# Bias Analysis Tool

CSV File

Choose file | NYPD_Complaint_Data_Historic_20250515_preprocessed.csv

Using file: NYPD_Complaint_Data_Historic_20250515_preprocessed.csv (cached)

Columns Description

* STATION_NAME: transit station name, if incident occurred on transit property.
* VIC_AGE_GROUP: victim's age group.
* VIC_RACE: victim's race category.
* VIC_SEX: victim's sex.

Describe the columns in your dataset for the LLM to understand.

Number of Rows

10000

0 for all rows

Test Size

0.3

Maximum Categories

10

LLM Model

llama_3_3

☑ Use all 3 LLMs for analysis (get results from llama_3_3, deepseek_r1, and mistral_nemo)

When enabled, the analysis will use all available LLM models and provide a comparison of their recommendations and bias assessments.

Analyze

# LLM Recommendations

**Multi-LLM Analysis Results:**

| Model | Target Column | Protected Attributes | Excluded Columns | Race Column | Privileged Groups | Unprivileged Gr |
|---|---|---|---|---|---|---|
| **deepseek_r1** | CRM_ATPT_CPTD_CD | SUSP_AGE_GROUP, SUSP_RACE, SUSP_SEX, VIC_AGE_GROUP, VIC_RACE, VIC_SEX | OFNS_DESC, PD_DESC, LAW_CAT_CD, KY_CD, CMPLNT_FR_TM, CMPLNT_TO_TM, TRANSIT_DISTRICT, Lat_Lon, CMPLNT_TO_DT, HOUSING_PSA | SUSP_RACE | WHITE | UNKNOWN, (null), BLACK, WHITE HISPANIC, BLACK HISPANIC, ASIAN / PACIFIC ISLANDER, AMERICAN INDIAN/ALASKAN NATIVE |
| **llama_3_3** | LAW_CAT_CD | SUSP_RACE, SUSP_SEX, SUSP_AGE_GROUP, VIC_RACE, VIC_SEX, VIC_AGE_GROUP | KY_CD, OFNS_DESC, PD_DESC, PD_CD | SUSP_RACE | WHITE, WHITE HISPANIC, ASIAN / PACIFIC ISLANDER | BLACK, BLACK HISPANIC, AMERIC INDIAN/ALASKAN NATIVE |
| **mistral_nemo** | LAW_CAT_CD | SUSP_RACE, SUSP_SEX, VIC_RACE, VIC_SEX, SUSP_AGE_GROUP, VIC_AGE_GROUP | KY_CD, OFNS_DESC, PD_DESC, PD_CD, BORO_NM, PATROL_BORO, Longitude, X_COORD_CD, PREM_TYP_DESC, LOC_OF_OCCUR_DESC | SUSP_RACE | WHITE, WHITE HISPANIC | UNKNOWN, (null), BLACK, BLACK HISPANIC, ASIAN / PACIFIC ISLANDER, AMERICAN INDIAN/ALASKAN NATIVE |

## Detailed Results:

### deepseek_r1

**Target Column:** CRM_ATPT_CPTD_CD

**Protected Attributes:** SUSP_AGE_GROUP, SUSP_RACE, SUSP_SEX, VIC_AGE_GROUP, VIC_RACE, VIC_SEX

**Excluded Columns:** OFNS_DESC, PD_DESC, LAW_CAT_CD, KY_CD, CMPLNT_FR_TM, CMPLNT_TO_TM, TRANSIT_DISTRICT, Lat_Lon, CMPLNT_TO_DT, HOUSING_PSA

**Race Column:** SUSP_RACE

**Privileged Groups:** WHITE

**Unprivileged Groups:** UNKNOWN, (null), BLACK, WHITE HISPANIC, BLACK HISPANIC, ASIAN / PACIFIC ISLANDER, AMERICAN INDIAN/ALASKAN NATIVE

**Top Correlated Features:**

```
OFNS_DESC=0.164
PD_DESC=0.150
LAW_CAT_CD=0.116
KY_CD=0.101
CMPLNT_FR_TM=0.099
CMPLNT_TO_TM=0.099
TRANSIT_DISTRICT=0.077
Lat_Lon=0.070
CMPLNT_TO_DT=0.070
HOUSING_PSA=0.070
```

## llama_3_3

**Target Column:** LAW_CAT_CD

**Protected Attributes:** SUSP_RACE, SUSP_SEX, SUSP_AGE_GROUP, VIC_RACE, VIC_SEX, VIC_AGE_GROUP

**Excluded Columns:** KY_CD, OFNS_DESC, PD_DESC, PD_CD

**Race Column:** SUSP_RACE

**Privileged Groups:** WHITE, WHITE HISPANIC, ASIAN / PACIFIC ISLANDER

**Unprivileged Groups:** BLACK, BLACK HISPANIC, AMERICAN INDIAN/ALASKAN NATIVE

**Top Correlated Features:**

```
KY_CD=0.971
OFNS_DESC=0.696
PD_DESC=0.531
PD_CD=0.262
BORO_NM=0.159
PATROL_BORO=0.159
Longitude=0.132
X_COORD_CD=0.132
PREM_TYP_DESC=0.122
LOC_OF_OCCUR_DESC=0.111
```

## mistral_nemo

**Target Column:** LAW_CAT_CD

**Protected Attributes:** SUSP_RACE, SUSP_SEX, VIC_RACE, VIC_SEX, SUSP_AGE_GROUP, VIC_AGE_GROUP

**Excluded Columns:** KY_CD, OFNS_DESC, PD_DESC, PD_CD, BORO_NM, PATROL_BORO, Longitude, X_COORD_CD, PREM_TYP_DESC, LOC_OF_OCCUR_DESC

**Race Column:** SUSP_RACE

**Privileged Groups:** WHITE, WHITE HISPANIC

**Unprivileged Groups:** UNKNOWN, (null), BLACK, BLACK HISPANIC, ASIAN / PACIFIC ISLANDER, AMERICAN INDIAN/ALASKAN NATIVE

**Top Correlated Features:**

```
KY_CD=0.971
OFNS_DESC=0.696
PD_DESC=0.531
PD_CD=0.262
BORO_NM=0.159
PATROL_BORO=0.159
Longitude=0.132
X_COORD_CD=0.132
PREM_TYP_DESC=0.122
LOC_OF_OCCUR_DESC=0.111
```

## Key Differences Summary:

**Target Column Disagreement:** Models suggest different target columns: CRM_ATPT_CPTD_CD, LAW_CAT_CD

**Protected Attributes Variation:** Different models identified different protected attributes

# Multi-LLM Bias Analysis Comparison

| Protected Attribute | LLM Model | Bias Analysis Status | Key Findings | Bias Level |
|---|---|---|---|---|
| **SUSP_AGE_GROUP** | **deepseek_r1** | Success | Okay, so I need to figure out how to analyze the bias in this machine learning model based on the given data. Let me start by understanding what the data is showing. The context provided includes a bi... | High |
| **SUSP_AGE_GROUP** | **llama_3_3** | Success | ## Bias Level Classification Based on the global feature importance analysis, the bias level is classified as **MEDIUM**. This classification is due to the presence of moderate bias indicators, such a... | Moderate |
| **SUSP_AGE_GROUP** | **mistral_nemo** | Success | Okay, so I'm trying to analyze the bias in this machine learning model based on the provided data. The protected attribute here is SUSP_AGE_GROUP, which I assume refers to the age group of the suspect... | High |
| **SUSP_RACE** | **deepseek_r1** | Success | Okay, so I need to analyze the provided bias and classification analysis data for the protected attribute 'SUSP_RACE'. The user has given me a detailed breakdown of precision, recall, bias metrics, an... | High |
| **SUSP_RACE** | **llama_3_3** | Success | ## Bias Level Classification Based on the provided data, the bias level classification is **HIGH**. This is due to the significant disparity in importance scores for the `SUSP_RACE` feature across dif... | Fair |
| **SUSP_RACE** | **mistral_nemo** | Success | Okay, so I'm trying to figure out how to analyze the bias in this dataset based on the provided information. Let me start by understanding what each part means. First, the user provided a detailed br... | High |

| Protected Attribute | LLM Model | Bias Analysis Status | Key Findings | Bias Level |
|---|---|---|---|---|
| **SUSP_SEX** | **deepseek_r1** | Success | Okay, so I need to analyze the bias in a model based on the provided data. The protected attribute here is SUSP_SEX, which I assume stands for suspect sex. The analysis includes overall metrics, group... | High |
| **SUSP_SEX** | **llama_3_3** | Success | ## Bias Analysis Report ### Introduction This report provides a comprehensive analysis of bias in the given dataset, focusing on the protected attribute 'SUSP_SEX'. The analysis is based on the global... | Fair |
| **SUSP_SEX** | **mistral_nemo** | Success | Okay, so I need to figure out how to approach this bias and fairness analysis based on the data provided. Let me start by understanding the problem. The protected attribute here is SUSP_SEX, which I a... | High |
| **VIC_AGE_GROUP** | **deepseek_r1** | Success | Okay, so I need to analyze the bias in this dataset based on the provided feature importance and classification metrics. The protected attribute here is VIC_AGE_GROUP, but I also see other demographic... | High |
| **VIC_AGE_GROUP** | **llama_3_3** | Success | ### Bias Analysis Report #### 1. Bias Level Classification Based on the provided global feature importance analysis, the bias level classification is **MEDIUM**. This classification is due to the pre... | Fair |
| **VIC_AGE_GROUP** | **mistral_nemo** | Success | Alright, so I'm trying to analyze the bias in this dataset, focusing on the protected attribute VIC_AGE_GROUP. Let me start by going through the provided data and the questions I need to address. Fir... | Moderate |
| **VIC_RACE** | **deepseek_r1** | Success | Okay, so I'm trying to figure out how to analyze the bias in this dataset. Let me start by looking at the feature | High |

| Protected Attribute | LLM Model | Bias Analysis Status | Key Findings | Bias Level |
|---|---|---|---|---|
| | | | importance and the protected attribute, which is VIC_RACE. The analysis provided shows... | |
| **VIC_RACE** | **llama_3_3** | Success | ## Bias Analysis Report ### Introduction The provided data includes a global feature importance analysis for a classification model predicting crime classes (FELONY, MISDEMEANOR, VIOLATION) with a foc... | Unknown |
| **VIC_RACE** | **mistral_nemo** | Success | Okay, so I'm trying to analyze the bias in this dataset based on the provided metrics. Let me start by understanding what's given. The protected attribute here is 'VIC_RACE', which I assume refers to ... | High |
| **VIC_SEX** | **deepseek_r1** | Success | Alright, I need to analyze the provided bias and feature importance data to assess potential biases in the model, particularly focusing on the protected attribute VIC_SEX. Let's start by understanding... | High |
| **VIC_SEX** | **llama_3_3** | Success | ## Bias Analysis Report ### Introduction This report provides a comprehensive bias analysis of the given data, focusing on the protected attribute 'VIC_SEX'. The analysis covers feature ranking patter... | Fair |
| **VIC_SEX** | **mistral_nemo** | Success | Okay, so I need to analyze the bias and fairness in the given data. Let me start by understanding what's provided. The data includes bias and classification analysis for the protected attribute 'VIC_S... | High |

# Analysis for Protected Attribute: SUSP_AGE_GROUP

## Multi-LLM Bias Analysis

| LLM Model | Bias Analysis | Status |
|---|---|---|
| **deepseek_r1** | Okay, so I need to figure out how to analyze the bias in this machine learning model based on the given data. Let me start by understanding what the data is showing. The context provided includes a bias and classification analysis for the protected attribute 'SUSP_AGE_GROUP'. There's also a feature importance analysis broken down by class: Overall, FELONY, MISDEMEANOR, and VIOLATION.<br><br>First, I'll look at the feature importance. The top features across all classes are SUSP_SEX, SUSP_RACE, VIC_SEX, SUSP_AGE_GROUP, and VIC_AGE_GROUP. SUSP_RACE is the second most important overall, which is a protected attribute. That's a red flag because it might mean the model is relying too much on race, which could lead to bias.<br><br>Looking at each class specifically, for FELONY and VIOLATION, the feature importance is similar to the overall ranking. But for MISDEMEANOR, SUSP_RACE is the most important by a huge margin, which is concerning. It suggests that race is heavily influencing misdemeanor predictions, which isn't fair.<br><br>Now, considering bias indicators, SUSP_RACE and SUSP_SEX are both protected attributes. They might be acting as proxies for other demographic factors, which could lead to unfair treatment of certain groups. For example, if the model is using race to predict misdemeanors more than other factors, it might disproportionately affect specific racial groups.<br><br>Class-specific bias is evident in how SUSP_RACE dominates in MISDEMEANOR cases. This could mean that the model is making decisions based on race rather than other, more relevant factors. Also, the high importance of SUSP_SEX in | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | FELONY and VIOLATION but not in MISDEMEANOR shows that gender might be influencing certain types of offenses more than others.<br><br>Features with high importance, like SUSP_RACE and SUSP_SEX, are above 15% and need scrutiny. They are likely influencing the model's decisions significantly, which raises fairness concerns. Their interaction with SUSP_AGE_GROUP might mean that the model is considering age along with race and gender, potentially leading to compounded bias.<br><br>For the bias level classification, since SUSP_RACE is a top feature and especially dominant in MISDEMEANOR, this points to a HIGH bias level. There's clear unfair treatment, particularly in how race influences outcomes differently across offense types.<br><br>Specific bias patterns show that Black individuals have lower recall in FELONY and VIOLATION, meaning they might be less likely to be correctly classified, leading to potential under-policing or misclassification. On the other hand, White individuals have higher recall in VIOLATION, which might result in over-policing.<br><br>The risk assessment is significant because relying on race and gender can perpetuate existing biases in the criminal justice system. This could lead to unfair treatment of certain groups, undermining trust in the system and potentially causing real harm to individuals.<br><br>Overall, the model's heavy reliance on protected attributes indicates a critical need for fairness assessments and possible adjustments to reduce bias. | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | ### Bias Analysis Report<br><br>#### 1. Bias Level Classification<br>- **HIGH**: The model exhibits significant bias, particularly in how race influences predictions across different offense types, indicating unfair treatment.<br><br>#### 2. Specific Bias Patterns<br>- **Racial Disparities**: SUSP_RACE is a top feature, especially dominant in misdemeanor predictions, suggesting race heavily influences outcomes. Black individuals face lower recall in felony and violation classifications, indicating potential under-policing or misclassification. Conversely, White individuals have higher recall in violations, possibly leading to over-policing.<br>- **Gender Influence**: SUSP_SEX is highly important in felony and violation predictions but less so in misdemeanors, showing gender's varied impact across offense types.<br><br>#### 3. Feature Influence Analysis<br>- **High-Importance Features**: SUSP_RACE and SUSP_SEX are critical, with SUSP_RACE being particularly influential in misdemeanors. These features may act as proxies for other demographics, raising fairness concerns.<br>- **Protected Attributes**: SUSP_RACE and SUSP_SEX are protected attributes, and their prominence in predictions suggests potential bias.<br><br>#### 4. Risk Assessment<br>- **Significant Risks**: The model's reliance on race and gender risks perpetuating biases in the criminal justice system, leading to unfair treatment and erosion of trust. This could result in real harm through misclassification and unequal enforcement. | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | ---<br><br>This analysis highlights the need for careful examination and adjustment to mitigate bias and ensure fairness in the model's decisions. | |
| **llama_3_3** | ## Bias Level Classification<br>Based on the global feature importance analysis, the bias level is classified as **MEDIUM**. This classification is due to the presence of moderate bias indicators, such as the high importance of demographic features like `SUSP_RACE` and `SUSP_SEX`, which could potentially serve as proxies for protected attributes. However, the importance scores do not indicate severe or critical bias levels that would suggest major fairness violations.<br><br>## Specific Bias Patterns<br>The analysis reveals specific patterns of unfair treatment across different groups, primarily driven by the varying importance of features like `SUSP_RACE` and `SUSP_SEX` across classes. For instance, `SUSP_RACE` is significantly more important for predicting `MISDEMEANOR` (83.8%) compared to `FELONY` and `VIOLATION`, where its importance is around 22.3%. This discrepancy suggests potential bias in how racial information is used in predictions for different offense types.<br><br>## Feature Influence Analysis<br>The feature influence analysis highlights the following key points:<br>- **Most Important Features**: `SUSP_SEX`, `SUSP_RACE`, `VIC_SEX`, `SUSP_AGE_GROUP`, and `VIC_AGE_GROUP` are among the top features influencing predictions, with `SUSP_RACE` being particularly dominant in | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | predicting `MISDEMEANOR`.<br>- **Demographic Proxies**: Features like `SUSP_RACE` and `SUSP_SEX` could act as proxies for protected attributes, potentially introducing bias into the predictions.<br>- **Bias in Importance Across Groups**: The importance of `SUSP_RACE` varies significantly across classes, being most important for `MISDEMEANOR` and less so for `FELONY` and `VIOLATION`. This variation indicates potential bias in how demographic information is utilized for different types of offenses.<br><br>## Risk Assessment<br>The detected biases pose a moderate risk, as they could lead to unfair treatment of individuals based on their demographic characteristics, particularly race. The significant importance of `SUSP_RACE` in predicting `MISDEMEANOR` suggests that individuals from certain racial backgrounds might be more likely to be predicted as committing misdemeanors, potentially due to biased patterns in the data or the model's reliance on demographic proxies. This risk necessitates careful examination and potential adjustment of the model to ensure fairness and equity in predictions across all groups. | |
| **mistral_nemo** | Okay, so I'm trying to analyze the bias in this machine learning model based on the provided data. The protected attribute here is SUSP_AGE_GROUP, which I assume refers to the age group of the suspect. The data includes precision, recall, and support metrics for different classes (FELONY, MISDEMEANOR, VIOLATION) across various race groups. There's also a feature importance analysis that shows which features are most influential in the model's predictions.<br><br>First, I need to understand the overall performance. Looking at the overall metrics, MISDEMEANOR has the highest recall, which means the model is good at predicting that class correctly. FELONY and VIOLATION have lower recalls, | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | especially VIOLATION, which is really low. This might indicate that the model struggles with those classes, possibly due to less data or inherent bias.<br><br>Now, looking at the group-wise data, I notice that for FELONY, the precision is higher for BLACK (0.581) compared to others, but the recall is lower. This could mean that when the model predicts FELONY for BLACK suspects, it's more accurate, but it doesn't predict it often. For MISDEMEANOR, BLACK has a lower precision but higher recall, which might mean the model overpredicts this class for BLACK individuals. VIOLATION has low precision and recall across all groups, which is concerning.<br><br>The feature importance shows that SUSP_RACE is the second most important feature overall, which is a red flag because it's a protected attribute. SUSP_SEX is the most important, which is also a protected attribute. This suggests the model might be relying too much on demographic factors rather than other, more relevant features.<br><br>Looking at each class specifically, for FELONY and VIOLATION, the top features are similar, with SUSP_RACE and SUSP_SEX leading. But for MISDEMEANOR, SUSP_RACE is by far the most important, which is problematic because it's a protected attribute and shouldn't be so influential.<br><br>I'm also noticing that SUSP_AGE_GROUP, the protected attribute we're focusing on, is the third most important overall and in FELONY and VIOLATION, but it's not as important in MISDEMEANOR. This could mean age is influencing predictions more in certain offense types.<br><br>There's a risk that the model is using race and sex as proxies for other factors, | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | which can lead to unfair treatment. For example, if the model is more likely to predict MISDEMEANOR for BLACK individuals because of their race, that's a bias. Also, the low recall for VIOLATION across all races, especially the complete lack of predictions for some groups, indicates potential neglect of certain cases.<br><br>I think the bias level here is medium to high because the model heavily relies on protected attributes, and there are significant disparities in how different race groups are treated, especially for MISDEMEANOR. The high importance of SUSP_RACE and SUSP_SEX is a clear indicator of potential bias. The specific patterns show that BLACK individuals are more likely to be misclassified for MISDEMEANOR, and certain groups like AMERICAN INDIAN/ALASKAN NATIVE are almost ignored in FELONY and VIOLATION predictions.<br><br>Overall, the model's reliance on protected attributes and the performance disparities across different groups suggest that there's a significant risk of unfair treatment, which needs to be addressed to ensure fairness and equity.<br><br><br>### Bias Analysis Report<br><br>#### 1. Bias Level Classification:<br>- **HIGH**: The model exhibits significant bias, particularly due to the high importance of protected attributes and performance disparities across groups.<br><br>#### 2. Specific Bias Patterns:<br>- **Race Influence**: SUSP_RACE is the second most important feature, indicating potential racial bias, especially in MISDEMEANOR predictions where it's overwhelmingly influential. | |

| LLM Model | Bias Analysis | Status |
|---|---|---|
| | - **Overprediction for BLACK Individuals**: The model overpredicts MISDEMEANOR for BLACK suspects, shown by high recall but lower precision, suggesting possible unfair treatment.<br>- **Neglect of Certain Groups**: Groups like AMERICAN INDIAN/ALASKAN NATIVE are underrepresented in FELONY and VIOLATION predictions, indicating neglect.<br><br>#### 3. Feature Influence Analysis:<br>- **High-Importance Features**: SUSP_SEX and SUSP_RACE are top features, raising concerns of demographic bias.<br>- **Protected Attributes as Proxies**: These features might act as proxies for other factors, leading to unfair predictions.<br>- **Class-Specific Bias**: SUSP_RACE's dominance in MISDEMEANOR suggests bias, while SUSP_AGE_GROUP's role varies across classes.<br><br>#### 4. Risk Assessment:<br>- **Unfair Treatment Risk**: High reliance on protected attributes risks systemic bias, affecting decision-making fairness.<br>- **Disparate Impact**: Performance disparities across race groups may lead to unequal treatment in legal outcomes.<br>- **Model Reliability**: Low recall in VIOLATION indicates potential failure in certain cases, undermining model reliability.<br><br>This analysis highlights the need for mitigation strategies to reduce bias and ensure equitable predictions. | |

## Group-wise Metrics

| SUSP_AGE_GROUP | Class | Precision | Recall |
|---|---|---|---|
| undefined | FELONY | 0.429 | 0.005 |
| undefined | MISDEMEANOR | 0.669 | 0.997 |
| undefined | VIOLATION | 0.000 | 0.000 |
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 1.000 | 1.000 |
| undefined | VIOLATION | 0.000 | 0.000 |
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 0.385 | 1.000 |
| undefined | VIOLATION | 0.000 | 0.000 |
| undefined | FELONY | 0.582 | 0.451 |
| undefined | MISDEMEANOR | 0.412 | 0.670 |
| undefined | VIOLATION | 0.446 | 0.224 |
| undefined | FELONY | 0.400 | 0.400 |
| undefined | MISDEMEANOR | 0.563 | 0.750 |
| undefined | VIOLATION | 0.462 | 0.286 |
| undefined | FELONY | 0.409 | 0.119 |

| SUSP_AGE_GROUP | Class | Precision | Recall |
|---|---|---|---|
| undefined | MISDEMEANOR | 0.457 | 0.894 |
| undefined | VIOLATION | 0.167 | 0.019 |
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 0.412 | 0.326 |
| undefined | VIOLATION | 0.479 | 0.680 |
| undefined | FELONY | 0.273 | 0.081 |
| undefined | MISDEMEANOR | 0.507 | 0.764 |
| undefined | VIOLATION | 0.490 | 0.357 |

# Analysis for Protected Attribute: SUSP_RACE

## Multi-LLM Bias Analysis

| LLM Model | Bias Analysis | Status |
|---|---|---|
| **deepseek_r1** | Okay, so I need to analyze the provided bias and classification analysis data for the protected attribute 'SUSP_RACE'. The user has given me a detailed breakdown of | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | precision, recall, bias metrics, and feature importance across different classes: FELONY, MISDEMEANOR, and VIOLATION. My task is to carefully go through this data and identify potential biases, focusing on feature importance, possible proxies, and class-specific issues.<br><br>First, I'll start by looking at the overall metrics. The bias level for FELONY is HIGH, MISDEMEANOR is LOW, and VIOLATION is HIGH. This tells me that FELONY and VIOLATION classes have significant bias issues, while MISDEMEANOR is within acceptable ranges. The statistical parity differences and disparate impact ratios indicate that certain groups are being treated unfairly, especially in FELONY cases where the disparate impact is nearly double, suggesting a significant bias against unprivileged groups.<br><br>Next, I'll examine the group-wise performance. For FELONY, the BLACK group has a recall of 0.450549, which is higher than some others, but I need to see if this is consistent across other metrics. The WHITE group in FELONY has zero recall, which is concerning. In MISDEMEANOR, BLACK HISPANIC has a high recall, but BLACK itself has a lower recall compared to others. VIOLATION shows that BLACK and WHITE HISPANIC have higher recalls than others, but the support is low for some groups, which might indicate issues with data representation.<br><br>Looking at the feature importance, SUSP_RACE is the second most important feature overall and in FELONY and VIOLATION, but it's the most important in MISDEMEANOR with 83.8% importance. This is a red flag because a protected attribute being so crucial in predictions can lead to direct bias. SUSP_SEX is also highly important, which might be a proxy for race if certain genders are disproportionately represented in specific racial groups. | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | I notice that in MISDEMEANOR cases, SUSP_RACE overshadows other features, which could mean the model is relying too much on race, leading to unfair outcomes. For VIOLATION, the importance of SUSP_RACE and SUSP_SEX together is quite high, suggesting potential compounding biases based on both race and gender.<br><br>The feature interactions are also concerning. Since SUSP_RACE is a protected attribute, its high importance indicates the model might be discrimininating based on race. The combination with SUSP_SEX could exacerbate this, creating intersectional biases, especially affecting certain demographics more than others.<br><br>In terms of risk assessment, the high importance of protected attributes and possible proxies poses a significant risk of unfair treatment, particularly in FELONY and VIOLATION cases. The model may be systematically favoring or disadvantaging specific racial groups, leading to real-world impacts like unequal access to justice or harsher penalties.<br><br>Overall, the analysis points to a HIGH bias level, especially in FELONY and VIOLATION classes, with SUSP_RACE being a major contributor. The model's reliance on this protected attribute and potential proxies like SUSP_SEX indicates a critical need for further investigation and possible model adjustments to ensure fairness.<br><br><br>### Bias Analysis for Protected Attribute 'SUSP_RACE'<br><br>#### 1. Bias Level Classification:<br>- **HIGH**: Significant bias detected in FELONY and VIOLATION classes, with the | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | model showing clear unfair treatment across different racial groups.<br>MISDEMEANOR class shows minimal bias.<br><br>#### 2. Specific Bias Patterns:<br>- **FELONY Class**: High disparate impact ratio (1.914) indicates significant bias against unprivileged groups. Certain races, like BLACK and BLACK HISPANIC, have lower recall, suggesting unfair treatment.<br>- **VIOLATION Class**: High statistical parity difference (-0.120) points to bias against certain groups, with some races having zero recall.<br>- **MISDEMEANOR Class**: Despite low bias metrics, reliance on SUSP_RACE is concerning, indicating potential unfair treatment.<br><br>#### 3. Feature Influence Analysis:<br>- **Highest Importance**: SUSP_RACE is the second most important feature overall and in FELONY and VIOLATION, and most important in MISDEMEANOR (83.8%). SUSP_SEX is also highly influential.<br>- **Proxies**: SUSP_SEX may act as a proxy for race, potentially leading to intersectional biases.<br>- **Class-Specific Bias**: SUSP_RACE's dominance in MISDEMEANOR suggests over-reliance on race, indicating potential bias.<br><br>#### 4. Risk Assessment:<br>- **FELONY and VIOLATION**: High risk of unfair treatment due to model reliance on SUSP_RACE and SUSP_SEX, leading to potential real-world impacts like unequal justice.<br>- **MISDEMEANOR**: Despite low bias metrics, high reliance on SUSP_RACE poses risks of unfair outcomes. | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | ### Conclusion:<br>The model exhibits significant bias, particularly in FELONY and VIOLATION classes, with SUSP_RACE being a major contributor. The high importance of protected attributes and potential proxies indicates a critical need for further investigation to ensure fairness. | |
| **llama_3_3** | ## Bias Level Classification<br>Based on the provided data, the bias level classification is **HIGH**. This is due to the significant disparity in importance scores for the `SUSP_RACE` feature across different classes, as well as the high statistical parity difference and disparate impact ratios for certain classes, such as FELONY and VIOLATION.<br><br>## Specific Bias Patterns<br>The data reveals specific patterns of unfair treatment across different groups. For instance:<br>- The `SUSP_RACE` feature has a significantly higher importance score for the MISDEMEANOR class (83.8%) compared to other classes, indicating potential bias in misdemeanor predictions based on race.<br>- The importance of `SUSP_RACE` varies substantially across classes, suggesting that race may be influencing predictions differently depending on the type of offense.<br>- Certain racial groups have notably different precision and recall rates for specific classes, such as the high recall rate for MISDEMEANOR among BLACK individuals (0.670330) compared to other groups.<br><br>## Feature Influence Analysis<br>The feature influence analysis highlights the following points:<br>- **Most Important Features**: The top features across all classes include | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | `SUSP_SEX`, `SUSP_RACE`, `VIC_SEX`, `SUSP_AGE_GROUP`, and `VIC_AGE_GROUP`. `SUSP_RACE` is particularly important for the MISDEMEANOR class.<br>- **Demographic Proxies**: Features like `SUSP_SEX` and `SUSP_AGE_GROUP` could potentially serve as demographic proxies, influencing predictions in a way that might perpetuate biases.<br>- **Bias in Importance Across Groups**: The importance of `SUSP_RACE` varies significantly across different classes, indicating potential bias in how race influences predictions for different types of offenses.<br><br>## Risk Assessment<br>The detected biases pose a significant risk of unfair treatment in the prediction of offenses based on race. The high importance of `SUSP_RACE` in certain classes, combined with the disparity in statistical metrics across racial groups, suggests that the system may be perpetuating existing biases. This could lead to discriminatory outcomes, where individuals from certain racial backgrounds are more likely to be misclassified or receive harsher predictions than others. The potential impact includes:<br>- Discriminatory treatment of individuals based on race.<br>- Erosion of trust in the justice system due to perceived bias.<br>- Potential legal and reputational risks associated with biased decision-making processes. | |
| **mistral_nemo** | Okay, so I'm trying to figure out how to analyze the bias in this dataset based on the provided information. Let me start by understanding what each part means.<br><br>First, the user provided a detailed breakdown of the bias and classification analysis for the protected attribute 'SUSP_RACE'. They also included feature importance data for different classes: Overall, FELONY, MISDEMEANOR, and VIOLATION. My | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | task is to analyze this data to detect and explain any biases present.<br><br>I'll begin by looking at the overall metrics. The precision and recall for each class vary. FELONY has lower recall, which might mean it's harder to detect felonies correctly. MISDEMEANOR has high recall, so it's being predicted more accurately. VIOLATION has low recall, similar to FELONY.<br><br>Looking at the group-wise analysis, some groups have very low support, like AMERICAN INDIAN/ALASKAN NATIVE with only 2 cases in MISDEMEANOR. That's a small sample size, which might not be reliable. The BLACK group has higher support but lower recall in FELONY and VIOLATION, which could indicate underprediction for these classes.<br><br>The Bias Metrics table shows that FELONY and VIOLATION have HIGH bias levels, while MISDEMEANOR is LOW. The statistical parity difference and disparate impact ratios suggest that unprivileged groups might be treated more harshly in FELONY and VIOLATION cases.<br><br>Moving to the feature importance analysis, SUSP_RACE is the second most important feature overall and in FELONY and VIOLATION, but it's the most important in MISDEMEANOR. SUSP_SEX is also highly important across all classes. This makes me think that these features might be influencing predictions significantly, possibly introducing bias.<br><br>I should check if any features could be proxies for SUSP_RACE. SUSP_SEX and SUSP_AGE_GROUP might be correlated with race, acting as proxies. For example, if certain age groups are more represented in specific races, the model might be using age as a stand-in for race. | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | Looking at the class-specific patterns, MISDEMEANOR relies heavily on SUSP_RACE, which is concerning because it might mean race is a major factor in those predictions. FELONY and VIOLATION have more balanced feature importance, but SUSP_RACE is still a top factor.<br><br>High-importance features like SUSP_RACE and SUSP_SEX need scrutiny. Their high influence could mean the model is biased towards certain groups. For example, if being Black affects the prediction of FELONY more than other factors, that's a problem.<br><br>The risk assessment is crucial. High bias levels in FELONY and VIOLATION could lead to unfair treatment of certain racial groups, perpetuating discrimination. This can have serious real-world consequences, like wrongful arrests or harsher penalties.<br><br>I also notice that some groups, like AMERICAN INDIAN/ALASKAN NATIVE, have very low support, which might make the model less accurate for them. This underrepresentation can exacerbate biases.<br><br>In summary, the analysis points to significant bias in FELONY and VIOLATION classes, with SUSP_RACE and SUSP_SEX being key factors. These could be proxies for race, leading to unfair treatment of certain groups. The model's reliance on these features raises ethical concerns and potential legal issues.<br><br>### Bias Analysis Report for Protected Attribute 'SUSP_RACE' | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | #### 1. Bias Level Classification<br>- **FELONY**: HIGH bias detected due to a significant statistical parity difference (0.158215) and a disparate impact ratio of 1.914897, indicating potential unfair treatment.<br>- **MISDEMEANOR**: LOW bias detected, as the statistical parity difference (-0.038190) and disparate impact ratio (0.906803) are within acceptable ranges.<br>- **VIOLATION**: HIGH bias detected with a notable statistical parity difference (-0.120025) and a disparate impact ratio of 0.712371, suggesting unfair treatment.<br><br>#### 2. Specific Bias Patterns<br>- **FELONY Class**: The BLACK group has a lower recall (0.450549) compared to other groups, indicating potential underprediction for this group.<br>- **VIOLATION Class**: The BLACK group again shows a lower recall (0.224490), suggesting potential unfair treatment.<br>- **MISDEMEANOR Class**: The BLACK HISPANIC group has a higher recall (0.750000) compared to other groups, indicating overprediction.<br><br>#### 3. Feature Influence Analysis<br>- **Highest Importance Features**: SUSP_RACE and SUSP_SEX are consistently among the top features across all classes, with SUSP_RACE being the most important in MISDEMEANOR (83.8% importance).<br>- **Potential Proxies**: SUSP_SEX and SUSP_AGE_GROUP may act as proxies for SUSP_RACE, as demographic factors often correlate with race.<br>- **Class-Specific Bias**: SUSP_RACE is highly influential in MISDEMEANOR predictions, raising concerns about racial bias in this class.<br><br>#### 4. Risk Assessment<br>- The high bias levels in FELONY and VIOLATION classes pose significant risks, | |

| LLM Model | Bias Analysis | Status |
|---|---|---|
| | potentially leading to unfair treatment of certain racial groups. The reliance on SUSP_RACE and SUSP_SEX as key features increases the likelihood of biased outcomes, affecting fairness and equity in decision-making processes. | |

## Group-wise Metrics

| SUSP_RACE | Class | Precision | Recall |
|---|---|---|---|
| (null) | FELONY | 0.429 | 0.005 |
| (null) | MISDEMEANOR | 0.669 | 0.997 |
| (null) | VIOLATION | 0.000 | 0.000 |
| AMERICAN INDIAN/ALASKAN NATIVE | FELONY | 0.000 | 0.000 |
| AMERICAN INDIAN/ALASKAN NATIVE | MISDEMEANOR | 1.000 | 1.000 |
| AMERICAN INDIAN/ALASKAN NATIVE | VIOLATION | 0.000 | 0.000 |
| ASIAN / PACIFIC ISLANDER | FELONY | 0.000 | 0.000 |
| ASIAN / PACIFIC ISLANDER | MISDEMEANOR | 0.385 | 1.000 |
| ASIAN / PACIFIC ISLANDER | VIOLATION | 0.000 | 0.000 |
| BLACK | FELONY | 0.582 | 0.451 |
| BLACK | MISDEMEANOR | 0.412 | 0.670 |

| SUSP_RACE | Class | Precision | Recall |
|---|---|---|---|
| BLACK | VIOLATION | 0.446 | 0.224 |
| BLACK HISPANIC | FELONY | 0.400 | 0.400 |
| BLACK HISPANIC | MISDEMEANOR | 0.563 | 0.750 |
| BLACK HISPANIC | VIOLATION | 0.462 | 0.286 |
| UNKNOWN | FELONY | 0.409 | 0.119 |
| UNKNOWN | MISDEMEANOR | 0.457 | 0.894 |
| UNKNOWN | VIOLATION | 0.167 | 0.019 |
| WHITE | FELONY | 0.000 | 0.000 |
| WHITE | MISDEMEANOR | 0.412 | 0.326 |
| WHITE | VIOLATION | 0.479 | 0.680 |
| WHITE HISPANIC | FELONY | 0.273 | 0.081 |
| WHITE HISPANIC | MISDEMEANOR | 0.507 | 0.764 |
| WHITE HISPANIC | VIOLATION | 0.490 | 0.357 |

## Bias Metrics

| Category | Privileged Rate | Unprivileged Rate | Statistical Parity Difference | Disparate Impact | Mean Difference |
|---|---|---|---|---|---|
| FELONY | 0.173 | 0.331 | 0.158 | 1.915 | 0.158 |
| MISDEMEANOR | 0.410 | 0.372 | -0.038 | 0.907 | -0.038 |
| VIOLATION | 0.417 | 0.297 | -0.120 | 0.712 | -0.120 |

# Analysis for Protected Attribute: SUSP_SEX

## Multi-LLM Bias Analysis

| LLM Model | Bias Analysis | Status |
|---|---|---|
| **deepseek_r1** | Okay, so I need to analyze the bias in a model based on the provided data. The protected attribute here is SUSP_SEX, which I assume stands for suspect sex. The analysis includes overall metrics, group-wise data, and feature importance across different classes: FELONY, MISDEMEANOR, and VIOLATION.<br><br>First, looking at the overall precision and recall, I see that for FELONY, precision is about 50% and recall is low at around 12%. For MISDEMEANOR, both precision and recall are higher, around 60% and 92% respectively. VIOLATION has lower precision and recall, around 46% and 27%. This suggests that the model is better at predicting MISDEMEANOR cases but struggles with FELONY and VIOLATION. | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | Now, the group-wise analysis shows how different races perform within each class. For example, in FELONY, the precision for BLACK is around 58% and recall is 45%, which is better than some other groups. However, for VIOLATION, BLACK has a precision of 45% and a very low recall of 22%. The WHITE group in VIOLATION has a higher recall of 68%, which might indicate some disparity.<br><br>Looking at feature importance, SUSP_SEX is the second most important feature overall at 29.2%, and it's also high in FELONY and VIOLATION. SUSP_RACE is the most important overall and in FELONY and VIOLATION, but in MISDEMEANOR, SUSP_RACE is even more dominant at 83.8%. This makes me think that race is a strong predictor in MISDEMEANOR cases, which could be a bias indicator.<br><br>The fact that SUSP_SEX and SUSP_RACE are both in the top features suggests they might be demographic proxies. Their high importance could mean the model is relying heavily on these protected attributes, leading to potential bias. For instance, in MISDEMEANOR, SUSP_RACE's high importance might indicate that race is heavily influencing those predictions, which isn't fair.<br><br>I also notice that for MISDEMEANOR, the model's recall for BLACK is 67%, while for WHITE it's 33%. That's a significant difference. Similarly, in VIOLATION, BLACK has a lower recall compared to WHITE. This could mean the model is under-predicting violations for BLACK individuals, which is a concern.<br><br>The feature importance varies a lot between classes. In MISDEMEANOR, SUSP_RACE is overwhelmingly important, whereas in FELONY and VIOLATION, both SUSP_SEX and SUSP_RACE are top features. This inconsistency could indicate that the model's behavior changes across different offense types, possibly leading | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | to biased outcomes in specific cases.<br><br>Since both SUSP_SEX and SUSP_RACE are protected attributes, their high importance is a red flag. They might be directly influencing predictions, leading to unfair treatment of certain groups. The model seems to rely more on these attributes in some classes more than others, which could result in systemic biases.<br><br>Overall, the model shows significant reliance on protected attributes, especially race, which varies across different offense types. This points to a high level of bias, potentially leading to unfair outcomes for certain racial and gender groups.<br><br><br>### Bias Analysis Report<br><br>#### 1. Bias Level Classification:<br>- **HIGH**: The model exhibits significant bias, particularly in the MISDEMEANOR class where SUSP_RACE has an importance of 83.8%, indicating a strong reliance on a protected attribute. This suggests potential unfair treatment across different demographic groups.<br><br>#### 2. Specific Bias Patterns:<br>- **Racial Disparities**: In MISDEMEANOR cases, the recall for BLACK individuals is 67%, compared to 33% for WHITE individuals, showing a notable disparity.<br>- **Under-prediction in VIOLATION**: BLACK individuals have a lower recall (22%) in VIOLATION cases compared to WHITE individuals (68%), indicating potential under-prediction.<br><br>#### 3. Feature Influence Analysis: | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | - **High-Importance Features**: SUSP_RACE and SUSP_SEX are consistently among the top features, with SUSP_RACE being dominant in MISDEMEANOR.<br>- **Proxies for Protected Attributes**: The high importance of SUSP_RACE and SUSP_SEX suggests they may act as proxies, influencing predictions unfairly.<br>- **Class-Specific Bias**: The model's reliance on SUSP_RACE varies across classes, with significant influence in MISDEMEANOR, indicating potential bias in that category.<br><br>#### 4. Risk Assessment:<br>- The model's reliance on protected attributes poses a high risk of unfair outcomes, particularly affecting certain racial and gender groups disproportionately. This could lead to systemic biases in legal proceedings, undermining fairness. | |
| **llama_3_3** | ## Bias Analysis Report<br>### Introduction<br>This report provides a comprehensive analysis of bias in the given dataset, focusing on the protected attribute 'SUSP_SEX'. The analysis is based on the global feature importance data provided.<br><br>### 1. Bias Level Classification<br>Based on the feature importance scores and patterns observed, the bias level in this dataset is classified as **MEDIUM**. While there are some concerning patterns, such as the high importance of 'SUSP_SEX' and 'SUSP_RACE' across classes, the metrics do not indicate severe or critical bias.<br><br>### 2. Specific Bias Patterns<br>The analysis reveals the following specific patterns of unfair treatment: | Success |

| LLM Model | Bias Analysis | Status |
|---|---|---|
| | - **Consistent Importance of Protected Attributes**: 'SUSP_SEX' and 'SUSP_RACE' are among the top 5 most important features across classes, indicating potential bias.<br>- **Class-Specific Bias**: The importance of 'SUSP_RACE' is significantly higher for the 'MISDEMEANOR' class (83.8%) compared to other classes, suggesting potential bias in predicting misdemeanors based on race.<br>- **Demographic Proxies**: Features like 'SUSP_RACE' and 'VIC_AGE_GROUP' could be acting as demographic proxies, potentially leading to biased predictions.<br><br>### 3. Feature Influence Analysis<br>The feature influence analysis reveals the following:<br>- **Most Important Features**: 'SUSP_SEX', 'SUSP_RACE', 'VIC_SEX', 'SUSP_AGE_GROUP', and 'VIC_AGE_GROUP' are the top 5 most important features across classes.<br>- **Potential Demographic Proxies**: 'SUSP_RACE' and 'VIC_AGE_GROUP' might be proxies for protected attributes, as they are highly important for predicting certain classes.<br>- **Bias in Feature Importance**: The importance of 'SUSP_RACE' varies significantly across classes, with a high importance for 'MISDEMEANOR' and relatively lower importance for 'FELONY' and 'VIOLATION'.<br><br>### 4. Risk Assessment<br>The detected biases pose a moderate risk to fairness in the predictions. The high importance of protected attributes and potential demographic proxies could lead to unfair treatment of certain groups. Specifically, the class-specific bias in predicting misdemeanors based on race could result in disproportionate outcomes for individuals from certain racial backgrounds. The potential impact of these biases includes: | |

| LLM Model | Bias Analysis | Status |
|---|---|---|
| | - **Unfair Treatment**: Biased predictions could lead to unfair treatment of individuals based on | |
| **mistral_nemo** | Okay, so I need to figure out how to approach this bias and fairness analysis based on the data provided. Let me start by understanding the problem. The protected attribute here is SUSP_SEX, which I assume stands for suspect sex. The analysis includes overall and group-wise performance metrics for three classes: FELONY, MISDEMEANOR, and VIOLATION. There's also a feature importance analysis that shows which features are most influential in the model's predictions.<br><br>First, I should look at the overall metrics. The precision and recall vary across the classes. For FELONY, precision is about 50% and recall is low at around 12%. For MISDEMEANOR, both precision and recall are higher, with recall being 91.8%. VIOLATION has lower precision and recall. This might indicate that the model is better at predicting MISDEMEANOR cases but struggles with FELONY and VIOLATION.<br><br>Next, looking at the group-wise data. For SUSP_RACE, there are significant differences. For example, in FELONY cases, the precision is higher for BLACK (58%) compared to others, but recall is lower. For MISDEMEANOR, BLACK has lower precision (41%) but higher recall (67%). The support varies a lot, with some groups having very small numbers, like AMERICAN INDIAN/ALASKAN NATIVE with only 2 cases in MISDEMEANOR. This could mean the model might not generalize well for smaller groups.<br><br>Now, the feature importance analysis shows SUSP_SEX is the second most important feature overall and in FELONY and VIOLATION, but less important in MISDEMEANOR. SUSP_RACE is the most important overall and in MISDEMEANOR. | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | VIC_SEX and age groups also feature in the top 5. This suggests that sex and race are significant factors in the model's predictions, which might indicate potential bias.<br><br>I need to check if these features could be proxies for protected attributes. Since SUSP_SEX is a protected attribute itself, its high importance is a red flag. SUSP_RACE is also a protected attribute, so relying heavily on it could lead to biased outcomes. Other features like VIC_SEX might also relate to sex, indirectly influencing the protected attribute.<br><br>Looking at class-specific patterns, SUSP_RACE is overwhelmingly important for MISDEMEANOR (83.8%), which is concerning. This could mean the model is making predictions based heavily on race for that class, leading to unfair treatment of certain groups.<br><br>The high-importance features (above 15%) are SUSP_SEX and SUSP_RACE. These are both protected attributes, so their prominence is a significant bias indicator. Their influence across different classes varies, but they're consistently in the top features, which isn't good for fairness.<br><br>The feature interactions are also a point. Since SUSP_SEX and SUSP_RACE are both top features, they might be interacting in ways that unfairly affect certain groups. For example, being in a particular race and sex could compound biases.<br><br>Now, for the bias level classification, considering that protected attributes are highly influential and there are significant differences in performance across groups, especially in MISDEMEANOR where race is extremely important, I think the bias level is HIGH. There's clear unfair treatment, especially towards BLACK | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | individuals in FELONY cases and the heavy reliance on race in MISDEMEANOR.<br><br>Specific bias patterns include lower recall for BLACK in FELONY and VIOLATION, meaning they might be less likely to be correctly classified. Also, smaller groups like AMERICAN INDIAN/ALASKAN NATIVE have very low support, which can lead to poor model performance for them.<br><br>In terms of feature influence, SUSP_SEX and SUSP_RACE are the most important, and they're protected attributes. VIC_SEX might be acting as a proxy for SUSP_SEX, which is another concern. The model's reliance on these features could perpetuate existing biases.<br><br>The risk assessment is that these biases could lead to unfair treatment in legal outcomes, with certain groups being disproportionately affected. This can undermine trust in the system and lead to real-world harm.<br><br>### Bias Analysis Report<br><br>#### 1. Bias Level Classification:<br>- **HIGH**: Significant bias detected, indicating clear unfair treatment across different groups, particularly influenced by protected attributes.<br><br>#### 2. Specific Bias Patterns:<br>- **Racial Disparities**: The model shows varying performance across different races. For instance, BLACK individuals have lower recall in FELONY and VIOLATION cases, suggesting they might be less likely to be correctly classified.<br>- **Class-Specific Bias**: In MISDEMEANOR cases, SUSP_RACE has an importance | |

| LLM Model | Bias Analysis | Status |
|---|---|---|
| | of 83.8%, indicating potential over-reliance on race for predictions, which could lead to unfair treatment.<br>- **Small Group Neglect**: Groups like AMERICAN INDIAN/ALASKAN NATIVE have minimal support, risking poor model performance and neglect of these groups.<br><br>#### 3. Feature Influence Analysis:<br>- **High-Importance Features**: SUSP_SEX and SUSP_RACE are the most influential, both being protected attributes. Their prominence raises concerns about direct bias.<br>- **Proxy Features**: VIC_SEX might act as a proxy for SUSP_SEX, indirectly influencing bias.<br>- **Feature Interactions**: The interaction between SUSP_SEX and SUSP_RACE could compound biases, affecting certain groups disproportionately.<br><br>#### 4. Risk Assessment:<br>- The model's reliance on protected attributes poses a high risk of perpetuating biases, leading to unfair legal outcomes. This can erode trust and cause real-world harm to affected groups. | |

## Group-wise Metrics

| SUSP_SEX | Class | Precision | Recall |
|---|---|---|---|
| undefined | FELONY | 0.429 | 0.005 |
| undefined | MISDEMEANOR | 0.669 | 0.997 |

| SUSP_SEX | Class | Precision | Recall |
|---|---|---|---|
| undefined | VIOLATION | 0.000 | 0.000 |
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 1.000 | 1.000 |
| undefined | VIOLATION | 0.000 | 0.000 |
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 0.385 | 1.000 |
| undefined | VIOLATION | 0.000 | 0.000 |
| undefined | FELONY | 0.582 | 0.451 |
| undefined | MISDEMEANOR | 0.412 | 0.670 |
| undefined | VIOLATION | 0.446 | 0.224 |
| undefined | FELONY | 0.400 | 0.400 |
| undefined | MISDEMEANOR | 0.563 | 0.750 |
| undefined | VIOLATION | 0.462 | 0.286 |
| undefined | FELONY | 0.409 | 0.119 |
| undefined | MISDEMEANOR | 0.457 | 0.894 |
| undefined | VIOLATION | 0.167 | 0.019 |

| SUSP_SEX | Class | Precision | Recall |
|---|---|---|---|
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 0.412 | 0.326 |
| undefined | VIOLATION | 0.479 | 0.680 |
| undefined | FELONY | 0.273 | 0.081 |
| undefined | MISDEMEANOR | 0.507 | 0.764 |
| undefined | VIOLATION | 0.490 | 0.357 |

# Analysis for Protected Attribute: VIC_AGE_GROUP

## Multi-LLM Bias Analysis

| LLM Model | Bias Analysis | Status |
|---|---|---|
| **deepseek_r1** | Okay, so I need to analyze the bias in this dataset based on the provided feature importance and classification metrics. The protected attribute here is VIC_AGE_GROUP, but I also see other demographic features like SUSP_RACE, SUSP_SEX, and VIC_SEX. | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | First, looking at the overall feature importance, SUSP_RACE is the second most important feature at 22.3%. That's pretty high and might indicate potential bias since race shouldn't be a significant factor in legal classifications. Similarly, SUSP_SEX is the top feature, which is also a protected attribute. This makes me think that the model might be relying too much on demographic factors.<br><br>Breaking it down by class, for FELONY, SUSP_RACE is the second most important, which is concerning because it could mean race is influencing felony charges. For MISDEMEANOR, SUSP_RACE is even more dominant at 83.8%, which is really high and suggests a strong bias. VIOLATION has SUSP_RACE as the second most important, so same issue there.<br><br>I also notice that VIC_AGE_GROUP is in the top features for all classes, which is the protected attribute we're focusing on. It's less important than SUSP_RACE and SUSP_SEX, but still, its presence might mean age is being used in predictions, which could be problematic if it's a proxy for other factors.<br><br>Looking at the group-wise metrics, there's a lot of variation. For example, in MISDEMEANOR, the precision is much higher for some race groups than others. Black and Black Hispanic groups have lower precision compared to others, which might indicate they're being misclassified more often. Also, the support varies a lot, with some groups having very small numbers, which can lead to biased models.<br><br>The feature importance shows that SUSP_RACE and SUSP_SEX are consistently important across all classes, which is a red flag. These shouldn't be so influential in legal decisions. The high importance of these features suggests the model might be unfair, especially towards certain race groups. | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | So, putting this together, the model seems to have a high bias level because it's heavily relying on protected attributes, leading to different outcomes for different demographic groups. This could result in unfair treatment in how crimes are classified, affecting certain communities more harshly.<br><br>### Bias Analysis Report<br><br>#### 1. Bias Level Classification:<br>- **HIGH**: Significant bias detected. The model exhibits concerning patterns where protected attributes like SUSP_RACE and SUSP_SEX are highly influential, leading to potential unfair treatment across different demographic groups.<br><br>#### 2. Specific Bias Patterns:<br>- **Racial Disparities**: SUSP_RACE is a top feature across all classes, particularly dominant in MISDEMEANOR cases, suggesting potential racial bias in classifications.<br>- **Sex Bias**: SUSP_SEX is the most important feature overall and across FELONY and VIOLATION classes, indicating possible sex-based disparities.<br>- **Age Group Influence**: VIC_AGE_GROUP, a protected attribute, is consistently among top features, raising concerns about age-related bias.<br><br>#### 3. Feature Influence Analysis:<br>- **High-Importance Features**: SUSP_RACE and SUSP_SEX are consistently top features, indicating their significant influence on model predictions.<br>- **Proxy Indicators**: SUSP_RACE and SUSP_SEX may act as proxies for other demographic factors, potentially leading to biased outcomes. | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | - **Group Bias**: Metrics show varying precision across race groups, with Black and Black Hispanic groups having lower precision in MISDEMEANOR cases, suggesting misclassification issues.<br><br>#### 4. Risk Assessment:<br>- The reliance on protected attributes poses a high risk of unfair treatment, particularly affecting certain racial and demographic groups. This could lead to disproportionate legal outcomes and reinforce systemic biases. | |
| **llama_3_3** | ### Bias Analysis Report<br><br>#### 1. Bias Level Classification<br>Based on the provided global feature importance analysis, the bias level classification is **MEDIUM**. This classification is due to the presence of demographic indicators among the top features, which could potentially act as proxies for protected attributes, and the varying importance of features across different classes.<br><br>#### 2. Specific Bias Patterns<br>The analysis reveals specific patterns of unfair treatment across different groups, particularly in how certain features are weighted more heavily for predictions in specific classes. For instance:<br>- **SUSP_RACE** is highly important for **MISDEMEANOR** predictions, suggesting a potential bias in how race influences misdemeanor charges.<br>- The consistent high importance of **SUSP_SEX** and **SUSP_RACE** across **FELONY** and **VIOLATION** classes indicates that these demographic features may be influencing the prediction of these classes disproportionately.<br>- The feature **VIC_AGE_GROUP**, which is the protected attribute being | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | analyzed, shows a relatively low importance score across all classes, but its interaction with other features (like **SUSP_AGE_GROUP**) could still contribute to biased outcomes.<br><br>#### 3. Feature Influence Analysis<br>- **Most Important Features**: **SUSP_SEX**, **SUSP_RACE**, **VIC_SEX**, **SUSP_AGE_GROUP**, and **VIC_AGE_GROUP** are among the top features influencing predictions. **SUSP_RACE** stands out, especially for **MISDEMEANOR** predictions, with an importance score of 83.8%.<br>- **Demographic Proxies**: Features like **SUSP_RACE** and **SUSP_SEX** could act as proxies for protected attributes, potentially leading to biased predictions.<br>- **Bias in Importance Across Groups**: The importance of **SUSP_RACE** varies significantly across classes, being most crucial for **MISDEMEANOR** predictions. This disparity suggests that the model may treat different racial groups unevenly depending on the class of the offense.<br><br>#### 4. Risk Assessment<br>The detected biases pose a significant risk to fairness in the prediction system. The high importance of demographic features like **SUSP_RACE** and **SUSP_SEX** in predicting offense classes could lead to disproportionate outcomes for certain groups, potentially perpetuating existing social inequalities. The medium bias level indicates a need for careful examination and potential adjustment of the model to mitigate these risks and ensure fair treatment across all groups. | |
| **mistral_nemo** | Alright, so I'm trying to analyze the bias in this dataset, focusing on the protected attribute VIC_AGE_GROUP. Let me start by going through the provided data and the questions I need to address. | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | First, I see that the overall metrics for each class (Felony, Misdemeanor, Violation) vary quite a bit. Felony has low recall, which might mean it's not being predicted often, while Misdemeanor has high recall but lower precision. Violation's metrics are also low. This could indicate that the model is better at predicting Misdemeanors than the other classes.<br><br>Looking at the group-wise analysis, I notice that for SUSP_RACE, most groups have low recall for Felony and Violation, except for Black and White Hispanic groups. For Misdemeanor, most have high recall. This might mean that the model is more accurate for Misdemeanors across races but struggles with Felonies and Violations, which could hint at bias in how different races are treated in the model's predictions.<br><br>Now, moving to the feature importance analysis. SUSP_SEX is the top feature overall and for Felony and Violation, but not for Misdemeanor, where SUSP_RACE dominates. VIC_AGE_GROUP is in the top 5 for overall and Felony but not as crucial for Misdemeanor. This suggests that the model relies heavily on race in predicting Misdemeanors, which is a protected attribute and could be a bias indicator.<br><br>I'm concerned that SUSP_RACE is a protected attribute and is highly important, especially in the Misdemeanor class. This could mean the model is using race as a proxy for other factors, leading to unfair treatment. Also, SUSP_SEX being a top feature across classes might indicate gender bias, even though it's not a protected attribute here, but it's still a demographic factor.<br><br>The fact that VIC_AGE_GROUP is important in some classes but not others shows that age might be influencing predictions differently depending on the offense | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | type. However, since it's a protected attribute, its influence should be scrutinized.<br><br>I'm also noticing that for Misdemeanors, SUSP_RACE has a very high importance (83.8%), which is way higher than other features. This could mean the model is heavily reliant on race when predicting Misdemeanors, which is a red flag for bias. The same applies to SUSP_SEX in other classes.<br><br>The risk here is that the model might be treating different demographic groups unfairly, especially in Misdemeanor cases where race plays a huge role. This could lead to discrimination in how individuals are classified based on their race and age group.<br><br>So, putting it all together, I think the bias level is medium because while there are concerning patterns, it's not the highest severity. The specific patterns show that race is a major factor in Misdemeanor predictions, and gender is influential in others. Features like SUSP_RACE and SUSP_SEX are acting as proxies for protected attributes, leading to potential unfair treatment. The risk is significant because it can result in biased outcomes affecting different demographic groups differently.<br><br>### Bias Analysis for Protected Attribute VIC_AGE_GROUP<br><br>#### 1. **Bias Level Classification**<br>- **MEDIUM**: Moderate bias detected, with concerning patterns in feature importance and class-specific predictions.<br><br>#### 2. **Specific Bias Patterns**<br>- **Race as a Predictor**: SUSP_RACE is highly influential, especially in | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | Misdemeanor predictions, suggesting potential racial bias.<br>- **Gender Influence**: SUSP_SEX is a top feature across classes, indicating possible gender-based bias.<br>- **Age Group Variability**: VIC_AGE_GROUP's varying importance across classes may lead to age-related disparities.<br><br>#### 3. **Feature Influence Analysis**<br>- **High-Importance Features**: SUSP_RACE and SUSP_SEX are consistently top features, raising concerns as proxies for protected attributes.<br>- **Protected Attributes**: SUSP_RACE is a protected attribute with significant influence, particularly in Misdemeanor cases.<br>- **Class-Specific Bias**: SUSP_RACE dominates Misdemeanor predictions, while SUSP_SEX is more influential in Felony and Violation.<br><br>#### 4. **Risk Assessment**<br>- **Potential Impact**: The model may unfairly treat individuals based on race and gender, leading to biased outcomes in legal classifications.<br>- **Consequences**: This could result in discrimination, particularly affecting certain demographic groups in Misdemeanor cases.<br><br>This analysis highlights the need for careful examination of feature influence and potential biases, especially regarding race and gender, to ensure fair and unbiased predictions. | |

## Group-wise Metrics

| VIC_AGE_GROUP | Class | Precision | Recall |
|---|---|---|---|
| undefined | FELONY | 0.429 | 0.005 |
| undefined | MISDEMEANOR | 0.669 | 0.997 |
| undefined | VIOLATION | 0.000 | 0.000 |
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 1.000 | 1.000 |
| undefined | VIOLATION | 0.000 | 0.000 |
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 0.385 | 1.000 |
| undefined | VIOLATION | 0.000 | 0.000 |
| undefined | FELONY | 0.582 | 0.451 |
| undefined | MISDEMEANOR | 0.412 | 0.670 |
| undefined | VIOLATION | 0.446 | 0.224 |
| undefined | FELONY | 0.400 | 0.400 |
| undefined | MISDEMEANOR | 0.563 | 0.750 |
| undefined | VIOLATION | 0.462 | 0.286 |
| undefined | FELONY | 0.409 | 0.119 |

| VIC_AGE_GROUP | Class | Precision | Recall |
|---|---|---|---|
| undefined | MISDEMEANOR | 0.457 | 0.894 |
| undefined | VIOLATION | 0.167 | 0.019 |
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 0.412 | 0.326 |
| undefined | VIOLATION | 0.479 | 0.680 |
| undefined | FELONY | 0.273 | 0.081 |
| undefined | MISDEMEANOR | 0.507 | 0.764 |
| undefined | VIOLATION | 0.490 | 0.357 |

# Analysis for Protected Attribute: VIC_RACE

## Multi-LLM Bias Analysis

| LLM Model | Bias Analysis | Status |
|---|---|---|
| **deepseek_r1** | Okay, so I'm trying to figure out how to analyze the bias in this dataset. Let me start by looking at the feature importance and the protected attribute, which is | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | VIC_RACE. The analysis provided shows the importance of different features across three classes: FELONY, MISDEMEANOR, and VIOLATION.<br><br>First, I notice that SUSP_RACE is the second most important feature overall and also in FELONY and VIOLATION classes. That's concerning because it's a protected attribute, and if it's highly influential, it might indicate bias. Then there's SUSP_SEX, which is the most important in FELONY and VIOLATION but less so in MISDEMEANOR. VIC_SEX and age groups are also up there, which are related to demographics.<br><br>Looking at the class-wise breakdown, for MISDEMEANOR, SUSP_RACE has a very high importance of 83.8%, which is way higher than other classes. This makes me think that race is a major factor in predicting misdemeanors, which could mean biased outcomes for certain groups. For example, the Black group has lower precision and recall in FELONY but higher in MISDEMEANOR. Maybe they're being misclassified more often into misdemeanors.<br><br>I also see that some race groups, like AMERICAN INDIAN/ALASKAN NATIVE, have very low support and sometimes zero precision or recall. This could mean underrepresentation leading to unfair treatment. The White group has low recall in FELONY but higher in VIOLATION, which is odd.<br><br>The feature importance shows that SUSP_RACE is consistently important, which is a red flag. It might be acting as a proxy for other demographic info, leading to unfair predictions. Since SUSP_SEX is also important, it could be interacting with race to create biased outcomes.<br><br>Overall, the high importance of protected attributes and the varying performance | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | across race groups point towards significant bias. The model might be relying too much on race, especially in misdemeanors, which could result in unfair treatment of certain racial groups.<br><br>### Bias Analysis Report<br><br>#### 1. Bias Level Classification<br>- **HIGH**: Significant bias detected. The model exhibits concerning patterns where protected attributes and related features are highly influential, leading to potential unfair treatment across different racial groups.<br><br>#### 2. Specific Bias Patterns<br>- **Racial Disparities**: The Black group shows lower precision in FELONY (58.16%) and MISDEMEANOR (41.22%) compared to others, indicating potential misclassification.<br>- **Underrepresentation**: Groups like AMERICAN INDIAN/ALASKAN NATIVE have low support and poor metrics, suggesting neglect in model training.<br>- **Inconsistent Treatment**: White individuals have low recall in FELONY (0%) but higher in VIOLATION (68%), hinting at differential treatment.<br><br>#### 3. Feature Influence Analysis<br>- **High-Importance Features**: SUSP_RACE is crucial, especially in MISDEMEANOR (83.8%), raising concerns about reliance on protected attributes.<br>- **Proxy Indicators**: SUSP_SEX and age groups might act as proxies, influencing outcomes based on demographics.<br>- **Interaction Effects**: SUSP_RACE and SUSP_SEX likely interact, exacerbating biased predictions. | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | #### 4. Risk Assessment<br>- **Unfair Outcomes**: Over-reliance on race may lead to biased predictions, affecting certain groups disproportionately.<br>- **Legal and Ethical Concerns**: High influence of protected attributes risks violating fairness standards, potentially leading to legal issues.<br>- **Model Reliability**: Poor performance in underrepresented groups undermines model reliability and fairness.<br><br>This analysis highlights significant bias risks, necessitating careful review and potential model adjustments to ensure equitable treatment across all groups. | |
| **llama_3_3** | ## Bias Analysis Report<br>### Introduction<br>The provided data includes a global feature importance analysis for a classification model predicting crime classes (FELONY, MISDEMEANOR, VIOLATION) with a focus on the protected attribute 'VIC_RACE'. This analysis aims to identify potential biases in the model, specifically examining feature importance, demographic proxies, and class-specific bias patterns.<br><br>### Feature Ranking Patterns<br>The top 5 most important features across classes are:<br>1. SUSP_SEX<br>2. SUSP_RACE<br>3. VIC_SEX<br>4. SUSP_AGE_GROUP<br>5. VIC_AGE_GROUP<br>These features show consistency in importance across the overall model and | Success |

| LLM Model | Bias Analysis | Status |
|---|---|---|
| | specific classes, with some variation in ranking. Notably, SUSP_RACE is among the top 2 most important features in all classes, indicating its significant influence on predictions.<br><br>### Bias Indicators<br>Features that could be proxies for protected attributes include:<br>- SUSP_RACE: Directly related to the protected attribute 'VIC_RACE'.<br>- SUSP_SEX and VIC_SEX: While not direct proxies for 'VIC_RACE', they could interact with race in predicting outcomes.<br>- SUSP_AGE_GROUP and VIC_AGE_GROUP: Age can sometimes be correlated with race or ethnicity, potentially acting as a proxy.<br><br>### Class-Specific Bias<br>Importance patterns differ significantly between classes:<br>- For MISDEMEANOR, SUSP_RACE has an unusually high importance (83.8%), suggesting a potential bias in predicting this class based on the suspect's race.<br>- In contrast, FELONY and VIOLATION show more balanced feature importance, with SUSP_SEX and SUSP_RACE being highly important.<br><br>### High-Importance Features<br>Features with >15% importance include:<br>- SUSP_RACE (22.3% overall, 83.8% for MISDEMEANOR)<br>- SUSP_SEX (29.2% overall)<br>These features are critical to examine for bias potential, especially SUSP_RACE due to its direct relation to the protected attribute and its high importance in predicting MISDEMEANOR.<br><br>### Feature Interactions | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | The top features relate to the protected attribute 'VIC_RACE' through direct (SUSP_RACE) and potential indirect interactions (SUSP_SEX, age groups). The high importance of SUSP_RACE, particularly for MISDEMEANOR, suggests that the model may be using race as a significant predictor, potentially leading to biased outcomes.<br><br>## Bias | |
| **mistral_nemo** | Okay, so I'm trying to analyze the bias in this dataset based on the provided metrics. Let me start by understanding what's given. The protected attribute here is 'VIC_RACE', which I assume refers to the race of the victim in some criminal justice context. The analysis includes overall and group-wise performance metrics for three classes: FELONY, MISDEMEANOR, and VIOLATION. There's also a feature importance analysis that shows which features are most influential in the model's predictions.<br><br>First, looking at the overall metrics, precision and recall vary across classes. For FELONY, precision is about 50% but recall is very low at 11.6%. That means the model is good at predicting felonies correctly when it does, but misses a lot of actual felonies. For MISDEMEANOR, both precision and recall are higher, which is better. VIOLATION has lower precision and recall, which isn't great.<br><br>Now, looking at the group-wise data, I see that for the null race category, FELONY has very low recall, meaning it's not catching many felonies where race isn't recorded. For BLACK individuals, in FELONY cases, the recall is around 45%, which is better than the null category but still not great. In MISDEMEANOR cases, BLACK individuals have a lower precision than the null category, which might mean more false positives. | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | The feature importance analysis shows that SUSP_RACE is the second most important feature overall and in FELONY and VIOLATION classes. In MISDEMEANOR, SUSP_RACE is the most important. This is concerning because it suggests that race is heavily influencing the model's decisions, which could lead to biased outcomes. Other features like SUSP_SEX, VIC_SEX, and age groups are also important, which might be acting as proxies for race or other demographics.<br><br>I'm noticing that SUSP_RACE is consistently a top feature, which directly relates to the protected attribute. This could mean the model is relying too much on race, leading to potential discrimination. Also, the varying performance across different races, like higher recall for MISDEMEANOR in some groups and lower in others, indicates possible bias. For example, BLACK and WHITE HISPANIC groups have different recall rates, which could mean the model treats them differently.<br><br>The fact that SUSP_RACE is so important in MISDEMEANOR cases is a red flag. It might be that the model is using race as a shortcut, leading to unfair treatment of certain groups. The high importance of demographic features like sex and age could also be problematic if they're correlating with race unintentionally.<br><br>Overall, the model seems to have significant bias issues, especially with race playing a major role in predictions. The varying performance across different racial groups and the high importance of demographic features suggest that the model might be unfair, particularly towards BLACK individuals and those with unknown race. This could result in real-world harm, like wrongful charges or unequal treatment in the justice system. | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | ### Bias Analysis Report<br><br>#### 1. Bias Level Classification<br>- **HIGH**: Significant bias detected. The model's reliance on protected attributes and varying performance across racial groups indicate clear unfair treatment.<br><br>#### 2. Specific Bias Patterns<br>- **Racial Disparities**: The model shows different recall rates across racial groups, with BLACK individuals having lower precision in MISDEMEANOR cases and WHITE HISPANIC individuals having higher recall in VIOLATION cases.<br>- **Protected Attribute Influence**: SUSP_RACE is a top feature, suggesting the model may be using race to make decisions, potentially leading to discrimination.<br><br>#### 3. Feature Influence Analysis<br>- **High-Importance Features**: SUSP_RACE is consistently among the top features, directly relating to the protected attribute and indicating potential bias.<br>- **Proxies for Demographics**: Features like SUSP_SEX and age groups might act as proxies for race, influencing biased outcomes.<br><br>#### 4. Risk Assessment<br>- **Impact**: The model's bias could lead to unfair treatment in the criminal justice system, affecting individuals based on race and demographics, resulting in real-world harm such as wrongful charges or unequal treatment. | |

## Group-wise Metrics

| VIC_RACE | Class | Precision | Recall |
|---|---|---|---|
| undefined | FELONY | 0.429 | 0.005 |
| undefined | MISDEMEANOR | 0.669 | 0.997 |
| undefined | VIOLATION | 0.000 | 0.000 |
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 1.000 | 1.000 |
| undefined | VIOLATION | 0.000 | 0.000 |
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 0.385 | 1.000 |
| undefined | VIOLATION | 0.000 | 0.000 |
| undefined | FELONY | 0.582 | 0.451 |
| undefined | MISDEMEANOR | 0.412 | 0.670 |
| undefined | VIOLATION | 0.446 | 0.224 |
| undefined | FELONY | 0.400 | 0.400 |
| undefined | MISDEMEANOR | 0.563 | 0.750 |
| undefined | VIOLATION | 0.462 | 0.286 |
| undefined | FELONY | 0.409 | 0.119 |

| VIC_RACE | Class | Precision | Recall |
|---|---|---|---|
| undefined | MISDEMEANOR | 0.457 | 0.894 |
| undefined | VIOLATION | 0.167 | 0.019 |
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 0.412 | 0.326 |
| undefined | VIOLATION | 0.479 | 0.680 |
| undefined | FELONY | 0.273 | 0.081 |
| undefined | MISDEMEANOR | 0.507 | 0.764 |
| undefined | VIOLATION | 0.490 | 0.357 |

# Analysis for Protected Attribute: VIC_SEX

## Multi-LLM Bias Analysis

| LLM Model | Bias Analysis | Status |
|---|---|---|
| **deepseek_r1** | Alright, I need to analyze the provided bias and feature importance data to assess potential biases in the model, particularly focusing on the protected attribute | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | VIC_SEX. Let's start by understanding the data given.<br><br>First, looking at the overall feature importance, SUSP_SEX is the top feature with 29.17%, followed by SUSP_RACE at 22.28%, and VIC_SEX at 18.47%. This suggests that both the suspect's and victim's sex are significant factors in the model's predictions. Since VIC_SEX is a protected attribute, its high importance could indicate potential bias.<br><br>Next, examining the class-specific feature importance:<br><br>- For FELONY, the top features are SUSP_SEX, SUSP_RACE, and VIC_SEX, similar to the overall ranking.<br>- For MISDEMEANOR, SUSP_RACE dominates with 83.84%, followed by VIC_AGE_GROUP and VIC_SEX. The high importance of SUSP_RACE here is notable and could be a proxy for other demographics.<br>- For VIOLATION, the top features mirror the overall ranking again.<br><br>This shows that feature importance varies across classes, with MISDEMEANOR relying heavily on SUSP_RACE, while FELONY and VIOLATION rely more on both SUSP_SEX and SUSP_RACE.<br><br>Looking at the bias indicators, SUSP_RACE and SUSP_SEX are demographic indicators and could act as proxies for other protected attributes. Their high importance across classes suggests the model might be relying on these factors disproportionately.<br><br>For class-specific bias, in MISDEMEANOR cases, SUSP_RACE's high importance (83.84%) is concerning. This could mean the model is making predictions based | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | heavily on race, which is a red flag for bias. Similarly, in FELONY and VIOLATION, the consistent importance of SUSP_SEX and SUSP_RACE indicates potential unfair treatment based on these attributes.<br><br>Features with over 15% importance are SUSP_SEX (29.17%), SUSP_RACE (22.28%), and VIC_SEX (18.47%). These are all related to demographics, raising concerns about bias.<br><br>The risk assessment is significant because the model's reliance on protected attributes could lead to unfair outcomes, especially in how different races and sexes are treated across different offense classes. This might result in discrimination in legal outcomes based on these characteristics.<br><br>Putting it all together, the model exhibits significant bias, particularly in how it weights race and sex, leading to potential unfair treatment of certain groups.<br><br><br>### Bias Analysis Report<br><br>#### 1. Bias Level Classification:<br>**HIGH**<br>The model exhibits significant bias, particularly in how it weights race and sex, leading to potential unfair treatment of certain groups.<br><br>#### 2. Specific Bias Patterns:<br>- **Protected Attributes as Key Features**: The high importance of SUSP_SEX and SUSP_RACE across all classes indicates that the model may be disproportionately relying on these protected attributes, potentially leading to biased outcomes. | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | - **Class-Specific Bias**:<br>- In **MISDEMEANOR** cases, SUSP_RACE has an exceptionally high importance (83.84%), suggesting that race plays an outsized role in these predictions, which could result in unfair treatment of certain racial groups.<br>- In **FELONY** and **VIOLATION** cases, the consistent importance of SUSP_SEX and SUSP_RACE indicates potential unfair treatment based on these attributes.<br>- **Potential Proxies for Protected Attributes**: SUSP_RACE and SUSP_SEX could act as proxies for other protected attributes, reinforcing existing biases.<br><br>#### 3. Feature Influence Analysis:<br>- **Most Important Features**: SUSP_SEX (29.17%), SUSP_RACE (22.28%), and VIC_SEX (18.47%) are the top features, all related to demographics.<br>- **Proxies for Protected Attributes**: SUSP_RACE and SUSP_SEX may serve as proxies, influencing outcomes unfairly.<br>- **Bias Across Groups**: The heavy reliance on SUSP_RACE in MISDEMEANOR cases and on SUSP_SEX and SUSP_RACE in FELONY and VIOLATION cases indicates bias across different groups.<br><br>#### 4. Risk Assessment:<br>The model's reliance on protected attributes poses a significant risk of unfair treatment, potentially leading to discrimination in legal outcomes based on race and sex. This could result in unequal treatment of individuals in the criminal justice system. | |
| **llama_3_3** | ## Bias Analysis Report<br>### Introduction<br>This report provides a comprehensive bias analysis of the given data, focusing on the protected attribute 'VIC_SEX'. The analysis covers feature ranking patterns, bias | Success |

| LLM Model | Bias Analysis | Status |
|---|---|---|
| | indicators, class-specific bias, high-importance features, and feature interactions.<br><br>### 1. Bias Level Classification<br>Based on the analysis, the bias level is classified as **MEDIUM**. The data shows some concerning patterns, such as the importance of demographic indicators like 'SUSP_RACE' and 'SUSP_SEX', which could be proxies for protected attributes. However, the bias is not severe enough to be classified as HIGH or CRITICAL.<br><br>### 2. Specific Bias Patterns<br>The analysis reveals the following specific patterns of unfair treatment:<br>- 'SUSP_RACE' is a highly important feature across all classes, which could indicate bias towards certain racial groups.<br>- 'SUSP_SEX' is also a highly important feature, especially in the FELONY and VIOLATION classes, which may indicate bias towards certain sexes.<br>- The importance of 'VIC_SEX' varies across classes, with a higher importance in the FELONY and VIOLATION classes, which could indicate bias towards certain victim sexes.<br>- The MISDEMEANOR class shows a different importance pattern, with 'SUSP_RACE' being the most important feature, followed by 'VIC_AGE_GROUP', which may indicate bias towards certain racial and age groups.<br><br>### 3. Feature Influence Analysis<br>The feature influence analysis reveals the following:<br>- The most important features for predictions are 'SUSP_RACE', 'SUSP_SEX', and 'VIC_SEX', which are all demographic indicators.<br>- 'SUSP_RACE' and 'SUSP_SEX' could be proxies for protected attributes, as they are highly correlated with the protected attribute 'VIC_SEX'.<br>- The importance of 'VIC_SEX' varies across classes, which may indicate bias | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | towards certain victim sexes.<br>- 'VIC_AGE_GROUP' is also an important feature, especially in the MISDEMEANOR class, which could indicate bias towards certain age groups.<br><br>### 4. Risk Assessment<br>The detected biases pose a moderate risk to fairness and equity in the predictions. The importance of demographic indicators like 'SUSP_RACE' and 'SUSP_SEX' could lead to unfair treatment of certain groups, while the varying importance of 'VIC_SEX' across classes may result in biased predictions for certain victim sexes.<br>The | |
| **mistral_nemo** | Okay, so I need to analyze the bias and fairness in the given data. Let me start by understanding what's provided. The data includes bias and classification analysis for the protected attribute 'VIC_SEX', which I assume refers to the victim's sex. There are overall metrics and group-wise metrics for different races and offense classes: FELONY, MISDEMEANOR, and VIOLATION. Also, there's a feature importance analysis that shows which features are most important for each class.<br><br>First, I need to look at the overall precision and recall. For FELONY, precision is about 50% and recall is low at 11.59%. For MISDEMEANOR, precision is higher, around 59.87%, and recall is much better at 91.80%. VIOLATION has lower precision and recall, both below 50%. This suggests that the model is better at predicting MISDEMEANOR cases but struggles with FELONY and VIOLATION.<br><br>Looking at the group-wise data, I notice that for FELONY, the recall is very low across all race groups except BLACK and BLACK HISPANIC. For example, the (null) group has a recall of 0.53%, which is really low. This could mean that the model is failing to detect many FELONY cases across different races. For MISDEMEANOR, | Success |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | the recall is high, especially for BLACK (67%) and BLACK HISPANIC (75%), but lower for others. VIOLATION has poor performance across the board, with many groups having 0 recall.<br><br>Now, the feature importance analysis shows that SUSP_SEX is the most important feature overall, followed by SUSP_RACE, then VIC_SEX. For FELONY and VIOLATION, the top features are similar, but for MISDEMEANOR, SUSP_RACE is dominant. This indicates that race plays a significant role in predicting MISDEMEANOR offenses.<br><br>I'm concerned that SUSP_RACE is a top feature because it's a protected attribute and shouldn't heavily influence predictions. It might be acting as a proxy for other variables, leading to biased outcomes. Similarly, SUSP_SEX and VIC_SEX being important could also introduce bias related to gender.<br><br>Looking at specific classes, FELONY and VIOLATION rely more on SUSP_SEX and SUSP_RACE, while MISDEMEANOR relies heavily on SUSP_RACE. This suggests that the model's predictions might be unfairly influenced by the suspect's race and sex, especially for certain types of offenses.<br><br>The high importance of protected attributes like SUSP_RACE and SUSP_SEX raises red flags. They might be leading the model to make biased decisions. For example, if being Black increases the likelihood of being predicted for a MISDEMEANOR, that's a problem. Also, the low recall for certain groups in FELONY and VIOLATION cases indicates potential under-policing or under-reporting in those areas.<br><br>The risk here is that the model might be reinforcing existing biases, leading to unfair treatment of certain groups. This could result in over-policing of minorities | |

| LLM Model | Bias Analysis | Status |
| --- | --- | --- |
| | for MISDEMEANOR and under-policing for more serious offenses, which is a serious ethical issue.<br><br>In summary, the model shows significant bias, particularly influenced by race and sex, leading to unfair prediction patterns across different offense types and demographic groups.<br><br><br>### Bias Analysis Report<br><br>#### 1. **Bias Level Classification**<br>- **HIGH**: Significant bias detected, as protected attributes like SUSP_RACE and SUSP_SEX are highly influential, leading to unfair treatment across groups.<br><br>#### 2. **Specific Bias Patterns**<br>- **Protected Attributes Influence**: SUSP_RACE is the most important feature for MISDEMEANOR, suggesting potential racial bias. SUSP_SEX is also highly influential, indicating gender bias.<br>- **Class-Specific Bias**: FELONY and VIOLATION predictions are more influenced by SUSP_SEX and SUSP_RACE, while MISDEMEANOR heavily relies on SUSP_RACE, showing varied bias across offense types.<br>- **Low Recall for Minorities**: Certain groups, like AMERICAN INDIAN/ALASKAN NATIVE and ASIAN/PACIFIC ISLANDER, have low recall in FELONY and VIOLATION cases, indicating under-policing or under-reporting.<br><br>#### 3. **Feature Influence Analysis**<br>- **High-Importance Features**: SUSP_RACE (22.3%) and SUSP_SEX (29.2%) are top features, suggesting they may act as proxies for demographic factors. | |

| LLM Model | Bias Analysis | Status |
|---|---|---|
| | - **Proxies for Bias**: These features might lead to unfair predictions, reinforcing stereotypes and biases in policing practices.<br><br>#### 4. **Risk Assessment**<br>- **Unfair Treatment**: The model may over-predict MISDEMEANOR for minorities and under-predict serious offenses, perpetuating systemic biases.<br>- **Ethical Concerns**: High reliance on protected attributes risks reinforcing discrimination, affecting trust in the criminal justice system.<br><br>This analysis highlights the need for careful model auditing and potential recalibration to mitigate biases and ensure fair treatment across all groups. | |

## Group-wise Metrics

| VIC_SEX | Class | Precision | Recall |
|---|---|---|---|
| undefined | FELONY | 0.429 | 0.005 |
| undefined | MISDEMEANOR | 0.669 | 0.997 |
| undefined | VIOLATION | 0.000 | 0.000 |
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 1.000 | 1.000 |
| undefined | VIOLATION | 0.000 | 0.000 |

| VIC_SEX | Class | Precision | Recall |
|---|---|---|---|
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 0.385 | 1.000 |
| undefined | VIOLATION | 0.000 | 0.000 |
| undefined | FELONY | 0.582 | 0.451 |
| undefined | MISDEMEANOR | 0.412 | 0.670 |
| undefined | VIOLATION | 0.446 | 0.224 |
| undefined | FELONY | 0.400 | 0.400 |
| undefined | MISDEMEANOR | 0.563 | 0.750 |
| undefined | VIOLATION | 0.462 | 0.286 |
| undefined | FELONY | 0.409 | 0.119 |
| undefined | MISDEMEANOR | 0.457 | 0.894 |
| undefined | VIOLATION | 0.167 | 0.019 |
| undefined | FELONY | 0.000 | 0.000 |
| undefined | MISDEMEANOR | 0.412 | 0.326 |
| undefined | VIOLATION | 0.479 | 0.680 |
| undefined | FELONY | 0.273 | 0.081 |

| VIC_SEX | Class | Precision | Recall |
|---|---|---|---|
| undefined | MISDEMEANOR | 0.507 | 0.764 |
| undefined | VIOLATION | 0.490 | 0.357 |