KATHMANDU UNIVERSITY

Department of Artificial Intelligence

Dhulikhel, Kavre

Lab Report On: Data Mining and Ethics

Submitted By:

Name: Aaryan Shakya

Roll No: 20

Subject Code: AICC 301

Submitted To:

Mr. Sunil Regmi

Lecturer, Department of Artificial Intelligence

Submission Date:

6th July 2025

Objectives

1. Understand privacy and security concerns in data mining
2. Examine social impacts such as bias and discrimination in data mining models
3. Explore the importance of accountability and transparency in model development

Introduction

1. Privacy and Security

Privacy and security are fundamental concerns in data mining because the process often involves collecting and analyzing sensitive personal data. Protecting this information is critical to prevent misuse or unauthorized access.

- **Anonymization:** This technique involves removing or masking personally identifiable information (PII) such as names, social security numbers, or addresses from datasets before analysis. For example, a hospital sharing patient data for research would anonymize it to protect patient identities.
- **Encryption:** Data encryption transforms data into a coded form that only authorized parties can decode, ensuring data confidentiality during storage and transmission. For instance, financial institutions encrypt transaction records to secure customer data from hackers.
- **Consent:** Ethical data mining requires obtaining informed consent from individuals before collecting or using their data. This means clearly explaining how data will be used and giving people the choice to opt-in or opt-out. For example, apps that collect user data for personalization must request permission and explain what data is gathered.

2. Social Impact: Discrimination and Bias in Models

Data mining models can unintentionally reflect or amplify existing social biases present in the data, leading to discriminatory outcomes.

- **Discrimination:** When models unfairly disadvantage certain groups based on attributes like race, gender, or age, it results in discrimination. For example, an employment screening algorithm trained on biased historical hiring data might systematically reject qualified candidates from minority groups.
- **Bias:** Bias can occur during data collection (e.g., underrepresenting certain populations), sampling, or labeling. If the data does not fairly represent all groups, the model's predictions will be skewed. For instance, facial recognition systems trained mostly on lighter-skinned faces have shown lower accuracy for darker-skinned individuals.

Mitigating these biases involves careful data auditing, diverse sampling, and fairness-aware algorithms to promote equitable outcomes.

3. Accountability: Transparency in Model Building and Decision Making

Accountability means that data miners, developers, and organizations are responsible for their models and decisions, ensuring these are transparent and explainable.

- **Transparent Model Building:** Models should be designed and documented so stakeholders can understand how they work. For example, simpler models like decision trees can be easily interpreted compared to complex “black-box” neural networks.
- **Decision Making:** Decisions made by models should be explainable so that affected individuals can understand why a particular outcome occurred. For instance, in loan approvals, providing clear explanations for rejections builds trust and allows for appeals or corrections.

Accountability promotes trust, ethical compliance, and enables detection of errors or unfairness in automated systems.

Experiment

Analyzing Fairness and Bias in Data Mining Models with Ethical Case Studies

Problem:

• Analyze a biased dataset and demonstrate fairness issues.

• Compare outputs of models trained on biased vs debiased data.

• Discuss ethical case studies (e.g., COMPAS recidivism algorithm).

Solution:

Step 1: Analyze a Biased Dataset and Demonstrate Fairness Issues

- **Dataset: Use the Adult Census Income dataset, known for containing gender and racial biases in income prediction.**
- **Load the dataset and identify sensitive attributes (e.g., gender, race).**
- **Train a classification model (e.g., Logistic Regression) to predict income level.**
- **Evaluate model performance overall and separately for different demographic groups to detect disparities.**

Python Code

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report

# Load dataset (assumed preprocessed with sensitive attributes)
data = pd.read_csv('adult.csv')

# Sensitive attribute: gender
X = data.drop(columns=['income', 'gender'])
y = data['income']
gender = data['gender']

# Split data
X_train, X_test, y_train, y_test, gender_train, gender_test = train_test_split(
    X, y, gender, test_size=0.3, random_state=42)

# Train model on biased data
model = LogisticRegression(max_iter=200)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

```
# Overall accuracy
print("Overall Accuracy:", accuracy_score(y_test, y_pred))

# Accuracy by gender
for g in gender_test.unique():
    idx = gender_test == g
    print(f"Accuracy for gender {g}: {accuracy_score(y_test[idx], y_pred[idx])}")

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

# Load preprocessed Adult dataset with 'gender' as a column and 'income' as target
data = pd.read_csv('adult_preprocessed.csv')

# Define features and target
X = data.drop(columns=['income'])
y = data['income']

# Split dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Train on biased data (with gender)
model_biased = LogisticRegression(max_iter=200)
model_biased.fit(X_train, y_train)
y_pred_biased = model_biased.predict(X_test)

print("Biased Model Accuracy:", accuracy_score(y_test, y_pred_biased))

# Evaluate accuracy by gender group
for gender in X_test['gender'].unique():
    idx = X_test['gender'] == gender
    acc = accuracy_score(y_test[idx], y_pred_biased[idx])
    print(f"Accuracy for gender {gender} (biased model): {acc:.4f}")

# Debias by removing gender
X_train_debiased = X_train.drop(columns=['gender'])
X_test_debiased = X_test.drop(columns=['gender'])

model_debiased = LogisticRegression(max_iter=200)
model_debiased.fit(X_train_debiased, y_train)
y_pred_debiased = model_debiased.predict(X_test_debiased)
```

```
print("Debiased Model Accuracy:", accuracy_score(y_test, y_pred_debiased))

for gender in X_test['gender'].unique():
    idx = X_test['gender'] == gender
    acc = accuracy_score(y_test[idx], y_pred_debiased[idx])
    print(f"Accuracy for gender {gender} (debiased model): {acc:.4f}")
```

Step 2: Debiasing and Comparing Models

- **Apply simple debiasing by removing sensitive attributes or using resampling techniques to balance the dataset.**
- **Retrain the model and compare fairness metrics such as accuracy parity or disparate impact.**

Python Code

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

# Load preprocessed Adult dataset with 'gender' as a column and 'income' as target
data = pd.read_csv('adult_preprocessed.csv')

# Define features and target
X = data.drop(columns=['income'])
y = data['income']

# Split dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Train on biased data (with gender)
model_biased = LogisticRegression(max_iter=200)
model_biased.fit(X_train, y_train)
y_pred_biased = model_biased.predict(X_test)

print("Biased Model Accuracy:", accuracy_score(y_test, y_pred_biased))

# Evaluate accuracy by gender group
for gender in X_test['gender'].unique():
    idx = X_test['gender'] == gender
    acc = accuracy_score(y_test[idx], y_pred_biased[idx])
    print(f"Accuracy for gender {gender} (biased model): {acc:.4f}")
```

```
# Debias by removing gender
X_train_debiased = X_train.drop(columns=['gender'])
X_test_debiased = X_test.drop(columns=['gender'])

model_debiased = LogisticRegression(max_iter=200)
model_debiased.fit(X_train_debiased, y_train)
y_pred_debiased = model_debiased.predict(X_test_debiased)

print("Debiased Model Accuracy:", accuracy_score(y_test, y_pred_debiased))

for gender in X_test['gender'].unique():
   idx = X_test['gender'] == gender
   acc = accuracy_score(y_test[idx], y_pred_debiased[idx])
   print(f"Accuracy for gender {gender} (debiased model): {acc:.4f}")
```

Observations:

The biased model often demonstrates varying levels of accuracy across gender groups, highlighting issues of unfairness. In contrast, the debiased model—trained without using gender as a feature—tends to minimize these disparities. While this may lead to a slight reduction in overall accuracy, it generally results in improved fairness.

Step 3: Ethical Case Study – The COMPAS Algorithm

The COMPAS algorithm, commonly used in the criminal justice system to assess the risk of recidivism, has been the subject of extensive ethical scrutiny:

- Racial Bias: Studies have found that Black defendants are disproportionately labeled as high-risk compared to white defendants, even after accounting for prior criminal records and other relevant factors.
- Misclassification Issues: Black defendants who do not reoffend are nearly twice as likely to be incorrectly classified as high-risk. Conversely, white defendants who do reoffend are more frequently misclassified as low-risk.
- Accuracy vs. Fairness: Although COMPAS achieves a moderate accuracy rate of approximately 61–63%, it exhibits significant disparities in error rates between racial groups, raising concerns about fairness.
- Influence on Human Decisions: Research suggests that judges and parole officers who rely on COMPAS scores may unknowingly reinforce these biases, potentially affecting sentencing and parole outcomes.

Conclusion

This experiment highlighted the critical ethical challenges posed by bias and fairness in data mining. By analyzing a biased dataset, we observed how prejudiced data can lead to unfair model outcomes that disadvantage certain groups. Comparing models trained on biased versus debiased data demonstrated the importance of carefully preprocessing data to reduce discrimination and promote equitable results.

Furthermore, discussing real-world ethical case studies, such as the COMPAS recidivism algorithm, underscored the real impact these issues have on individuals and society, emphasizing the need for transparency, accountability, and ongoing vigilance in model development. Overall, the lab reinforced that ethical considerations are essential for building responsible, fair, and trustworthy data mining systems.