Potential Biases in the Dataset

In the context of predicting issue priority (as shown in the model), several types of bias can arise if not proactively addressed:

1. **Underrepresented Teams or Departments**
   If certain teams (e.g., smaller business units or remote departments) submit fewer issues or have fewer labeled high-priority issues, the model may learn patterns that undervalue their submissions. This results in systematic under-prioritization of valid concerns from these groups, reinforcing existing inequalities.

2. **Labeling Bias**
   The target variable (`issue_priority`) may have been assigned subjectively by individuals or teams, introducing human bias into the labels. If, for example, issues raised by senior staff are consistently prioritized higher, the model may learn to favor patterns unrelated to the issue's technical urgency.

3. **Feature Imbalance**
   Features like team name, location, or platform could be unevenly distributed, causing the model to weight frequent categories more heavily. This can distort predictions for less-represented groups.

How IBM AI Fairness 360 Can Help

IBM AI Fairness 360 (AIF360) is an open-source toolkit specifically designed to detect, measure, and mitigate bias in machine learning models. It can be integrated into your workflow in the following ways:

1. **Bias Detection**
   AIF360 provides over 70 fairness metrics (e.g., disparate impact, equal opportunity difference) that evaluate whether your model treats different groups (e.g., teams, regions) fairly. You can use these metrics post-training to analyze how predictions differ across sensitive attributes.

2. **Preprocessing Mitigation**
   If bias is detected in your dataset (e.g., class imbalance), AIF360 offers techniques like **reweighting** or **resampling** that adjust the training data to make it more equitable without distorting the underlying truth.

3. **In-Processing Algorithms**
   Some fairness techniques (like adversarial debiasing) modify the training process to penalize biased decisions, ensuring that fairness is baked directly into the model during

training.

4. **Post-Processing Corrections**
   If the model is already deployed, AIF360 can apply fairness-aware adjustments to the outputs (e.g., equalized odds post-processing) without retraining the model.