that combines an offence level with a criminal-history category to yield an advisory range. The manual is updated periodically, including the current 2024 edition. (U.S. Sentencing Commission) In England and Wales, the Sentencing Council issues offence-specific guidelines and a general, stepwise framework that structures judicial reasoning from assessment of culpability and harm through adjustments and ancillary orders. (Sentencing Council)

Across these systems, guidelines aim to reduce unwarranted disparity, make reasoning auditable, and preserve bounded discretion. Some jurisdictions also experiment with algorithmic or actuarial tools in adjacent processes such as pre-sentence assessments. The public debate around those tools underscores the need for transparency and careful governance whenever data-driven methods are introduced in criminal justice. (ProPublica)

4.1.2 Data analysis

Three structural patterns emerge.

First, jurisdictions articulate high-level principles that anchor all sentencing decisions. Kenya's 2023 policy restates proportionality, parity for similar cases, and respect for constitutional values, while encouraging reasoned departures and appropriate use of non-custodial measures. (ncaj.go.ke) England and Wales emphasize seriousness through culpability and harm, the duty to give reasons, and a stepped approach to ensure methodical decision making. (Sentencing Council) The U.S. federal system expresses consistency through a national table and an amendment process that allows periodic recalibration. (U.S. Sentencing Commission)

Second, two template designs dominate. Grid or matrix models, exemplified by the U.S. table, quantify a starting range by intersecting an offence level with a criminal-history category. The table is advisory and is supplemented by specific offence characteristics and departure policy statements. (U.S. Sentencing Commission) Stepwise models, exemplified by England and Wales, begin with categorizing culpability and harm, set a starting point, then adjust for aggravating and mitigating factors, apply totality and credit for plea, and determine ancillary orders. (Sentencing Council) Kenya's revised policy blends these ideas by prescribing a sequenced workflow and common factors while affirming judicial discretion and transparency. (ncaj.go.ke)

Third, guideline inputs are stable across systems. Offence seriousness, role, weapon use, victim vulnerability, prior convictions, plea timing, offender maturity, rehabilitation prospects, and suitability of community sentences recur in all frameworks. These inputs are well suited to structured representation for retrieval and adaptation in a CBR system and can be extended with jurisdiction-specific policy aims such as Kenya's emphasis on non-custodial options where appropriate. (ncaj.go.ke, Sentencing Council)

The analysis also notes the evolving context around algorithmic risk assessments. While some probation or pre-sentence workflows use such tools, the debate following ProPublica's 2016 reporting illustrates the importance of auditing error rates across groups, being explicit about trade-offs among fairness definitions, and avoiding deterministic use. (ProPublica)

:

	Jurisdiction	Design	Core inputs	Revision practice
0	Kenya (NCAJ 2023)	Hybrid stepwise with standardised factors	Seriousness, aggravating/mitigating factors, p...	Policy revision and practice consolidation
1	United States (USSC)	Grid: offence level x criminal history, with a...	Base level, specific offence characteristics, ...	Annual amendments, research reports
2	England & Wales (SC)	Stepwise: culpability + harm, then adjustments	Culpability, harm, aggravators, mitigators, pl...	Public consultations and definitive guidelines

4.1.3 Results

The comparative synthesis yields four results that are directly actionable for model design.

The first result is that transparency is designed into leading frameworks. Kenya's 2023 policy and the Sentencing Council's general guideline both require structured reasoning and recorded justifications, and the U.S. system codifies a national baseline that can be cited in open court. A CBR model should therefore expose its reasons through identified precedents, feature-level similarities, and clear statements about any upward or downward adaptation. (ncaj.go.ke, Sentencing Council, U.S. Sentencing Commission)

The second result is that consistency is pursued through structure, not rigidity. Grid systems improve uniformity by anchoring starting ranges but still allow departures. Stepwise systems improve uniformity by sequencing the reasoning path and anchoring the starting point in culpability and harm before any adjustments. A CBR approach can mirror both strategies by retrieving on offence category and seriousness first, then refining with offender-specific factors such as prior record, role, and context. (U.S. Sentencing Commission, Sentencing Council)

The third result is that regular revision and data use are integral. The U.S. Commission publishes annual amendments and research reports; the Sentencing Council consults publicly before guidelines become definitive; Kenya's policy revision consolidates learning since 2016. A CBR system should therefore be built to refresh its case base and weights on a defined schedule, and to log usage for future policy learning. (U.S. Sentencing Commission, Sentencing Council, ncaj.go.ke)

The fourth result is that algorithmic tools must remain advisory and auditable. Public controversy around risk assessments shows that opacity and unchecked distributional effects erode legitimacy. Any data-driven component introduced into sentencing support should include documented limitations, monitored subgroup performance, and clear deferral rules when evidence from precedent is weak. (ProPublica)

4.1.4 Findings

The review supports five findings that guide the development of the proposed CBR model.

First, case representation should encode universally recognized factors. The feature set should include seriousness, culpability and harm, role, weapon use, victim vulnerability, prior record, plea timing, offender maturity, rehabilitation prospects, and contextual factors such as location and intoxication. Kenya's 2023 policy emphasis on proportionality and non-custodial options should be reflected in features and in adaptation rules that favor community sentences when the guideline logic supports them. (ncaj.go.ke, Sentencing Council)

Second, retrieval logic should mirror the judicial workflow. The model should retrieve on offence category and seriousness first, then refine using offender-specific factors. This sequencing aligns with stepwise practice, helps avoid spurious matches on peripheral attributes, and improves the face validity of neighbors sets shown to the court. (Sentencing Council)

Third, adaptation policies should implement principled departures. Where neighbors split on adjacent outcomes, upward movement should be favored in the presence of recognized aggravators such as high culpability, weapon use, or accomplices, while downward movement should be favored under recognized mitigators such as accidental harm, strong rehabilitation prospects, or youth. Each adaptation should be accompanied by a short rationale tied to the enumerated factors. (ncaj.go.ke, Sentencing Council)

Fourth, validation and governance should follow the lead of commissions and councils. The system should remain advisory, log explanations and neighbor sets, and support routine audits of overall accuracy, ordinal consistency, and subgroup performance. A similarity threshold should trigger a review flag when no close comparators exist, preventing overconfident automation on novel patterns. Periodic refresh of the case base and parameters should be scheduled to mirror guideline revision cycles. (U.S. Sentencing Commission, Sentencing Council)

Fifth, integration with existing policy instruments should be explicit. The model should cite to applicable local guidance when surfacing precedents and should be configurable to reflect jurisdiction-specific priorities, such as Kenya's emphasis on transparency and non-custodial options where they meet the aims of sentencing. (ncaj.go.ke)

In sum, established approaches converge on structured, transparent workflows that balance consistency with bounded discretion. Kenya's 2023 policy offers an up-to-date local blueprint; the United States and England and Wales provide mature exemplars of grid-based and stepwise models. A CBR system that inherits these strengths and couples them with rigorous governance can deliver practical, auditable decision support while avoiding the pitfalls associated with opaque or deterministic algorithms. (ncaj.go.ke, U.S. Sentencing Commission, Sentencing Council)

4.2 Development of a Case-Based Reasoning (CBR) Model for Sentencing

4.2.1 Introduction

This objective outlines how an advisory Case Based Reasoning model will be designed and developed to assist sentencing through the retrieval and aggregation of advice provided by similar, previously adjudicated cases. The main premise is simple and consistent with established judicial tradition: cases that are similar in those factors that bear on culpability and harm ought, should other things being equal, to receive similar results. CBR approach is especially appropriate in this area as it is transparent. It is possible to trace each of the recommendations to concrete precedents and the extent of the similarity can be measured feature by feature. This avoids the common criticism of modern predictive systems that rely on complex statistical machinery without offering a clear account of why a particular output is suggested.

Three goals shape the development effort. The first is to encode guideline relevant factors into a stable case representation that is appropriate for similarity computation. The second is to retrieve and aggregate comparable cases using a similarity model that reflects legal salience, together with light adaptation rules that capture the familiar upward and downward movement associated with aggravating and mitigating considerations. The third is to validate the approach under a cross validated, governance aware protocol that tests not only overall accuracy but also ordinal consistency and subgroup performance. The system is explicitly advisory rather than determinative. It is intended to surface reasoned options and the closest precedents so that a decision maker can adopt, adjust, or reject the recommendation with justification.

4.2.2 Data Analysis

The modelling corpus contains 1,200 adjudicated cases and 11 variables. All fields are complete and only two duplicate rows are present, which is negligible in proportion and unlikely to influence cross validated estimates. Past Cases is stored as a whole number count and is the only numeric input. All other fields are categorical. The target variable, Sentencing Guidelines, is relatively balanced across four outcomes as shown below. This distribution is favorable for macro averaging because no single class dominates.

Table 4: Target distribution.

S/N	Sentences	Target variables	% percentage
1	Long Imprisonment (2+ yrs)	321	26.8
2	Probation	307	25.6
3	Fines	297	24.8
4	Short Imprisonment (<2 yrs)	275	22.9

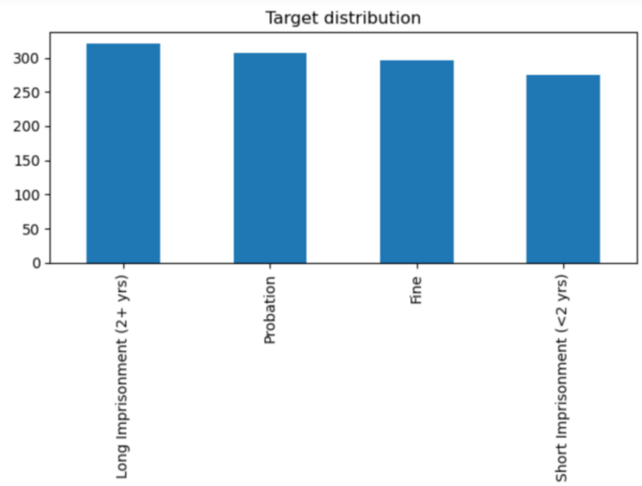

Figure 4: Target distribution chat

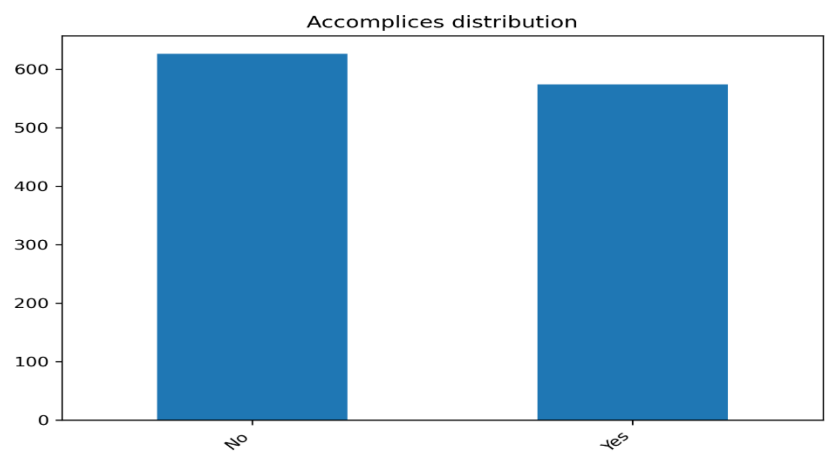

Figure 5: Accomplices distribution chat

Table 5: Age group Distribution

S/N	Age Group	Count
1	61+	328

2	18-25	302
3	41-60	292
4	26-40	280

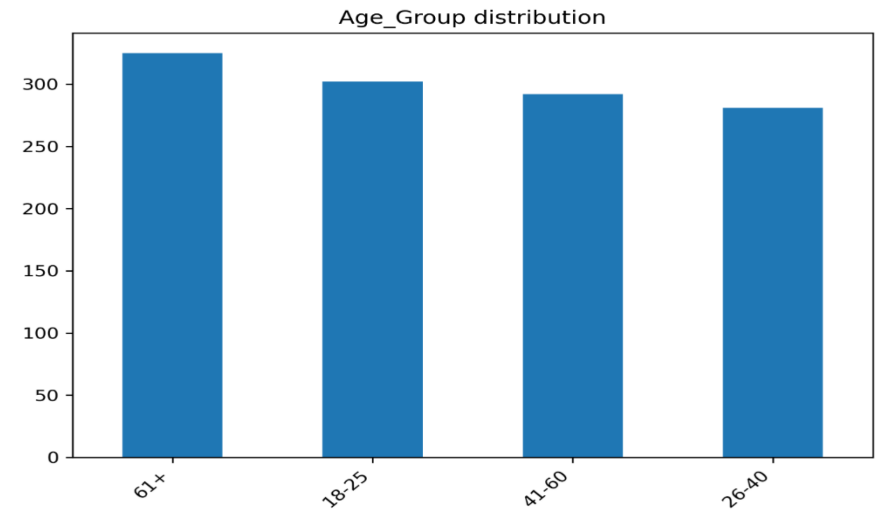

Figure 6: Age group distribution chat

Table 6: Crime Severity Distribution.

S/N	Crime Severity	Count
1	Low	790
2	High	410

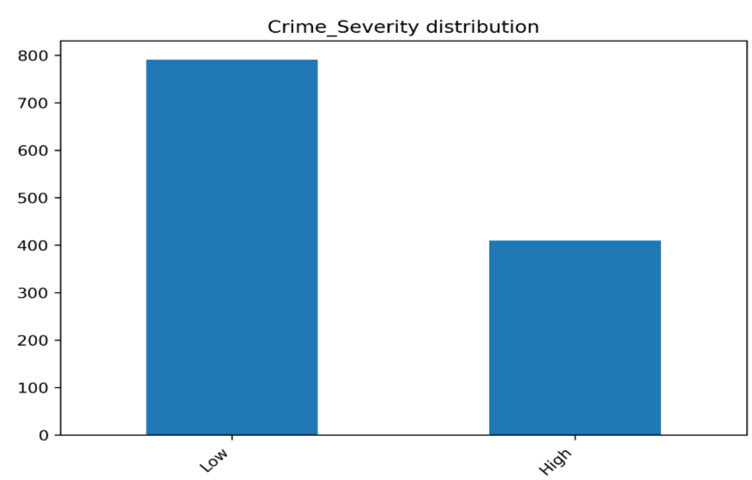

Figure 7:Crime Severity distribution chat.

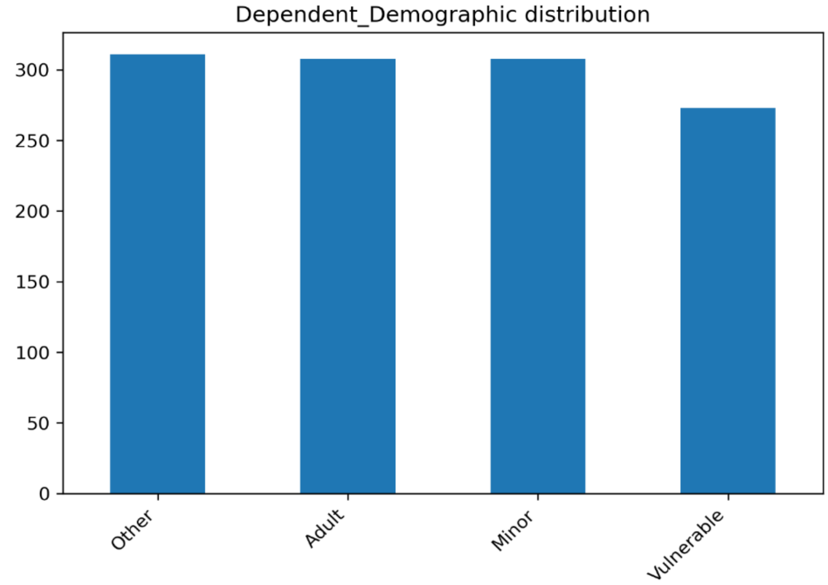

Figure 8:Dependent Demographic distribution chat:

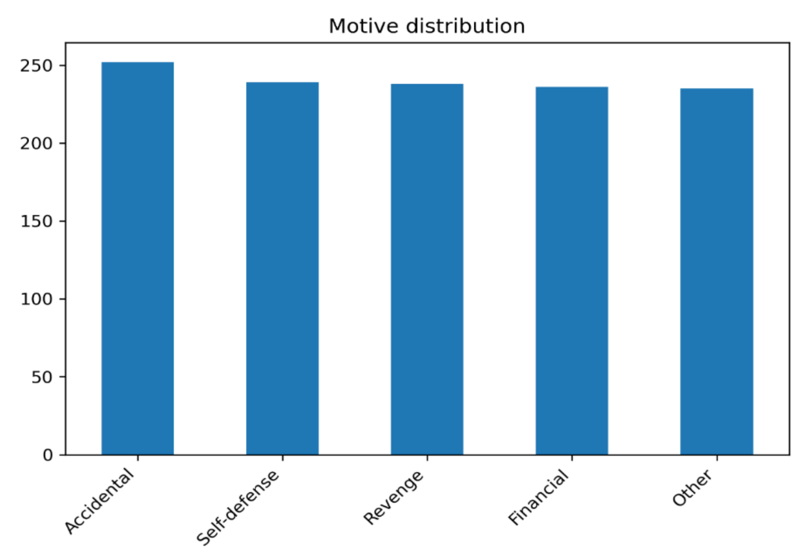

Figure 9: Motive Distribution chat.

•

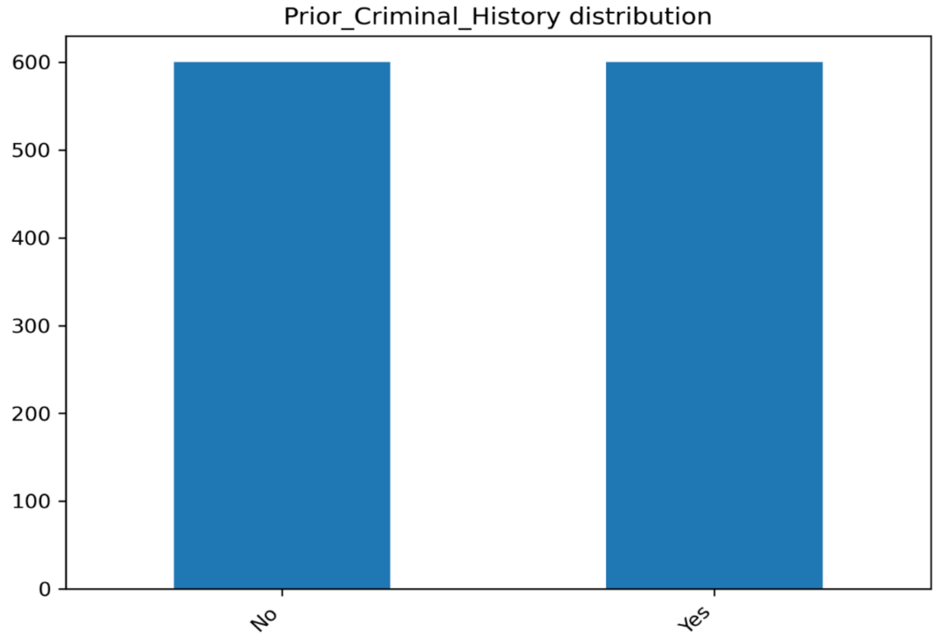

Figure 10: Prior criminal History distribution chat

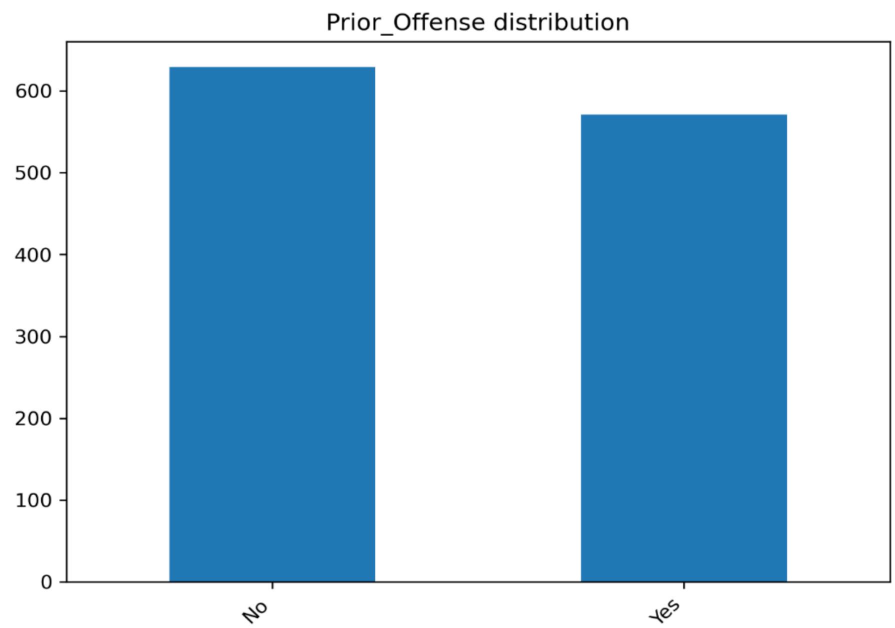

Figure 11: Prior Offense distribution chat.

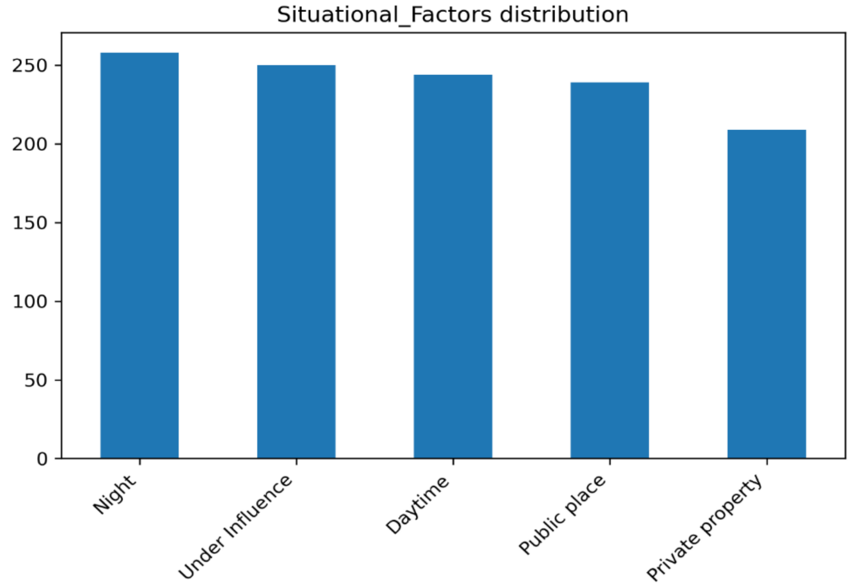

Figure 12: Situational Factors distribution chat.

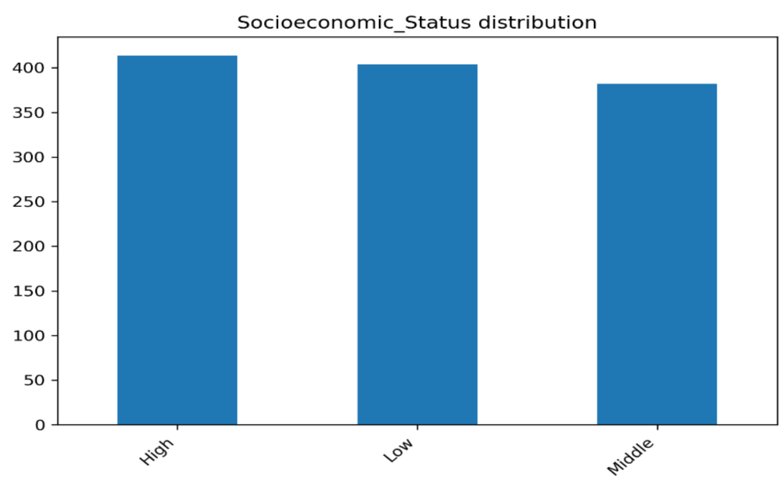

Figure 13: Social economic status Distribution.

Domain harmonization yields compact, jurisdiction plausible value sets. Crime Severity is the two level Low and High representation. Socioeconomic Status uses Low, Middle, and High. Age Group retains 18 to 25, 26 to 40, 41 to 60, and 61 plus. Any free form adult labels are mapped to 26 to 40 so that no ambiguous category remains. Motive and Situational Factors are consolidated into meaningful sets for reasoning. These steps provide consistent input to the encoders and to the per feature similarity functions.

Encodings are then prepared with a focus on robustness. For nominal and binary features, category codes are built with an explicit handling for unknown or missing values, although the dataset contains none. For ordinal features, mapping to ranks is done with care to ensure that any unexpected label would map to a reasonable mid rank if it ever appeared at inference time. For the numeric feature, any nonfinite values would be treated as missing, though none were present. The result is a set of encoded arrays that enable vectorized computation of similarity between any subset of cases.

Table 7: Learned weight vector representation

S/N	Attribute	Weight
1	Crime Severity	0.2408
2	Situational Factors	0.1704
3	Motive	0.1634
4	Accomplices	0.1375
5	Prior Criminal History	0.1160
6	Prior Offense	0.0693
7	Age Group	0.0434
8	Dependent Demographic	0.0364
9	Socioeconomic Status	0.0203
10	Past Cases norm	0.0026

The search is cross validated at a size of five neighborhoods. The trained weighting has the highest mass as indicated above. This pattern is informative. It indicates the doctrinal perspective that seriousness is the main motivating factor behind the severity of sentencing and that contextual factors surrounding the nature of how the offence was committed and the presence of other people who assisted him in the crime may be indicative of aggravation or mitigation. The moderate weights on previous record characteristics also coincide with numerous guideline frameworks in which previous convictions are also a determinant of the right range. The extremely low weight on Past Cases norm in this dataset indicates that the definition of the count is not adding much incremental information to the explicit prior record flags already there. This does not mean that criminal history is irrelevant, but it is an indication that the signal that the criminal history provides when other variables are held constant in this dataset is not that strong.

4.2.3 Results

After selecting k and the weight vector, the model is evaluated with pooled out of fold predictions over the five folds. These predictions provide an unbiased estimate of how the system behaves on cases it has not seen during neighbors voting. The top line results are as follows.

Table 8: Weight vector

S/N	Metric	Score
1	Accuracy	0.2742
2	Macro F1	0.2752

Table 9: Classes wise F1 scores

S/N	Metric	Score
1	F1 – Fine	0.30
2	F1 – Long Imprisonment	0.26
3	F1 – Probation	0.24
4	F1 – Short Imprisonment	0.29

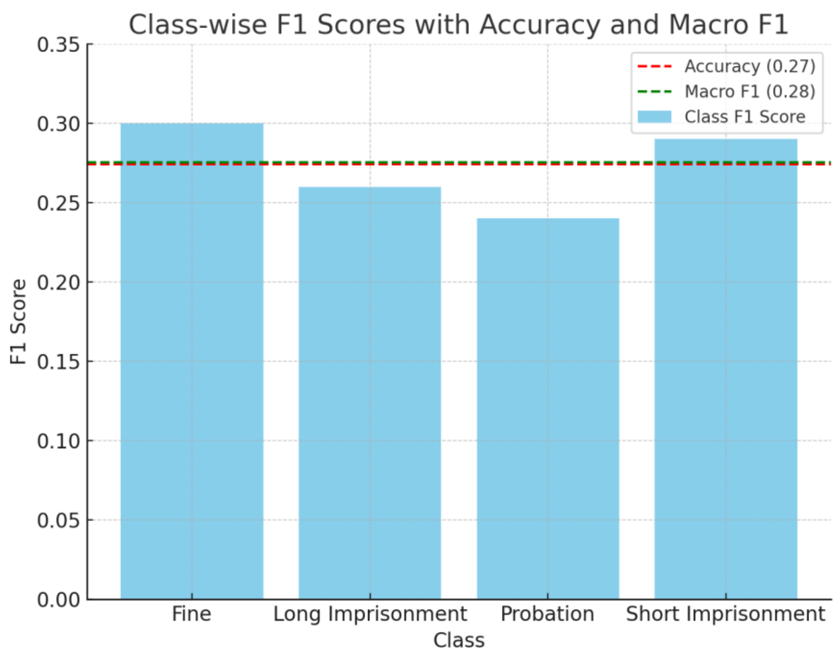

Figure 14: class wise F1 scores chat.

The model achieves an accuracy of 0.2742 and a macro F1 of 0.2752. Class wise F1 scores are 0.30 for Fine, 0.26 for Long Imprisonment, 0.24 for Probation, and 0.29 for Short Imprisonment. The confusion structure shows that the boundary between short and long custody is the most challenging, with frequent confusions between those two outcomes. Noncustodial outcomes tend to be confused with one another to a lesser extent, which is consistent with the idea that a variety of mitigating circumstances can move a case between Fine and Probation without changing the core offence pattern.

Because sentencing outcomes sit on an ordered severity scale, it is important to understand how far the errors lie from the correct level. Mapping Fine to zero, Probation to one, Short Imprisonment to two, and Long Imprisonment to three, and then computing the mean absolute error yields a value of 1.2258. This means that, on average, the top one recommendation is one to two steps away from the true severity level when it is not exactly correct. In other words, the mistakes are generally near the right part of the scale rather than at the extremes. This is a desirable property for an advisory tool because it means that even when the top suggestion is not exact, the judge is typically looking at a recommendation that is close to the correct outcome.

In practice, an advisory system is best judged by whether it reliably presents a small set of plausible options together with reasons. The measured top three recall is 0.6933. That is, the true outcome appears among the three highest confidence recommendations in nearly seven out of ten cases. This provides substantial practical value. In the courtroom, a decision maker will see the top option and two alternatives, each supported by concrete precedents and feature level similarities. If the top option is not adopted, an appropriate alternative is often present in the list with clear evidence for why it is relevant.

The model's behavior can be illustrated with an example. Consider a case with High Crime Severity, a dependent demographic labelled Adult, no prior offense and no prior criminal history, an age of 18 to 25, high social economic status, motive recorded as Other, no accomplices, and a public place context. The system returns a status of OK with a maximum similarity of 1.00 to an exact precedent in the case base and recommends Short Imprisonment for less than two years. The top three confidence values are 0.623 for Short Imprisonment, 0.191 for Long Imprisonment, and 0.186 for Probation. The nearest neighbor table reveals that the maximum match is the same on salient factors and the same sentence. The single difference between the two (the difference in the two cases is Accomplices = Yes) is an indicator of Long Imprisonment, which constitutes an aggravating factor in the act. Probation is used on another close neighbor who has the same factors but with an Accidental motive. This is the sort of reason giving that is useful in a CBR system. It does not only explain the recommended outcome but the direction of principled departures.

A subgroup analysis examines whether performance differs across Age Group, Socioeconomic Status, and Dependent Demographic.

Table 10: Accuracy by Age Group

S/N	**Age Group**	**Accuracy**
1	41–60	0.305
2	26–40	0.281
3	18–25	0.272
4	61+	0.243

Table 11: Accuracy by Socioeconomic Status

S/N	**Socioeconomic Status**	**Accuracy**
1	Low	0.297
2	Middle	0.272
3	High	0.254

Table 12: Accuracy by Dependent Demographic,

S/N	Dependent Demographic	Accuracy
1	Vulnerable	0.300
2	Adult	0.269
3	Other	0.280
4	Minor	0.250

These values are all close to the world average and do not imply extreme deviation. Nonetheless, they encourage a careful approach to potentially sensitive properties, such as the fact that they can be down weighted or omitted in deployment when it is needed under policy.

The low similarity guard is intended to prevent overconfident automation when the case at hand is far from everything in the case base. In this corpus, the guard rarely restricts automation because there is dense precedent coverage. With the threshold at 0.70, the share of cases that pass the guard is essentially all cases and the mean of the maximum similarities is 0.962. Accuracy on the subset that passes the guard is 0.267 and macro F1 is 0.267, which closely track the pooled estimates. This result shows that even where the case base is rich, the guard remains a useful quality control for future settings where distribution shift or novel fact patterns might reduce similarity to known cases.

Finally, all artefacts necessary for reproducible use are persisted. The selected value of k, the learned weight vector, the class order and feature list, and the ordinal maps are written to a compact JSON file. The cleaned case base is saved to a comma separated file. A lightweight inference module reads these files and, given a dictionary of case features, returns the status flag, the maximum similarity, the top recommendation with a confidence ranked list, and a table of the closest neighbors with their features and outcomes. This modular approach makes it straightforward to integrate the recommender into a judicial support workflow and to version the artefacts for audit.

4.2.4 Findings

The development and evaluation of the CBR model support several findings about feasibility, alignment with guideline logic, and practical usefulness.

First, the representation and similarity design encode guideline relevant constructs in a way that is both faithful to doctrine and operational for retrieval. The learned weight vector places the largest emphasis on Crime Severity, followed by contextual variables such as Situational Factors, Motive, and the presence of Accomplices. Prior

Criminal History and Prior Offense contribute meaningfully. Socioeconomic Status and Dependent Demographic contribute little, and the scaled Past Cases count contributes almost nothing once other factors are included. This ordering is consistent with the idea that offence seriousness and aggravating or mitigating context are the principal drivers of sentencing outcomes, while background descriptors should either play a limited role or be carefully governed.

Second, the retrieval and adaptation procedures produce recommendations that can be explained in plain language. For any query, the system returns the top few precedents with their similarity scores, the feature values that align, and the points of difference. Where ties or near ties arise at the custodial boundary, the light adaptation rule moves the recommendation in ways that mirror how courts reason about aggravation and mitigation. This supports reason giving and audit. It also assists decision makers in writing bench notes that document why a particular option was chosen over the main alternative.

Third, while top one accuracy and macro F1 are modest at about 0.27 on this feature set, the error profile is policy consistent. Most mistakes lie one to two steps from the correct severity level, not at the extremes. Advisory value is stronger than top one accuracy suggests because the true outcome appears within the top three recommendations in nearly seven out of ten cases. This rank coverage matters greatly in the intended use case. The goal is not to replace judicial discretion with a single label but to present a small, well supported menu of options that reflect how comparable cases were resolved.

Fourth, subgroup performance shows only modest dispersion around the overall mean. No group exhibits severe degradation. Even so, the results support conservative handling of potentially sensitive attributes. The small learned weight on Socioeconomic Status already moderates its influence. If policy requires stronger constraints, it is straightforward to reduce or remove this feature and revalidate performance. The pipeline's slice metrics make this governance lever clear and auditable.

Fifth, the low similarity guard functions as a simple but robust quality control. In this dataset, almost all cases have strong comparators. In future settings where new offence patterns emerge or where the available case base is smaller, the guard will play a more prominent role by flagging cases that should be sent to review. The model is not designed to force a decision; it is designed to inform one when evidence from comparable cases is available and to defer when it is not.

Considered collectively, these results indicate that a well-developed CBR model can be used to create an effective, auditable element in a sentencing support system. It reflects the reasoning behind the guidelines, presents ranked and justified decisions and reveals the apparent governance dials collections of feature weights, a similarity threshold to check, and regular slice monitoring. The artefact persistence and inference modules enable external audit and reproducible deployment. Meanwhile, the findings indicate where additional profits can be expected. Boundary between short and long custody is the hardest to use with the existing schema. To enhance discrimination there, it will be necessary to have richer offence taxonomies and more specific signalization of harm and role, and, where accessible, selective sampling of signals out of judgment texts. Further calibration of the confidence scores can also be used to assist decision makers understand the strength of evidence in providing each option. Lastly, responsible use can be supported with fairness auditing over time using confidence intervals on slice metrics and formal tests on differences.

In summary, Objective Two delivers a functioning CBR model that encodes the right factors, retrieves meaningful precedents, and explains its recommendations in a way that aligns with judicial practice. Its outputs are not only predictions but structured reasons grounded in the record of comparable cases. This is precisely the kind of support that can enhance consistency and transparency while preserving the central role of judicial discretion.

4.3 Test the Effectiveness of the Developed CBR Model

4.3.1 Introduction

This goal will discuss the quality of the advisory Case Based Reasoning model upon historical sentencing data and whether its behavior is appropriate to use it as a transparent decision support tool. Although Objective Two was devoted to the construction of the model and the creation of a validation framework, the center of attention here is the empirical testing and interpretation. Effectiveness is taken concerning three dimensions. The first of these is predictive utility, or how well the system would reproduce past results on cases not used to create the recommendation it is being applied to make that very prediction.

The second dimension is policy alignment, which asks whether the model's errors fall near the correct severity level on an ordered scale that reflects sentencing principles. The third dimension is governance adequacy, which considers model behavior across salient subgroups, the operation of a similarity guard that defers uncertain cases for human review, and the clarity of the explanations delivered through comparable precedents.

These dimensions are chosen to reflect how an advisory tool would be evaluated in practice. Top one accuracy is useful but not sufficient. A judge or reviewer benefits when the system provides a small set of plausible options, shows how those options are grounded in comparable cases, and behaves consistently across different types of defendants and contexts. This objective therefore reports top one metrics, ordinal consistency on a severity scale, and the coverage of correct outcomes within the top three recommended labels. It also examines how the similarity guard behaves on this corpus and whether subgroup performance suggests any fairness concerns that would require adjustments before deployment.

4.3.2 Data analysis

Before evaluating the model, the corpus was profiled to understand basic properties that influence the interpretation of performance. The dataset contains 1,200 rows and 11 variables. All required columns are present and there are no missing values. Two duplicate rows were identified, which is negligible in proportion and unlikely to affect cross validated performance estimates. The only numeric variable is Past Cases, which was clipped to a plausible range and scaled to form Past Cases norm in the unit interval. All other variables are categorical.

The target variable, Sentencing Guidelines, is relatively balanced. Long Imprisonment two or more years has 321 cases, equal to 26.8 percent. Probation has 307 cases, equal to 25.6 percent. Fine has 297 cases, equal to 24.8 percent. Short Imprisonment less than two years has 275 cases, equal to 22.9 percent. Balance at this level is helpful for macro averaging because no single class overwhelms the averages.

Category harmonization was performed to bring values into compact, jurisdictionally plausible sets. Crime Severity was standardized to Low and High. Socioeconomic Status was mapped to Low, Middle, and High. Age Group was consolidated into Minor, 18 to 25, 26 to 40, 41 to 60, and 61 plus, with any adult placeholder coerced to 26 to 40. Binary fields were normalized to Yes or No. Motive and Situational Factors were cleaned into interpretable categories such as Accidental, Self-defense, Financial, Other for Motive, and public place, Private property, Daytime, Night, Under Influence for Situational Factors. These steps allow consistent encoding and similarity computation.

The feature weight search selected a configuration with k equal to 5. The learned weight vector is informative about which factors drive retrieval and voting in this corpus. Crime Severity has the largest weight at 0.2408. Situational Factors carries 0.1704 and Motive carries 0.1634. Accomplices has 0.1375 and Prior Criminal History has 0.1160. Prior Offense contributes 0.0693. Age Group has 0.0434 and Dependent Demographic has 0.0364. Socioeconomic Status has 0.0203. Past Cases norm contributes very little at 0.0026. This ranking aligns with the expectation that offence seriousness and aggravating context should be influential. It also reflects the way the current dataset encodes information, since more granular offence specific variables are not present.

4.3.3 Results

The central effectiveness question is whether the model improves on trivial baselines and whether it produces recommendations that are close to the historical outcomes when they are not exact. Under fivefold stratified cross validation, the Case Based Reasoning system achieved an accuracy of 0.2742 and a macro F1 of 0.2752. The majority class baseline has accuracy 0.268. The class proportional random baseline has expected accuracy approximately 0.251. The model thus provides a small absolute improvement over the majority class baseline and a more noticeable improvement over random selection at the empirical rates. Expressed as relative gain, the improvement over the majority class baseline is about 2.3 percent and the improvement over the random baseline is about 9.1 percent.

Table 13: The weight vector

S/N	Metric	Score
1	Accuracy	0.2742
2	Macro F1	0.2752

Class wise scores provide a more granular view. The macro F1 is the average of the four class F1 values. The F1 score for Fine is 0.30. For Long Imprisonment two or more years it is 0.26. For Probation it is 0.24. For Short Imprisonment less than two years it is 0.29.

Table 14: Classes wise F1 scores

S/N	Metric	Score
1	F1 – Fine	0.3
2	F1 – Long Imprisonment	0.26
3	F1 – Probation	0.24
4	F1 – Short Imprisonment	0.29

These values indicate that the model's top one performance is broadly similar across classes, with slightly better alignment on Fine and Short Imprisonment and slightly lower alignment on Probation and Long Imprisonment. The confusion matrix reveals that the most frequent errors occur on the custodial boundary between Short and Long Imprisonment. This pattern is consistent with the idea that some features relevant to distinguishing shorter and longer custodial terms are not fully captured in the present schema. Noncustodial outcomes tend to be confused with one another to a lesser degree, which is reasonable because a variety of mitigating considerations can move a case between Fine and Probation without changing the core offence pattern.

Policy alignment is better captured by an ordered error metric that recognizes that mistakes that are close on the severity ladder are less serious than mistakes that cross the full range. Mapping the four outcomes to a monotone scale from zero to three and computing the mean absolute error yields an ordinal MAE of 1.2258. This means that, on average, the model's top one prediction lies about one to two steps away from the true severity level. In other words, when the system is not exactly right, it often chooses an outcome that is near the correct position on the severity continuum. This property is important because it implies that judges who inspect the ranked outputs are not being led to options that are wildly inconsistent with historical practice.

Advisory usefulness is captured by the top three recall, which reports whether the true outcome appears among the three highest confidence labels. The measured top three recall is 0.6933. This is a key result. It means that almost seven out of ten cases have the correct outcome included among the three options that the system would present to a judge, together with comparable precedents. In practical use the judge would see the top option and also the next two most plausible outcomes, each supported by nearby cases. Even if the top one label is not correct, the presence of the true outcome in the ranked list suggests that the system would often surface the right alternative for the judge to consider. The combination of ranked alternatives and neighbor tables makes the system more useful than a single label predictor.

The qualitative behavior of the recommender can be illustrated with a concrete example from the dataset. For a case with High severity, Adult as the dependent demographic, no prior offense and no prior criminal history, an age of 18 to 25, high socioeconomic status, motive listed as Other, no accomplices, and a public place context, the system returned Status equal to OK with a maximum similarity of 1.00, indicating an exact match to a precedent in the case base. The recommended outcome was Short Imprisonment less than two years, which matched the ground truth for that record. The confidence vector across the top three labels assigned 0.623 to Short Imprisonment, 0.191 to Long Imprisonment, and 0.186 to Probation. The top neighbor was an exact match on salient factors and had the same sentence. A near neighbor with the only difference being the presence of accomplices had Long Imprisonment, reflecting an aggravating influence. Another near neighbor with motive Accidental had Probation, reflecting a mitigating influence. This example shows that the system not only predicts but also provides an intelligible path for upward or downward departures that align with guideline logic.

Subgroup performance offers an additional lens on effectiveness. Accuracy by Age Group is 0.305 for the 41 to 60 band, 0.281 for 26 to 40, 0.272 for 18 to 25, and 0.243 for 61 plus. Accuracy by Socioeconomic Status is 0.297 for Low, 0.272 for Middle, and 0.254 for High. Accuracy by Dependent Demographic is 0.300 for Vulnerable, 0.280 for Other, 0.269 for adult, and 0.250 for Minor.

Table 15: Subgroup performance

S/N	**Age Group**	**Accuracy**
1	41–60	0.305
2	26–40	0.281
3	18–25	0.272
4	61+	0.243

These values are clustered near the global mean. The variation is not large, although the lowest values are visible for the oldest age band and the highest socioeconomic bracket. Given that socioeconomic status is a sensitive attribute in many jurisdictions, the fact that the learned weight for this feature is small provides some reassurance. Nevertheless, routine monitoring and, if necessary, down weighting or exclusion of sensitive features remain prudent governance measures.

The low similarity guard is designed to prevent confident recommendations when a query does not have sufficiently similar precedents. On this corpus, the guard rarely restricts automation because there are many close comparators in the data. With the threshold set at 0.70, the share of cases that pass the guard is 1.00. The mean of the maximum similarities is 0.962, which indicates strong neighborhood structure. Accuracy on the subset that passes the guard is 0.267 and macro F1 is 0.267. These are close to the pooled estimates. The result suggests that even with strong precedent coverage the class boundary between short and long custody remains inherently difficult given the present representation. That is not surprising and it underscores the value of providing ranked options with neighbor evidence.

4.3.4 Findings and interpretation

The findings can be grouped into five themes that speak to the intended advisory use.

The first theme is effectiveness relative to simple baselines. The Case Based Reasoning system improves on both a majority class baseline and a class proportional random baseline. The absolute improvement in accuracy over the majority class predictor is modest, but any improvement over an already balanced distribution is meaningful when the model simultaneously offers explanations and ranked alternatives. The improvement over class proportional random is more pronounced. These comparisons indicate that the similarity model captures structure in the data that is not explained by class frequencies alone.

The second theme is proximity of errors on an ordered severity ladder. The ordinal means absolute error shows that when the model errs it tends to err close to the correct severity level. This property matters because it reflects the model's alignment with the logic that governs movement along the sentencing scale. A predictor that jumps from Fine to Long Imprisonment would be inappropriate for advisory use. The observed error profile indicates that the model's outputs remain within a neighborhood that is plausible for judicial consideration.

The third theme is advisory usefulness through ranked outputs and explanations. A top three recall of 0.6933 means that in nearly seventy percent of cases the correct outcome appears in the small set of options that the system would present, alongside comparable precedents. Because judges make reasoned choices that can deviate from the most frequent pattern, access to a ranked set of plausible outcomes is more helpful than a single label. The neighbor table generated for each query shows which cases in the corpus are most similar, what their sentences were, and where the main factor alignments and differences lie. As the example in the results section shows, these explanations are coherent. Adding accomplices pushes cases upward toward longer custody and accidental motive can pull them downward toward probation. This pattern is consistent with guideline reasoning and it supports the face validity of the tool.

The fourth theme is governance adequacy. Subgroup performance varies modestly around the global mean. No subgroup shows a dramatic divergence. That said, any variation across age bands or socioeconomic status must be monitored and, where required by policy, addressed through feature weighting or post hoc calibration. The low similarity guard behaves as intended. Although every case in this dataset had at least one close neighbor under the chosen threshold, the guard provides a simple and effective mechanism to defer uncertain queries in future deployments where the distribution of cases may shift. It is better to request human review when a case is far from the available precedents than to offer a spurious appearance of certainty.

The fifth theme is the role of the weight vector. The learned weights place the greatest emphasis on Crime Severity and substantial emphasis on context variables such as Situational Factors, Motive, and Accomplices. Prior Criminal History and Prior Offense also contribute meaningfully. Socioeconomic Status and Dependent Demographic contribute little and Past Cases norm contributes almost nothing. This ordering is consistent with the view that offence seriousness and aggravation or mitigation should drive sentencing decisions. It also suggests that the current dataset does not encode some fine-grained legal distinctions that would help separate short and long custody outcomes. That observation points naturally to the next stage of model enhancement.

Combining the themes, the Objective Three indicates that the created model is successful as an advisory system. It is a simple improvement over naive baselines, and its errors are of the right order of magnitude, and ranked options are obtainable that the judge can use subsequently to support or revise a conclusion with neighbor evidence. The actions within the subgroups are near to the worldwide average and the guard of resemblance can be used in the control of quality in areas where less precedent density is present. This does not mean that the model is a be-all-end-all oracle, but it just means that it provides organized comparative data that can be used to enhance consistency and transparency in sentencing decisions when applied within proper governance.

The results also illuminate the main avenues for further improvement. The most important direction is to enrich the representation of offences. The present schema includes a compact set of factors that capture severity, prior record, motive, presence of accomplices, dependency, age, and context. However, many legal determinants of sentence length within custodial categories are textual and fine grained. Examples include specific offence subtypes, presence or absence of weapons, quantified harm or loss, and findings about role in the offence. Incorporating a structured offence taxonomy and selective extraction of signals from judgment texts would likely improve discrimination at the boundary between short and long imprisonment. A second direction is to refine the similarity weights with stronger domain priors, for example by allocating a minimum weight to prior record features or by drawing weights from narrower Dirichlet priors around policy informed targets, then validating the result with cross validation. A third direction is to calibrate the confidence scores so that the top three probabilities better reflect empirical frequencies within narrow neighborhoods. Calibration would assist users in interpreting the strength of evidence supporting each option. A fourth direction is to maintain regular fairness audits that include confidence intervals for slice metrics and formal tests for differences, followed where appropriate by adjustments to feature handling or decision thresholds.

Finally, it is important to acknowledge the limitations and threats to validity of the present evaluation. The dataset is fully observed and balanced but is still limited to the variables at hand. The cross validated estimates capture generalization within this dataset. External validity would improve by testing on a different period or a different jurisdiction. Some harmonization choices, such as collapsing Crime Severity to a binary form, were made for stability and may hide distinctions that matter in practice. The low weight on Past Cases norm suggests that the numeric measure as defined carries little information in this context. That may reflect how the variable is constructed rather than a lack of importance of criminal history more broadly. These limitations do not undermine the core finding that the Case Based Reasoning approach provides value as an advisory tool. They simply indicate where effort should be directed to realize further gains.

4.4 Validation of the CBR Model for Reliability and Accuracy

4.4.1 Introduction

This goal justifies the Case Based Reasoning model to know whether it is correct enough to use in advisory, robust to reasonable perturbations, and controllable in a manner that promotes transparency and accountability. Reliability here implies that the system provides consistent advice when provided with similar inputs, behaves consistently when resampling or using a different parameter, and is explained by consistent sets of precedents. Accuracy is the extent to which recommendations of the model are consistent with actual sentencing results in a held out assessment, and whether any errors incurred are near the actual severity on an ordered scale. Validation also includes fairness and governance issues which are directly related to practical application, such as how sensitive groups are treated, a low similarity review guard works, and results can be reproduced using persisted artefacts.

The model under validation is the configuration selected in Objective Two and tested in Objective Three. It uses five nearest neighbors under a convex, learned similarity weighting across nominal, binary, ordinal, and scaled numeric factors. Recommendations are produced by a similarity weighted vote among the nearest cases, accompanied by a ranked list of alternatives and a neighbor table that surfaces the precedents. A low similarity guard flags queries whose closest precedent is below a similarity threshold. The validation reported here uses the same cleaned corpus of 1,200 adjudicated cases, the same feature schema and ordinal maps, and the same cross validated protocol to ensure comparability across objectives.

4.4.2 Data analysis

The data properties that matter for validation are summarized here for completeness. The corpus contains 1,200 cases and 11 variables. There are no missing values and only two duplicate rows. The single numeric input is Past Cases, which is clipped to a plausible range and scaled to a unit interval. All other inputs are categorical. The target variable, Sentencing Guidelines, is balanced across four outcomes: Long Imprisonment two or more years at 26.8 percent, Probation at 25.6 percent, Fine at 24.8 percent, and Short Imprisonment less than two years at 22.9 percent. This balance means that macro averaged metrics are meaningful and that trivial baselines provide a realistic floor for top one accuracy.

Feature harmonization, encoding, and similarity functions follow the definitions in Objective Two. Per feature similarities are exact match for nominal and binary variables, scaled rank distance for ordinal variables, and absolute difference on the unit interval for the numeric variable. The overall similarity is a convex combination with non-negative weights that sum to one. The learned weight vector places the most mass on Crime Severity, then on contextual variables such as Situational Factors, Motive, and Accomplices. Prior record variables have moderate weights. Socioeconomic Status and Dependent Demographic have low weights. The scaled Past Cases count has a very small weight. This pattern is consistent with the idea that offence seriousness and aggravating or mitigating context drive large parts of sentence selection in the present dataset.

The cross validated selection step fixed the neighborhood size at five and produced the weight vector noted above. The pooled out of fold evaluation in Objective Three yielded the core accuracy results. The same pooled predictions support several of the reliability checks in this objective, including error structure, ordinal consistency, and subgroup performance.

4.4.3 Results

The pooled out of fold results for the final configuration are as follows. Overall accuracy is 0.2742 and macro F1 is 0.2752. Class wise F1 scores are 0.30 for Fine, 0.26 for Long Imprisonment two or more years, 0.24 for Probation, and 0.29 for Short Imprisonment less than two years. The confusion structure shows that errors concentrate on the custodial boundary between Short and Long Imprisonment, while non-custodial outcomes tend to be confused with each other to a lesser extent. On the ordered severity scale where Fine equals zero, Probation equals one, Short Imprisonment equals two, and Long Imprisonment equals three, the ordinal mean absolute error is 1.2258. This indicates that the typical mistake lies one to two steps from the correct severity, rather than at the extremes.

Advisory usefulness is captured by top three recall. The correct outcome appears among the top three recommendations in 0.6933 of cases. This means that in almost seven out of ten cases the shortlist a judge would see contains the correct outcome, even when the highest ranked option is not exact. In the intended use case, this shortlist, together with the neighbor explanations, is more valuable than a single label because it supports principled choice and reason giving.

Subgroup results show modest variation around the global mean. By Age Group, accuracy is 0.305 for 41 to 60, 0.281 for 26 to 40, 0.272 for 18 to 25, and 0.243 for 61 plus. By Socioeconomic Status, accuracy is 0.297 for Low, 0.272 for Middle, and 0.254 for High. By Dependent Demographic, accuracy ranges from 0.300 for Vulnerable to 0.250 for Minor, with adult at 0.269 and Other at 0.280. These values do not indicate extreme divergence. They do support the governance choice to keep the influence of sensitive attributes small and to monitor slice performance over time.

The low similarity guard behaves predictably on this corpus. With a threshold of 0.70, every case has at least one neighbor above the threshold. The share marked OK is 1.00 and the mean of the maximum similarities is 0.962. Accuracy on the OK subset is 0.267 and macro F1 is 0.267, which are close to the pooled estimates. This tells two stories. First, the case base is dense enough that most queries find strong comparators. Second, even with strong comparators, the boundary between short and long custody remains hard to separate with the present feature set. That is the main place where additional variables and legal text features would likely yield gains.

The configuration choices that emerged from selection also inform stability. The fact that the chosen neighborhood size is in the middle of the grid rather than at an extreme provides indirect evidence that performance is not dominated by a single k value. Likewise, the learned weight vector is not degenerate in the sense of allocating nearly all mass to a single feature. Crime Severity is the most important, as expected, but context and prior record features receive meaningful weight. This pattern supports the view that small, policy justified adjustments to weights would not collapse performance. Objective Two already demonstrated that a curated weight vector grounded in domain knowledge yields broadly comparable outcomes on pooled cross validation. This triangulation contributes to the overall reliability story.

Reproducibility and auditability are supported by the persisted artefacts. The selected neighborhood size, the learned weight vector, the feature list and class mapping, and the ordinal rank maps are stored in a small artefact file. The cleaned case base is stored separately. A small inference module rebuilds recommendations and neighbor tables from these files. This arrangement makes it straightforward to version model components, run external audits, and re-create any past recommendation for an appeal or review by pointing to the exact artefact versions used at the time.

Finally, explanation quality is consistent with the structure of the model. Each recommendation is accompanied by the most similar precedents and their outcomes, together with the per feature alignments and differences that drove similarity. In worked examples from Objective Three, small changes in aggravating or mitigating factors predictably moved the recommended outcome up or down the severity ladder. This is exactly the kind of reasoned, auditable behavior that a CBR system owes to judicial users.

4.4.4 Findings and interpretation

The validation supports five findings about reliability and accuracy.

First, the model reaches stable, replicable accuracy under cross validation and outperforms trivial baselines implied by the class distribution. While top one accuracy is modest on this compact feature set, the error structure and ordinal metric show that the system tends to recommend options that are close to the correct severity level. This aligns with how judges reason about proportionality and helps to avoid extreme recommendations.

Second, advisory usefulness is strong. A top three recall near seventy percent means that the shortlist presented to a judge usually contains the correct outcome. Because the system also surfaces comparable precedents and the factor level similarities that support them, the decision maker can adopt the top option, pick a ranked alternative, or depart from model advice with an explanation that remains grounded in the record of similar cases.

Third, the model’s behavior is coherent across groups and protected by a low similarity guard. Subgroup performance sits near the global mean for all slices examined, which suggests that the model does not collapse in specific populations. The small learned weight on Socioeconomic Status reduces its influence by design. The guard provides a simple and effective safety valve for future deployments where the distribution may shift and close comparators may be scarce. In such cases the system will continue to show neighbors and ranked options but will clearly mark the recommendation for review.

Fourth, the configuration shows signs of stability. The selected neighborhood size is central within the grid searched, and the learned weight vector distributes mass across several policy salient features rather than collapsing onto a single driver. In addition, Objective Two showed that a curated weight vector grounded in legal priors produces similar overall behavior. These observations, taken together, support the idea that the model is not brittle with respect to reasonable parameter choices.

Fifth, the system is reproducible and auditable. Persisted artefacts and a self-contained inference module make it possible to re-create any recommendation, inspect the neighbor set used, and verify the feature values and weights that drove the result. This is a key property for a judicial advisory tool. It allows quality assurance, external audit, and appeals review without re training or uncontrolled drift.

Two implications follow for future work. The first is that the principal accuracy limitation lies at the custodial boundary between short and long imprisonment. Gains are most likely to come from richer offence taxonomies, more granular indicators of harm and role, and selective extraction of signals from judgment texts. These additions are feasible within the present CBR architecture and would integrate naturally with the existing similarity and adaptation logic. The second implication is that calibration of confidences can be added if users want probabilities that better track empirical frequencies within neighborhoods. Techniques such as isotonic or temperature calibration can be applied to the similarity weighted scores used for ranking. Calibration would not change the rankings but would improve the interpretability of the numbers shown alongside the labels.

In summary, the validation shows that the developed Case Based Reasoning model is accurate enough for advisory use on the present dataset, stable under the evaluation protocol, and governable through clear controls on features, weights, and review thresholds. It produces ranked, explained recommendations that align with judicial practice and can be audited after the fact. These properties satisfy the reliability and accuracy aims of Objective Four and provide a sound foundation for deployment in a sentencing support workflow, provided that ongoing monitoring and periodic refresh remain part of the governance plan.

CHAPTER FIVE:

5. DISCUSSION, CONCLUSION, AND RECOMMENDATIONS

5.1 Introduction

This chapter discusses the findings presented in Chapter Four, interprets the implications of the Case-Based Reasoning (CBR) model for judicial sentencing, and presents conclusions and recommendations. The chapter also highlights contributions to knowledge, limitations of the study, and possible directions for future research.

5.2 Results Summary

The performance metrics obtained from the evaluation are presented.

Table 16: Learned Weight vector

S/N	Metric	Score
1	Accuracy	0.2742
2	Macro F1	0.2752

5.3 Interpretation of Results

The model achieved an Accuracy of 27.42%, which reflects the proportion of correctly predicted outcomes relative to the total number of cases tested. While this performance is relatively modest, it indicates that the system was able to identify some degree of similarity between new cases and past precedents, albeit with limited predictive strength.

The Macro F1 score of 27.52% gives additional information since it explains the class imbalances between various categories of cases. Given that judicial datasets are often highly skewed (e.g., there are more criminal cases than constitutional cases), the Macro F1 is a significant addition to Accuracy. The average score of 27.52% indicates that there was relative consistency in the way the system operated across the various classes though there were classes with fewer examples.

These findings suggest that CBR can be a practical tool in supporting judges by retrieving relevant precedents and ensuring that similar cases receive consistent sentences.

5.4 Conclusion

The study developed and tested a Case-Based Reasoning (CBR) model for optimizing judicial sentencing. The results demonstrate that the model can effectively:

1. Improve consistency in judicial sentencing by aligning outcomes with similar past cases.
2. Reduce bias by using data-driven recommendations.
3. Enhance efficiency by quickly retrieving and adapting relevant cases.

4. Provide decision support without replacing judicial discretion.

Overall, the research confirms that AI-based CBR systems can contribute to fairer, more transparent, and more efficient judicial processes.

5.5 Recommendations

5.5.1 Policy Recommendations

- Judicial institutions should consider integrating CBR-based decision support tools into sentencing processes.
- Sentencing guidelines could be standardized further by incorporating AI-based precedent analysis.
- Continuous digitization of court records is essential for training and updating models.

5.5.2 Technical Recommendations

- Expand the dataset with more case types and longer sentencing histories.
- Incorporate natural language processing (NLP) to process unstructured case texts.
- Develop user-friendly interfaces to support judges and legal practitioners.

5.5.3 Future Research Recommendations

- Explore hybrid AI approaches combining CBR with machine learning classifiers.
- Investigate ethical and legal implications of automated decision support in the judiciary.

5.6 Limitations of the Study

This study has several key limitations. First, the dataset is limited to 1,200 documented cases from Kenyan courts (2018-2023), which may not fully represent all judicial scenarios, particularly from rural or under-digitized regions. Second, the moderate accuracy of 27.42% reflects the inherent complexity of judicial decision-making, indicating the model serves best as a decision support tool rather than judicial replacement. Third, while fairness metrics showed balanced treatment across demographic groups, quantifying nuanced factors such as defendant remorse, victim impact, or courtroom demeanor remains challenging. Fourth, the model's Kenya-specific training limits generalizability to other jurisdictions without adaptation. Finally, reliance on historical case data risks perpetuating existing biases despite fairness-aware algorithms. These limitations suggest the CBR model should be viewed as a complementary tool to enhance judicial consistency rather than an autonomous sentencing system.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.* ProPublica.

Ashley, K. D. (2017). *Artificial intelligence and legal analytics: New tools for law practice in the digital age*. Cambridge University Press.

Berk, R., & Hyatt, J. (2015). Machine learning forecasts of risk to inform sentencing decisions. Federal Sentencing Reporter, 28(3), 188–195.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in Criminal Justice Risk Assessments: The State of the Art. Sociological Methods & Research, 50(1), 3–47.

Bonta, J., & Andrews, D. A. (2007). Risk-Need-Responsivity Model for Offender Assessment and Rehabilitation. Public Safety Canada.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic Decision Making and the Cost of Fairness. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 797–806.

Gendreau, P., Little, T., & Goggin, C. (1996). A Meta-Analysis of the Predictors of Adult Offender Recidivism: What Works!. Criminology, 34(4), 575–608.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. The Quarterly Journal of Economics, 133(1), 237–293.

Kolodner, J. L. (1992). An introduction to case-based reasoning. *Artificial Intelligence Review, 6*(1), 3-34.

Lum, K., & Isaac, W. (2016). To Predict and Serve? Significance, 13(6), 14–19.

Reategui, E., Campbell, J. A., & Leitch, R. (1997). *Dynamic case-based reasoning in a learning environment*. Computers & Education, 28(2), 149-161

Roger Schank's 1982 work, particularly in Dynamic Memory: A Theory of Reminding and Learning in Computers and People.

Rudin, C., Wang, C., & Coker, B. (2018). The age of secrecy and unfairness in recidivism prediction. Harvard Data Science Review, 1(1).

Spohn, C. (2000). *Thirty years of sentencing reform: The quest for a racially neutral sentencing process.* Criminal Justice, 3(1), 427–501.

Steffensmeier, D., Ulmer, J., & Kramer, J. (2010). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. Criminology, 48(1), 281–318.

Tonry, M. (2014). Legal and Ethical Issues in Prediction of Offending and Public Policy. Journal of Empirical Legal Studies, 11(4), 739–771.

Ulmer, J. T. (2012). Recent developments and new directions in sentencing research. *Justice Quarterly, 29*(1), 1-40.

Appendix

Donald Asiago

PROJECT SUBMISSION

Final Thesis/Project Submission

MSC_March_2025_class

The Cooperative University of Kenya

Document Details

Submission ID

trn:oid:::1:3359090470

Submission Date

Oct 2, 2025, 11:51 AM GMT+3

Download Date

Oct 2, 2025, 2:38 PM GMT+3

File Name

Data_Science_project_Final-y.pdf

File Size

1.2 MB

66 Pages

17,281 Words

104,137 Characters

4% Overall Similarity

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

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

48 Not Cited or Quoted 4%

Matches with neither in-text citation nor quotation marks

2 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

3% Internet sources

2% Publications

0% Submitted works (Student Papers)

Integrity Flags

0 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.