



Corso di Laurea Magistrale in Informatica

Machine Learning Fairness: Definizione di un Catalogo di Pattern e Anti-Pattern

Prof. Fabio Palomba
Dott.ssa Giulia Sellitto

Francesco Abate
Mat. 0522500993



Fairness: i sistemi software non sono equi



Un software di Amazon per la selezione del personale discriminava in base al sesso



Il generatore di sottotitoli di YouTube discriminava in base al sesso e la lingua



Un software della polizia statunitense discriminava in base al colore della pelle



Obiettivi

Stesura di un catalogo di root cause e practice



Scoprire i possibili attributi e i possibili aspetti del machine learning che potrebbero causare discriminazioni



Classificare diverse practice inerenti al machine learning per comprendere se sia meglio adottarle o meno



Studio Empirico



Struttura del survey

Domande poste agli esperti del settore



Progettazione tramite linee guida



Realizzato con Google Form

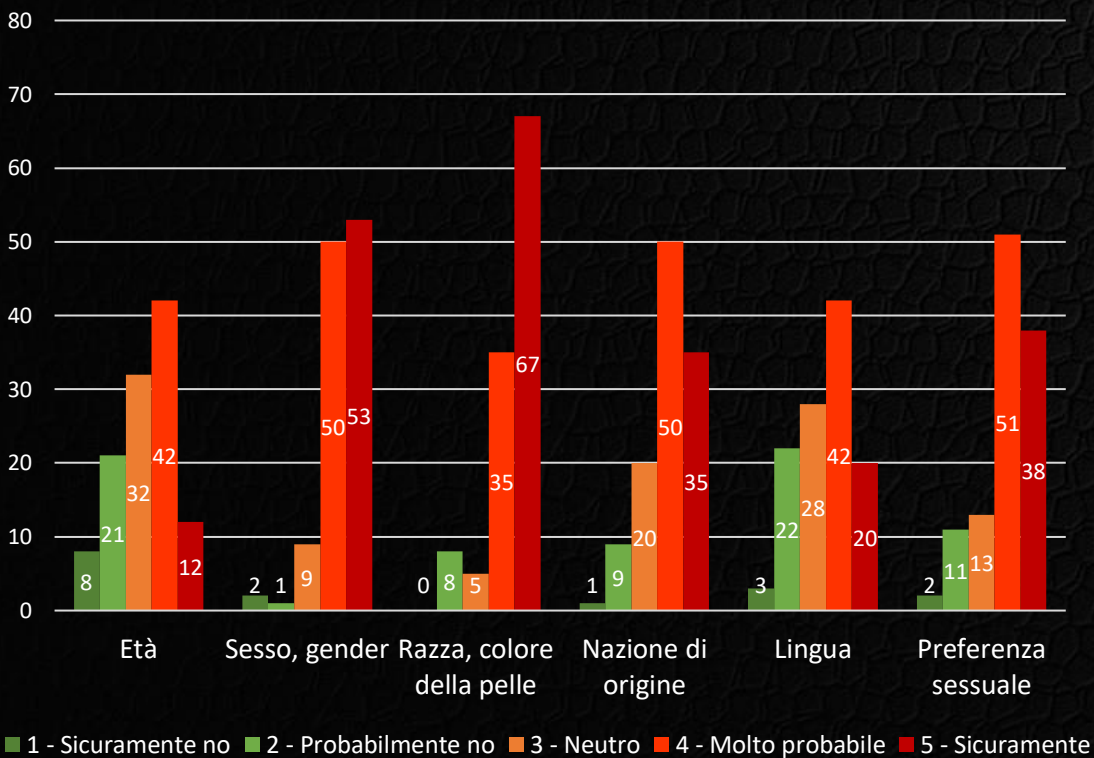


Reclutamento tramite Prolific

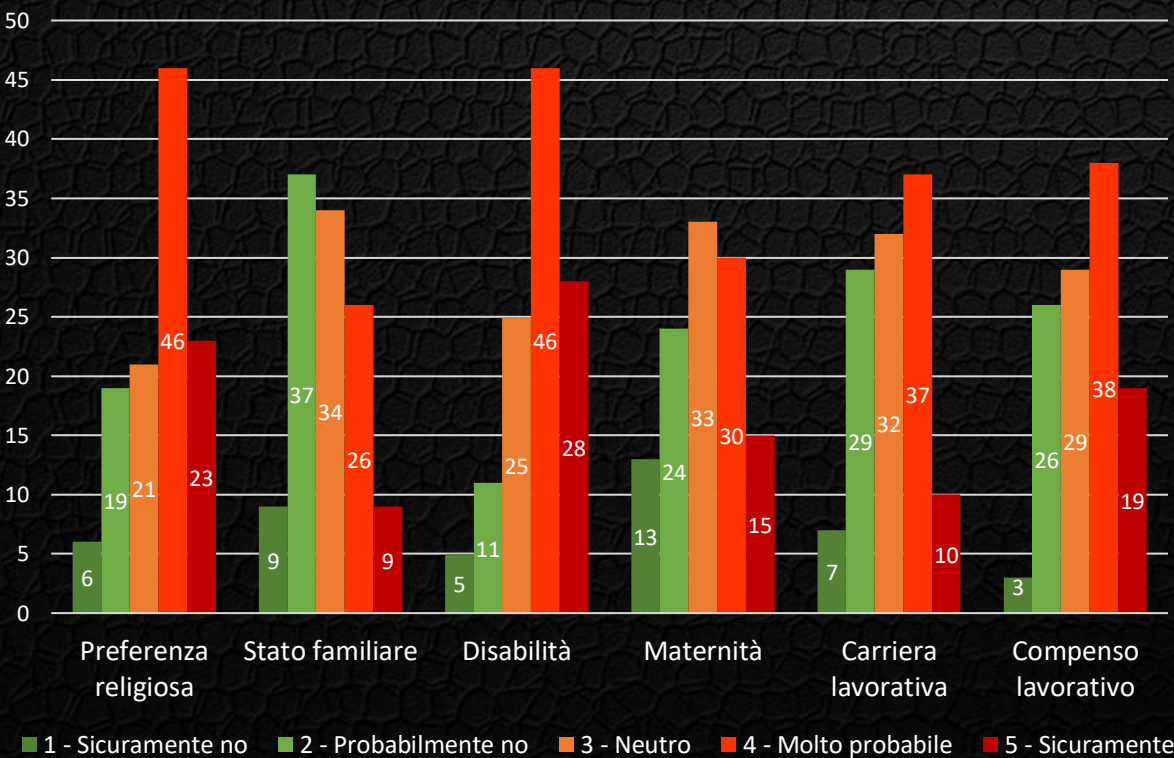


Risultati: Cause di Discriminazioni – Attributi Sensibili

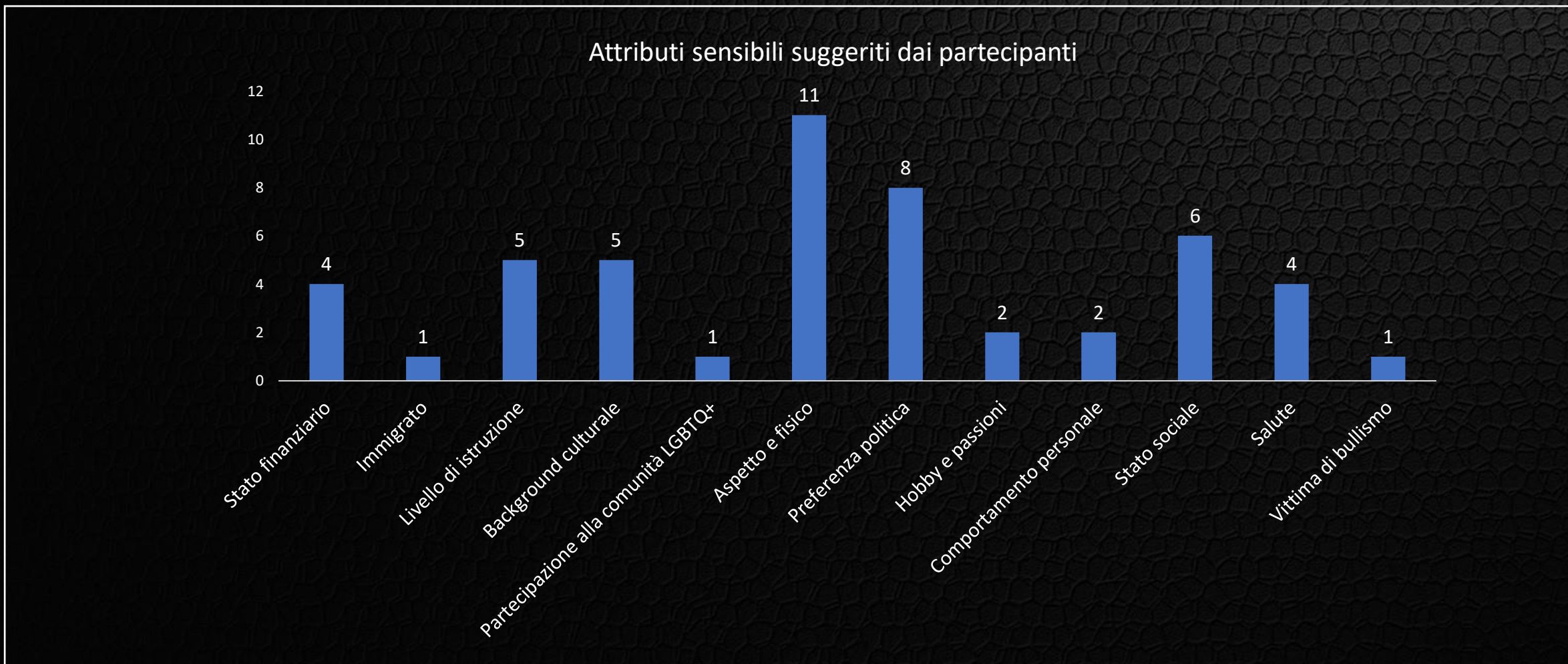
Con che probabilità i seguenti attributi sensibili potrebbero causare discriminazioni?



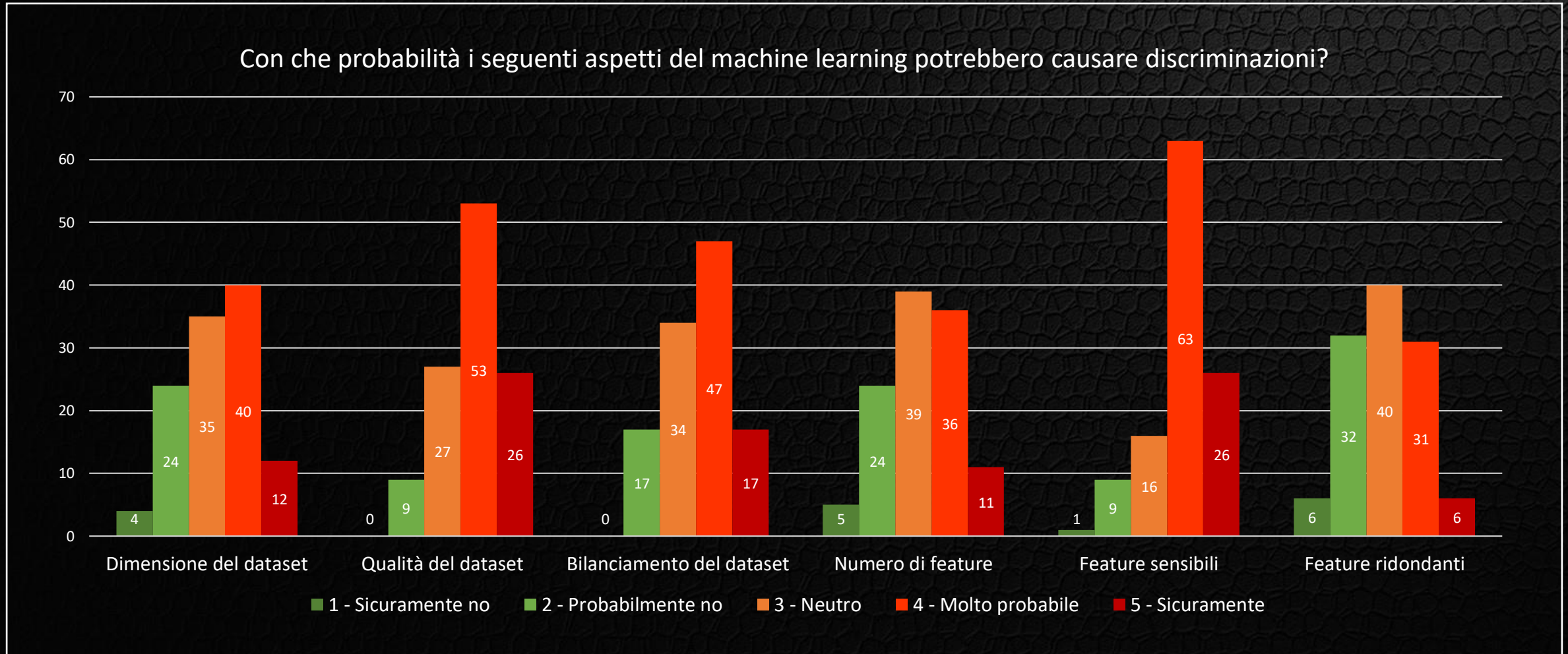
Con che probabilit  i seguenti attributi sensibili potrebbero causare discriminazioni?



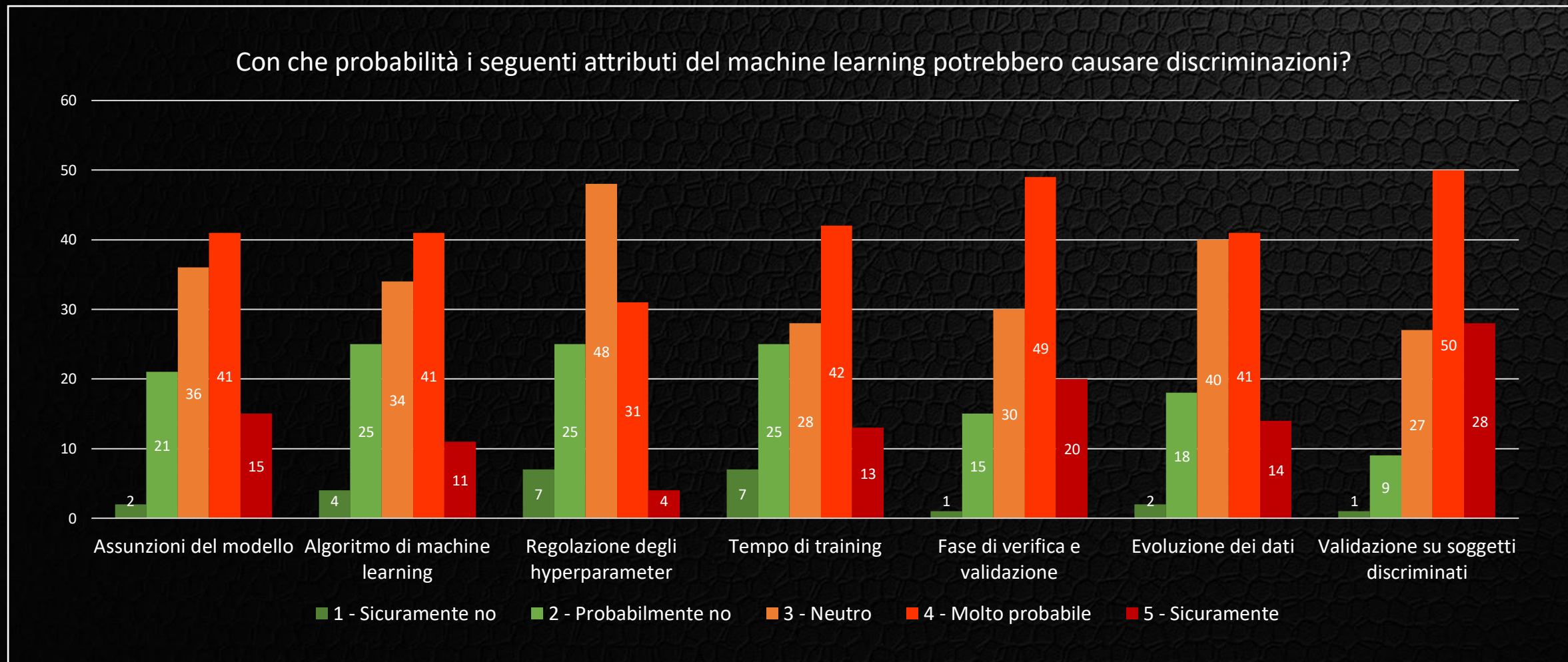
Risultati: Cause di Discriminazioni – Attributi Sensibili



Risultati: Cause di Discriminazioni – Aspetti del Machine Learning

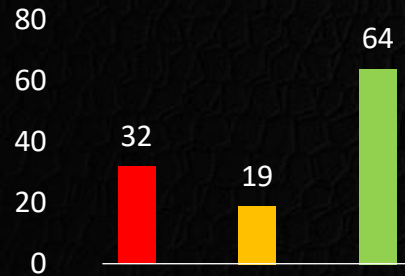


Risultati: Cause di Discriminazioni – Aspetti del Machine Learning

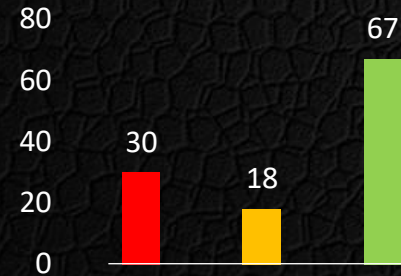


Risultati: Classificazione in Best Practice

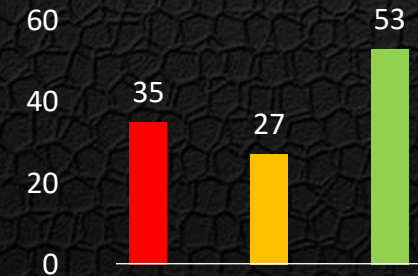
Svolgimento di interview e focus groups
al fine di elicitare requisiti inerenti la
fairness



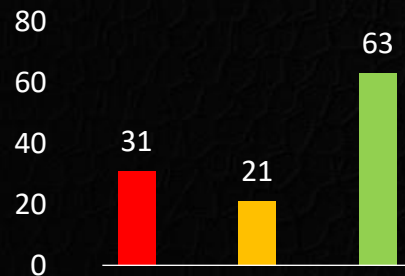
Considerare la fairness come un aspetto
prioritario durante la fase di analisi



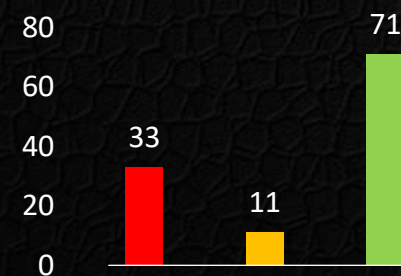
Utilizzo di combinazioni di modelli
(ensemble learning)



Training del modello su
un dataset bilanciato



Testing e validazione su
un dataset bilanciato



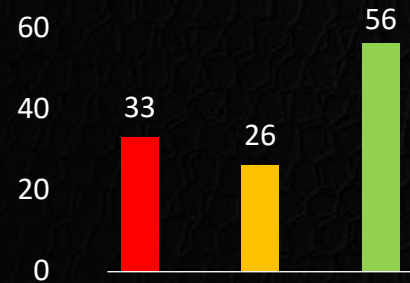
Legenda

- Classificazione in bad practice
- Nessuna classificazione
- Classificazione in best practice

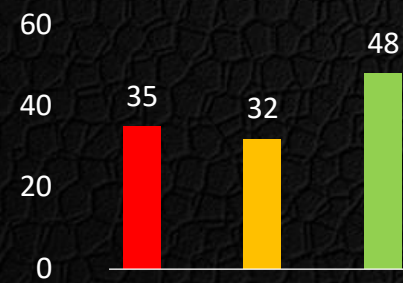


Risultati: Classificazione in Best Practice

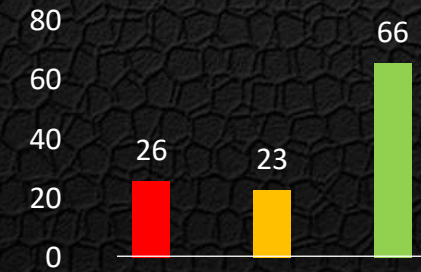
Rispetto delle assunzioni fair tramite la scelta dell'algoritmo di machine learning



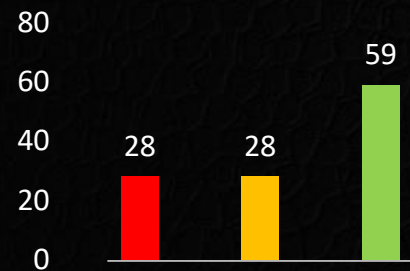
Regolazione degli hyperparameter



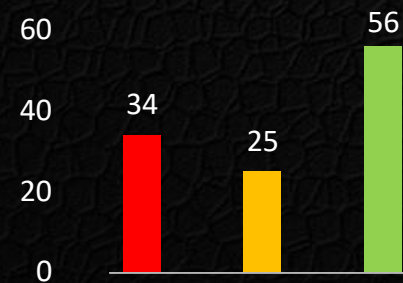
Manutenzione ed evoluzione considerando i cambiamenti dei dati nel tempo



Assegnamento dei pesi per rendere il modello più preciso



Assegnamento dei pesi per risolvere overfitting e underfitting



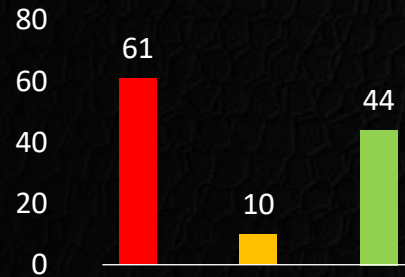
Legenda

- Classificazione in bad practice
- Nessuna classificazione
- Classificazione in best practice

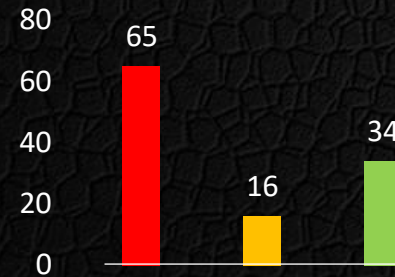


Risultati: Classificazione in Bad Practice

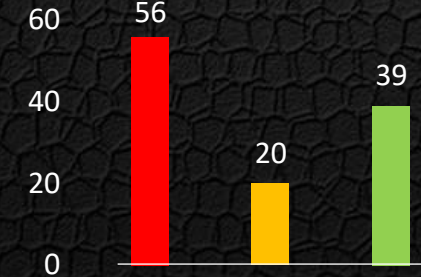
Training del modello basato principalmente su feature sensibili



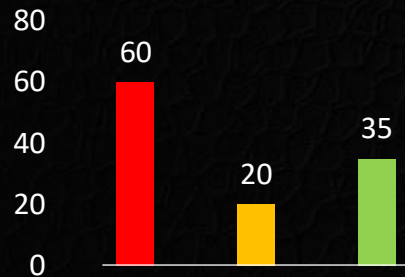
Rimozione delle feature sensibili dal dataset



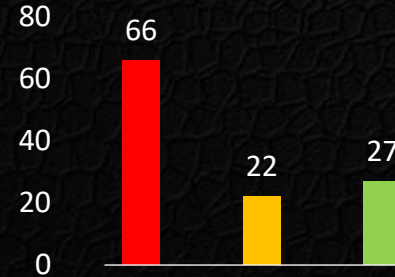
Validazione su un dataset contenente solo soggetti discriminati



Aggiunta di nuovi record nel dataset per avvantaggiare gli individui discriminati



Assegnazione di valori randomici ai pesi del modello



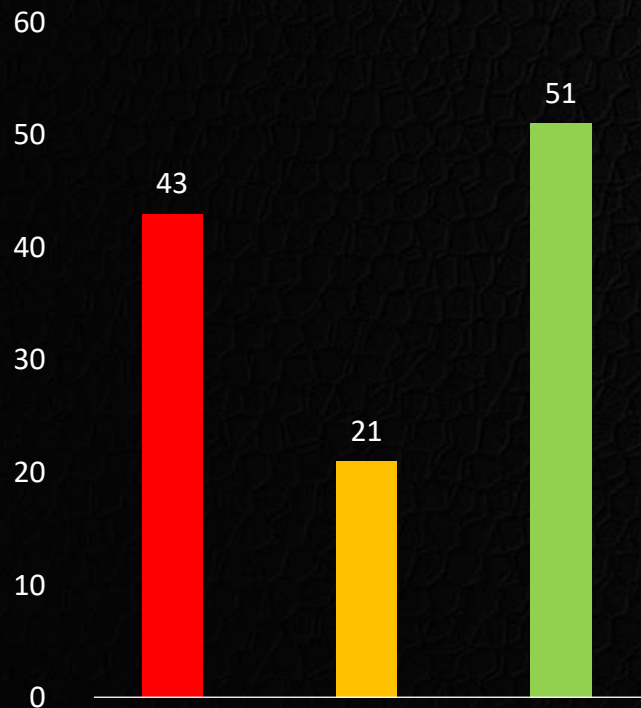
Legenda

- Classificazione in bad practice
- Nessuna classificazione
- Classificazione in best practice

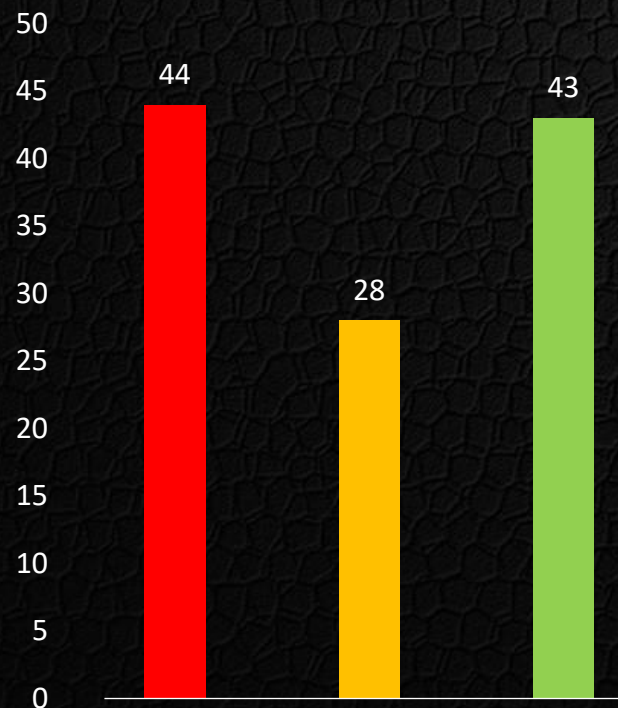


Risultati: Practice non classificabili

Confronto di un modello che fa uso di feature sensibili con un modello che non ne fa uso



Bilanciamento dei dati basato sulle feature sensibili



Legenda

- Classificazione in bad practice
- Nessuna classificazione
- Classificazione in best practice



Ricapitolazione delle Practice

Best practice

Svolgimento di interview e focus groups al fine di elicitare requisiti inerenti la fairness

Utilizzo di combinazioni di modelli (ensemble learning)

Regolazione degli hyperparameter

Rispetto delle assunzioni fair tramite la scelta dell'algoritmo di machine learning

Testing e validazione su un dataset bilanciato

Training del modello su un dataset bilanciato

Considerare la fairness come un aspetto prioritario durante la fase di analisi

Assegnamento dei pesi per risolvere overfitting e underfitting

Assegnamento dei pesi per rendere il modello più preciso

Manutenzione ed evoluzione considerando i cambiamenti dei dati nel tempo

Practice non classificabili

Confronto di un modello che fa uso di feature sensibili con un modello che non ne fa uso

Bilanciamento dei dati basato sulle feature sensibili



Bad practice

Training del modello basato principalmente su feature sensibili

Rimozione delle feature sensibili dal dataset

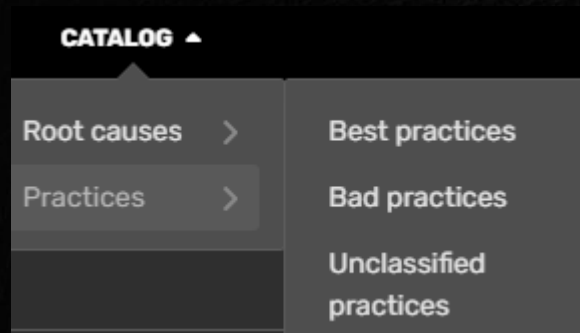
Validazione su un dataset contenente solo soggetti discriminati

Aggiunta di nuovi record nel dataset per avvantaggiare gli individui discriminati

Assegnazione di valori randomici ai pesi del modello



Catalogo Online e Sviluppi Futuri



- Informazioni inerenti la practice
- Classificazione dei partecipanti
- Argomenti inerenti la practice
- Fonti
- Come contribuire

Visiona il catalogo

fairness-guru.fandom.com/wiki/Fairness_Guru_Wiki

Best practices

 |  [VIEW SOURCE](#) | 

A best practice is a standard or set of guidelines that is known to produce good outcomes if followed [source].

The following practices are classified as "best". The classification is based on a set of responses obtained by approximately 120 people working with fair critical machine learning systems.

- Fairness requirements emerged with interviews
- Use of composite models
- Hyperparameters tuning
- Fair hypothesis respect through ML algorithm
- Validation and testing on a balanced dataset
- Model training on a balanced dataset
- Fairness as a priority aspect during analysis phase
- Weight assignment for underfitting and overfitting
- Weight assignment to make model more precise
- Maintenance considering data changes over time






Grazie per l'attenzione

Francesco Abate

 abatefrancesco98@gmail.com

 Crediti: Flaticon (icone), Storyset (illustrazioni)

sesa^{lab}
SOFTWARE ENGINEERING
SALERNO

freedom[®]
let's green the planet

Questa tesi ha contribuito a piantare
un albero nella SeSa Random Forest



Scansiona il QR Code e
contribuisci alla causa