Ethics statement:

The bias audit of the income prediction model using the UCI Adult dataset reveals a critical ethical challenge at the heart of applied machine learning: models trained on historical data can systematically perpetuate and amplify existing societal inequalities. Our analysis found significant gender bias, where the baseline model was 19% less likely to predict a high income (>$50K) for females compared to males, despite similar underlying qualifications potentially being present in the data. This technical finding is not merely a statistical anomaly; it carries profound real-world implications that connect directly to core principles of AI ethics.

Firstly, this bias goes against the fairness and justice principles. Women would be systematically excluded from possibilities for financial growth if a model for evaluating creditworthiness, job seekers, or loan applications were utilized. This results in a detrimental feedback loop whereby the model learns past income discrepancies, which are frequently caused by unequal access to education and work, and then codifies them into an automated system, making it more difficult to end the cycle of inequality. This model penalizes people based on their demographic group membership rather than their unique circumstances or qualities.

Second, the project emphasizes the conflict between the moral principles of non-maleficence (avoidance of damage) and beneficence (doing good). It can be shown that the underprivileged group suffers harm from a highly accurate model that represents biased realities. As a result, using a model that is "accurate" but biased is unethical. Our mitigation efforts, especially the Reweighing technique, show that bias can be reduced dramatically at a low accuracy cost. By actively preventing harm, this demonstrates the principle of non-maleficence and implies that fairness ought to be regarded as an absolute need rather than a choice.

The audit also emphasizes how crucial accountability and transparency are. If the model's bias isn't consciously examined with tools like AIF360, it may continue to be a hidden problem with unexplainable discriminatory results. Banks and human resources departments are examples of stakeholders who need to accept responsibility for the results of their automated systems. This necessitates openness regarding the constraints and any prejudices of the models they employ, guaranteeing that choices may be defended and disputed. This transparency is facilitated by our visuals, which help both technical and non-technical decision-makers comprehend complicated fairness metrics.

Lastly, this piece relates to the idea of human dignity and autonomy. People lose some of their influence over their own financial future when they are forced to rely on biased algorithms to make decisions for them. They are subjected to a system that appears to be impartial but is actually biased. To protect human dignity and make sure that AI empowers everyone equally rather than subjecting them to historical prejudices, bias mitigation is crucial.

To sum up, the scientific procedure of auditing and reducing bias is essentially an important one. It is a real-world application of the obligation to create systems that are not only wise but also fair and right. The advice for stakeholders is straightforward: putting justice first is a moral need that needs to be incorporated into all phases of the AI development lifecycle, from gathering data to deploying and monitoring models.

Reference List:

Barocas, S. and Selbst, A.D. (2016) ‘Big Data’s Disparate Impact’, *California Law Review*, 104, p. 671.

Bellamy, R.K.E. *et al.* (2018) ‘AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias’, *arXiv preprint*, arXiv:1810.01943.

Bird, S. *et al.* (2020) *Fairlearn: A Toolkit for Assessing and Improving Fairness in AI*. Microsoft. Available at: https://fairlearn.org/ (Accessed: 19 September 2025).

Dua, D. and Graff, C. (2019) *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science. Available at: http://archive.ics.uci.edu/ml (Accessed: 19 September 2025).

Hardt, M., Price, E. and Srebro, N. (2016) ‘Equality of Opportunity in Supervised Learning’, in *Advances in Neural Information Processing Systems 29 (NeurIPS)*. Available at: https://papers.nips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html (Accessed: 19 September 2025).

Jobin, A., Ienca, M. and Vayena, E. (2019) ‘The Global Landscape of AI Ethics Guidelines’, *Nature Machine Intelligence*, 1(9), pp. 389-399.

Kamiran, F. and Calders, T. (2012) ‘Data Preprocessing Techniques for Classification without Discrimination’, *Knowledge and Information Systems*, 33(1), pp. 1–33.