Introduction:

Predictive risk assessment tools are being popular in criminal legal system. They aid judges in categorising defendants as either "high risk" detention required or "low risk" allowing pre-trial release. The tool assigns a risk score, categorising individuals as high or low risk. This is based on the belief that a higher risk correlates with an increased probability of reoffending. In Australia, research has exposed racial bias contributing to the overrepresentation of indigenous in prison. Therefore, developing responsible risk assessment tools can lead the system to reduce the discrimination and wrong sentences. In summary the right tools offer a potential avenue to address sentencing disparities and improve the fairness of criminal proceedings in Australis.

Application of Principles of Responsible AI:

privacy and data protection:

To ensure that privacy and data protection are met, the dataset has been carefully processed to remove personally identifiable information, including names. Also, other potentially identifying information, such as id has been removed. The aim is to minimize the risk of disclosure while developing the analysis.

Reliability and safety assessments for the predictive risk assessment tool:

- Performance Evaluation
  The accuracy, precision, and overall performance of the predictive model are analysed. The outcome suggest that the model is not very accurate. The accuracy of the model is 29.50% if we consider the high-risk score is equal and greater than seven. Other metrics such as confusion matrix shows that the model predicts 4999, who should not be jailed but these people are jailed in the actual situation, are incorrectly. The model also forecast 87, who should be jailed but they are not arrested in the actual situation, are incorrectly. However, the model predicts 1908 should be arrested and 220 should not be arrested correctly.
  The model returns a better accuracy, if we choose high-risk score equal and greater than six. The accuracy of the model is 37.55%. The confusion matrix depicts that the model predicts 4388 should not be jailed but these people are jailed in the actual situation. The model also says 117 should be jailed but they are not arrested in the actual situation. The model predicts 2519 should be arrested and 190 should not be arrested correctly.

- Bias and Fairness Assessment:
  The AI360 is used to identify any disparities. The model is evaluated to discover potential biases and fairness issues, especially concerning protected attributes including age, gender, and race.
  Male, Caucasian, and age equal and greater than 40 are considered as privileged classes, and Female, African-American, and age less than 40 are considered as un-privileged classes. The selection is based on the historical discrimination towards Female and African-American. The age less than 40 is selected because the most of observation belongs to this class.

Findings:

<u>Difference Mean</u>:

<u>Age Category:</u> On average the unprivileged age group experiences 22.65% fewer favourable outcomes compared to the privileged age group.

Gender Category: A 5.96% higher average of favourable outcomes for the unprivileged gender category, Female, compared to the privileged gender category, Male.

Race Category: On average, the unprivileged race category, non-Caucasian, experiences 18.40% fewer favourable outcomes compared to the privileged race category, Caucasian.

Overall Population/ apply all categories: On average, the entire unprivileged population experiences 26.96% fewer favourable outcomes compared to the entire privileged population.

Original training dataset Difference in mean outcomes between unprivileged and privileged groups for age category	Original training dataset Difference in mean outcomes between unprivileged and privileged groups, Sex Cat	Original training dataset Difference in mean outcomes between unprivileged and privileged groups, Race	Original training dataset Difference in mean outcomes between all unprivileged and all privileged groups
-0.233277	0.059597	-0.183996	-0.269550

(Table 1: Mean difference)

Table 1: Mean difference

Disparate Impact:

Age Category: The unprivileged age group experiences approximately 70.71% of the favourable outcomes of the privileged age group.

Gender Category: Male benefits more from favourable outcomes than Female. Male experiences approximately 9.61% more favourable outcomes.

Race Category: Non-Caucasian has a higher probability of favourable outcomes compared to Caucasian. Non-Caucasian experiences approximately 75.55% of the favourable outcomes of the privileged race category.

Overall Population: Across all categories, the unprivileged population experiences approximately 69.70% of the favourable outcomes of the privileged population. (Table 2: Disparate Impact)

Original training dataset Disparate Impact age category	Original training dataset Disparate Impact, sex category	Original training dataset Disparate Impact (For Race)	Original training dataset Disparate Impact All categories
0.707142	1.096068	0.755471	0.696965

Table 2: Disparate Impact

Mean Differences

Age Category: Mean differences suggests a relatively balanced distribution of favourable outcomes between the two groups for the age category.

Sex Category: On average Male has a slightly higher rate of favourable outcomes compared to Female.

Race Category: On average, the non-Caucasian has a lower rate of favourable outcomes compared to Caucasian.

Overall Population/All Categories: On average the privileged group has a slightly higher rate of favourable outcomes compared to the unprivileged group across all categories. (Table 3: Mean Differences)

Difference in mean outcomes between unprivileged and privileged groups age category	Difference in mean outcomes between unprivileged and privileged groups, Sex Cat	Difference in mean outcomes between unprivileged and privileged groups, Race Cat	Difference in mean outcomes between unprivileged and privileged groups ALL Cats
-0.004168	0.068922	-0.087537	0.056356

Table 3: Mean Differences

Fairness Metrics:

Age Category: On average, the unprivileged group, less than 40 years old, has a lower rate of favourable outcomes compared to the privileged group, equal and greater than 40. Disparate Impact suggests that the unprivileged group has approximately 73.08% of the probability of favourable outcomes compared to the privileged group.

Sex Category: The positive mean difference indicates that, on average, Male has a slightly higher rate of favourable outcomes compared to Female. Disparate Impact suggests that Female has approximately 111.07% of the probability of favourable outcomes compared to Male.

Race Category: Mean Difference is -0.1548. This suggests that, on average, non-Caucasian has a lower rate of favourable outcomes compared to Caucasian. Disparate Impact suggests that non-Caucasian has approximately 79.22% of the probability of favourable outcomes compared to the Caucasian.

Overall Population/All Categories: Mean Difference is -0.1973. On average, the unprivileged group has a lower rate of favourable outcomes compared to the privileged group across all categories. Disparate Impact suggests that the unprivileged group has approximately 77.25% of the probability of favourable outcomes compared to the privileged group across all categories. (Table four: Fairness Metrics)

Age cat, Mean Difference: -0.2128 Disparate Impact: 0.7308	Mean Difference sex cat: 0.0694 Disparate Impact sex cat: 1.1107	Mean Difference Race Cat: -0.1548 Disparate Impact Race Cat: 0.7922	All Cats, Mean Difference ALL: -0.1973 Disparate Impact All: 0.7725

Table four: Fairness Metrics

Conclusion and Recommendation

The reliability and safety of the tool raised significant concerns. The model's accuracy is very low. The confusion matrix reveals inaccuracies including false predictions of people who should or should not be jailed. Furthermore, the finding indicates that disparities in favourable outcomes, with the unprivileged groups experience fewer positive outcomes. Moreover, across all categories, the unprivileged groups have lower rates of favourable outcomes. Unprivileged groups face a lower probability of favourable outcomes. Disparate Impact metrics quantify the likelihood of favourable outcomes for unprivileged groups compared to the other privileged groups. Mean differences highlight imbalances in favourable outcomes across age, gender, and race categories.

In conclusion, the predictive risk assessment tool demonstrates deficiencies in accuracy and fairness. The findings determines that the tool does not provide fair and reliable predictions. Finally, the recommendation is to implement measure to enhance model's accuracy, mitigate bias, and ensure transparency in the tool developments.