UNIVERSIDADE
BEIRA INTERIOR

A Justiça da Inteligência Artificial e Algoritmos de Apoio à Decisão em Sistemas de Gestão de Ofensores

Rafael Filipe Morais Pais da Silva

Dissertação para Obtenção do Grau de Mestre em
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Doutor Nuno Gonçalo Coelho Costa Pombo

Junho de 2023

Declaração de Integridade

Eu, Rafael Filipe Morais Pais da Silva, que abaixo assino, estudante com o número de inscrição M10669 de/o Mestrado em Engenharia Informática da Faculdade de Engenharias, declaro ter desenvolvido o presente trabalho e elaborado o presente texto em total consonância com o Código de Integridades da Universidade da Beira Interior.

Mais concretamente afirmo não ter incorrido em qualquer das variedades de Fraude Académica, e que aqui declaro conhecer, que em particular atendi à exigida referenciarão de frases, extratos, imagens e outras formas de trabalho intelectual, e assumindo assim na íntegra as responsabilidades da autoria.

Universidade da Beira Interior, Covilhã 11/06/2023

Rafael Filipe Morais Pais da Silva

Resumo

Sistemas de gestão de ofensores (Offender Management Systems) têm vindo a ser adotados cada vez mais em inúmeros países como Estados Unidos da América[1], Singapura[2] Finlândia [3] entre outros países. A sua implementação permite que sistemas prisionais forneçam uma melhor experiência de reabilitação para os ofensores, uma monitorização mais detalhada e o calculo de fatores de risco durante e após a sua sentença. Porém a literatura existente descreve que no que toca ao calculo de risco de reincidência, estes sistemas tendem a descriminar alguns grupos étnicos e favorecer outros. Esta dissertação tem como objetivo em primeira instância, reunir e analisar resultados obtidos na literatura e posteriormente comparar diferentes métodos de *machine learning* para este calculo de reincidência para comparar resultados e combater a descriminação presente atualmente.
A analise efetuada nesta dissertação, comparou 3 modelos de *machine learning* desenvolvidos com 3 métodos diferentes, *Adaboost Classifier, Logistic Regression* e *Random Forest Classifier*, cada um destes modelo visa o calculo de reincidência de um ofensor. Para a avaliação destes 3 modelos foram usadas as seguintes métricas *Score, Precission, Recall* e *F1 Score* e a comparação de características de entrada com maior importância na construção do modelo. Foi também feita a analise de dois *datasets*, *COMPAS Score dataset* e *NIJ Recidivism Challange Dataset*, em que foram comparadas e analisadas as características dos ofensores como a idade, género e etnia.

Palavras-chave

Ofensores, Inteligência artificial, Algoritmos de Apoio à Decisão, Equidade, Justiça, Imparcialidade, Etnia, Idade, Genero, *Dataset*, Gestão de Ofensores, Risco, Reincidência, *Adaboost Classifier, Logistic Regression, Random Forest,* LSI-R, COMPAS, SAVRY, NIJ Recidivism Challange Dataset.

Abstract

Offender Management Systems have been increasingly adopted in numerous countries such as the United States of America, Singapore, Finland and other countries. Their implementation allows prison systems to provide a better rehabilitative experience for offenders, more detailed monitoring and calculation of risk factors during and after their sentence. However, the existing literature describes that when it comes to calculating risk of recidivism, these systems tend to discriminate against some ethnic groups and favor others. This dissertation aims in the first instance, to gather and analyze results obtained in the literature and then compare different methods of *machine learning* for this recidivism calculation in order to compare results and combat the discrimination currently present.
The analysis performed in this dissertation, compared 3 models of *machine learning* developed with 3 different methods, *Adaboost Classifier, Logistic Regression* and *Random Forest Classifier*, each of these models aims at the calculation of recidivism of an offender. For the evaluation of these 3 models the following metrics *Score, Precission, Recall* and *F1 Score* were used and the comparison of input features with greater importance in the construction of the model. Two *datasets* were also analized, *COMPAS Score dataset* and *NIJ Recidivism Challange Dataset*, where the characteristics of offenders such as age, gender and ethnicity were compared and analyzed.

Keywords

Offenders, Artificial Intelligence, Decision Support Algorithms , Fairness, Justice, Impartiality, Ethnicity, Age, Gender, *Dataset*, Offender Management, Risk, Recidivism, *Adaboost Classifier, Logistic Regression, Random Forest,* LSI-R, COMPAS, SAVRY, NIJ Recidivism Challange Dataset.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AI	Artificial Inteligence
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
IA	Inteligência Artificial
LSI-R	Level of Service Inventory - Revised
SAVRY	Structured Assessment of Violence Risk in Youth
UBI	Universidade da Beira Interior

Capítulo 1

Introdução

Estabelecimentos prisionais inteligentes já são uma realidade em países como os Estados Unidos da América[1], Singapura[2], Finlândia[3], Hong Kong[9] entre outros. Estes sistemas permitem que seja feita uma melhor monitorização e gestão de ofensores, fornecem um apoio no que toca a decisões a serem tomadas. No entanto estas ferramentas de apoio a decisão por vezes são implementadas com base em algoritmos com falhas, nomeadamente falhas perante a justiça e o favorecimentos de certos grupos de indivíduos.
Neste capitulo introdutório será dada um breve descrição de sistemas de gestão de ofensores e de algoritmos de apoio à decisão. Será também apresentado o papel da inteligência artificial em sistemas de gestão de ofensores e a justiça nestes sistemas.

1.1 Sistemas de Gestão de Ofensores

Um sistema de gestão de ofensores é desenhado para armazenar dados referentes a todo o histórico criminal de um ofensor, desde da sua entrada inicial e classificação até a sua saída do sistema criminal. Fornece também informações detalhadas como o tempo cumprido e qualquer transferências entre instituições ao longo do tempo. O uso desta tecnologia proporciona uma forma de serem criados desfechos mais positivos e reduzir o crime em geral. Todos estes dados referentes ao ofensor são usados para tomar decisões de como gerir um ofensor da melhor forma. Com estes sistemas os ofensores mais problemáticos e persistentes são identificados e corretamente geridos em conjunto com forças policiais que trabalham em conjunto. Um sistema de gestão de ofensores é composto por vários subsistemas tais como:

- Classificação inicial;
- Avaliação de risco e necessidades;
- Alocação de prisão e cela.

A classificação inicial é feita assim que o ofensor é inserido no sistema. Nesta classificação são utilizados dados básicos como idade, morada, morada da família, sentenças anteriores, entre outros, em conjunto com as necessidades do ofensor para construir uma visão geral do mesmo.
Na avaliação de risco e necessidades o sistema tem em conta as necessidades do ofensor, medicas, psicológicas e outras necessidades relevantes, para determinar o melhor sitio para ser inserido e os melhores programas para determinado ofensor. É também feita uma avaliação de risco para cada ofensor que com base em diferentes fatores como por exemplo sentenças anteriores, tipo de crime, comportamento em geral, para calcular o seu nível de risco de voltar a reincidir em crime.

Todos estes dados recolhidos e avaliações são posteriormente usados na alocação do ofensor numa prisão e numa cela em especifico. O ofensor é inserido na prisão que oferece melhores programas que vão de encontro às suas necessidades e numa cela que visa em reduzir risco e problemas durante a sua sentença.

1.2 Algoritmos de Apoio à Decisão

No contexto de Sistemas de Gestão de Ofensores, os algoritmos de apoio à decisão são maioritariamente utilizados na avaliação de risco[10]. Estas ferramentas tornaram-se cada vez mais importantes em estabelecimentos prisionais, e foi adotadas uma grande variedade de ferramentas deste tipo para ajudar na classificação, gestão e tratamento de ofensores[11]. Os instrumentos utilizados inserem ofensores em grupos de acordo com o seu nível de risco (baixo, médio ou alto) com base em fatores estáticos, fatores que não podem ser alterados como por exemplo o seu histórico criminal, e fatores dinâmicos, que podem evoluir com o tempo como por exemplo abuso de substancias ou comportamentos anti-sociais[12]. Estas ferramentas podem ou não ser implementadas em sistemas inteligentes e alguns exemplos são:

- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions);
- PATTERN (Prisoner Assessment Tool Targeting Estimated Risk and Needs);
- LSI-R (Level of Service Inventory-Revised).

No entanto todos estes demonstram um certo de injustiça perante alguns grupos de indivíduos e favorecimento perante outros.

1.3 IA em Sistemas de Gestão de Ofensores

A implementação de inteligência artificial em estabelecimentos prisionais tem o potencial de impactar uma grande variedade de operações e facilitar a tomada de decisões. Estes sistemas são em grande parte sistemas de recomendação que ajudam os técnicos nestes estabelecimentos a tomar, aparentemente, as melhores decisões. Com base nas caraterísticas do ofensor e da prisão, a inteligência artificial, pode indicar por exemplo a prisão mais indicada para um determinado ofensor bem como a cela e ala mais apropriadas. Pode ser também utilizada para indicar programas que vão de encontro as necessidades do ofensor e até fazer o calculo do risco do ofensor com objetivo de reduzir a reincidência do mesmo. Tecnologias de IA podem também ser usadas para monitorizar comunicações entre ofensores, a sua localização, dados biométricos e detetar contrabando.
Tudo isto faz parte de um plano maior de tornar estabelecimentos prisionais mais inteligente e mais tecnológicos que visam o melhor tratamento dos ofensores, a diminuição da reincidência em crime bem como uma ajuda para os técnicos presentes nestes estabelecimentos a tomar a melhores decisões possíveis. Contudo ainda é algo um pouco longe da realidade sendo que as implantações presentes ainda mostram alguns problemas.

1.4 A Justiça em Sistemas de Gestão de Ofensores

Apesar dos recentes avanços e implementações para um sistema prisional mais inteligente e tecnológico ainda existem certas falhas e inconsistências que comprometem o uso destes sistemas.
Em 2016 foi conduzida uma investigação pela organização ProPublica[13] que tinha como objetivo analisar a ferramenta COMPAS, um algoritmo de suporte a decisão utilizado em sistemas prisionais. Esta analise obteve os seguintes resultados:

- Ofensores negros eram frequentemente classificados com um risco mais alto de reincidência em crime do que eram na realidade;
- Ofensores brancos eram frequentemente classificados com um nível de risco mais baixo do que realmente eram;
- No que toca a crimes violentos ofensores negros eram também duas vezes mais prováveis de serem classificados com um risco superior do que ofensores brancos;
- Ofensores brancos eram 63% mais prováveis de serem classificados com um baixo risco de reincidência em crimes violentos comparativamente com ofensores negros.

COMPAS foi um dos casos mais estudados mas grande parte destes algoritmos sofrem do mesmo problema.
No decorrer deste trabalho será aprofundado mais este tema em diferentes algoritmos e modelos de IA.

Capítulo 2

Estado da Arte

2.1 Introdução

Nesta secção intitulada de Estado da Arte irá ser ser feita uma analise de alguns artigos e publicações dentro do tema da justiça e imparcialidade da inteligência artificial e algoritmos de apoio a decisão no contexto de sistemas de gestão de ofensores.
Primeiramente irá ser feita uma breve apresentação dos algoritmos usados que incluirá o seu desenvolvimento, o método e outras informações pertinentes para o estudo. De seguida serão analisados artigos e publicações referentes a cada algoritmos.

2.2 LSI-R

O *Level of Service Inventory (LSI)* foi desenvolvido no inicio do anos 80 por dois psicólogos Canadianos, Don Andrews e James Bonta. Uma década depois, nos anos 90, a ferramenta foi atualizada e renomeada para *Level of Service Inventory-Revised (LSI-R)*, e foi considerada uma ferramenta de avaliação de risco de terceira geração.
Ferramentas de primeira geração são em grande parte baseadas em julgamentos subjetivos, feitos pelos técnicos encarregues de trabalhar com os ofensores. Estes julgamentos são baseados em experiência e intuição para decidir que risco do ofensor para com a sociedade. Apesar destas ferramentas não poderem ser descartadas, estudos mostram que decisões baseadas nestes instrumentos são menos precisas do que decisões baseadas em dados empíricos ou estatísticos.
Ferramentas de segunda geração surgiram com uma tentativa de avançar para além de julgamentos subjetivos e medir o risco do ofensor objetivamente. O ponto fraco destes instrumentos é que foram compostos por itens que principalmente mediam riscos estáticos. Estes risco não podem ser alterados como por exemplo o histórico criminal. Ainda que estes instrumentos possam prever futura reincidência, os riscos estáticos levam a uma fraca orientação para intervenção no tratamento dos ofensores.
Os instrumentos de terceira geração incluem tanto riscos estáticos como riscos dinâmicos que podem ser avaliados durante a pena, como comportamentos anti-sociais e atitudes em geral. Uma ferramenta deste tipo identifica não só o risco de reincidência mas também as necessidade criminológicas do ofensor, que são depois usadas para o seu tratamento através de intervenções e programas de reabilitação.
O LSI-R, é uma destas ferramentas, proporciona uma medida de violação de regras ao identificar o nível de risco de reincidência e características criminais associadas com tais comportamentos.

2.2.1 Método

O LSI-R é atualmente usado em inúmeras jurisdições internacionais, que incluem os Estados Unidos da América, O Reino Unido e Austrália. Este instrumento contem 54 itens e produz um valor de risco global, que é caraterizado entre 5 níveis de risco diferentes. Um valor de risco elevado representa uma maior tendência para cometer crimes futuros. As categorias de risco são:

- 1 - Baixo, com valores de risco entre 0 e 13 pontos;
- 2 - Baixo/Moderado, com valores de risco entre 14 e 23 pontos;
- 3 - Moderado, com valores de risco entre 24 e 33 pontos;
- 4 - Moderado/Alto, com valores de risco entre 34 a 40 pontos;
- 5 - Alto, com valores de risco entre 41 a 54 pontos

O valor de risco para cada ofensor é formado através da avaliação de 10 sub-escalas de necessidades diferentes:

- Histórico Criminal;
- Educação/Emprego;
- Finanças;
- Família/Conjugais;
- Alojamento;
- Lazer;
- Companhias;
- Problemas com Álcool/Drogas;
- Problemas Emocionais/Pessoais;
- Atitudes.

A informação é recolhidas através de uma entrevista estruturada conduzida por um praticante de justiça criminal. Durante a entrevista é administrado o instrumento, que pode levar de 45 minutos a 1 hora a completar. Os 54 itens são marcados como SIM ou NÃO ou numa escala de 0 a 3:

- 0 - situação muito insatisfatória com uma clara necessidade para melhoria;
- 1 - situação relativamente insatisfatória com necessidade de melhoria;
- 2 - situação relativamente satisfatória com algum espaço para melhoria;

- 3 - situação satisfatória sem necessidade de melhoria.

Após a entrevista o técnico criminal avalia o ofensor nos 54 itens. Por cada SIM e por cada o ou 1, o risco global é incrementado por 1 valor. No final o valor de risco obtido é comparado com os 5 níveis de risco e o ofensor é inserido num deles. Por fim o técnico pode delinear um plano adequado para o ofensor com base no seu risco e necessidades.

2.2.2 *Racial Bias and LSI-R Assessments in Probation Sentencing and Outcomes*

Instrumentos de avaliação de risco podem mostrar um potencial viés tanto na previsão de reincidência bem como no seu uso para recomendar sentenças, super visionamento e tratamento. No que toca a calculo de sentenças, estes instrumentos, podem mostrar desfavorecimento para grupos de minorias e atribuir sentenças mais longas e uma maior intervenção judicial. Neste artigo os autores, Evan M. Lowder, Megan M. Morrison, Daryl G. Kroner, e Sarah L. Desmarais[14], afirmam que ate à data foram efetuados poucos estudos que avaliem o uso de ferramentas de avaliação de risco no contexto de calculo de sentença e envolvimento judicial.
A amostra usada foi composta por 11.798 ofensores em Kansas que foram condenados a liberdade condicional. Eram maioritariamente brancos (8.811, 74.7%), em comparação com negros (2.981, 25.3%). Grande parte eram masculinos (9.276, 80.1%) e não hispânicos (10.404, 88.2%). A media de idades era 32.37 anos num intervalo de 15 a 84 anos na altura da avaliação. Os ofensores tinham em media 1.16 acusações de ofensa e um correspondente nível de segurança de 5.99. Todos os registos foram obtidos no *Kansas Department of Corrections* de 2003 a 2015.
Ofensores de raça branca foram ligeiramente mais classificados com risco baixo (1.117, 12.7%) e alto (183, 2.1%) em relação com ofensores de raça negra (41, 1.4%, risco baixo, 98, 1.4%, risco alto). Ofensores negros foram classificados em maior numero para risco moderado (1.219, 40.9%) em comparação com ofensores brancos (3.358, 38.1%). Não houve diferença entre as duas raças em questão no que toca a risco baixo/moderado (3.168, 36% e 1.102, 37.0% respetivamente) e risco moderado/alto (985, 11.2% e 334, 11.2%). Ofensores negros obtiveram um valor de risco do LSI-R em media de 24.17, superior ao obrtido pelos ofensores brancos, 23.78.

Tabela 2.1: Resultados obtidos relativos ao nível de risco de cada raça presente neste estudo.

Raça	Baixo	Baixo/Moderado	Moderado	Moderado/Alto	Alto
Caucasianos	1.117 (12.7%)	3.168 (36%)	3.358 (38.1%)	985 (11.2%)	183 (2.1%)
Afro-Americanos	41 (1.4%)	1.102 (37.0%)	1.219 (40.9%)	334 (11.2%)	98 (1.4%)

Com os resultados obtidos os autores afirmam que ofensores brancos receberam sentenças mais longas quando classificados com baixo risco (22.08), seguido pelo risco baixo/moderado (19.89) e moderado (19.89) e sentenças mais curtas quando classificados com risco moderado/alto (19.25) e alto (19.28). Em contraste esta tendência foi menos pronunciada em ofensores negros, que receberam sentenças mais curtas quando classificados com risco moderado (19.50) e sentenças mais longas com risco moderado/alto (19.58) e alto (19.56). Em

semelhança com ofensores brancos a sentença media foi mais longa com classificação de risco baixo (20.39) e baixo/moderado (19.99)

Tabela 2.2: Sentença, em meses, obtida em relação com o nível de risco

Raça	Baixo	Baixo/Moderado	Moderado	Moderado/Alto	Alto
Caucasianos	22.08	19.89	19.89	19.25	19.28
Afro-Americanos	20.39	19.99	19.50	19.58	19.56

Os autores concluíram que poderá existir uma potencial injustiça racial com o uso do LSI-R mas os seu efeitos são pequenos. Foram também encontrados poucos indícios de viés racial na habilidade do LSI-R prever consequências comunitárias entre ofensores brancos e negros.

2.2.3 *Can 14,737 women be wrong? A meta-analysis of the LSI-R and recidivism for female offenders*

Paula Smith, Francis T. Cullen and Edward J. Latessa (2009)[15] conduziram uma investigação ao uso do LSI-R para calcular o valor de risco para ofensores do sexo feminino. O seu método consistiu em usar um conjunto de estudos e dados relevantes ao seu tópico, obtidos com uma pesquisa em bases de dados de literatura de todos os estudos do LSI-R. Obtiveram também dados não publicados através de investigadores com artigos publicados anteriormente sobre o LSI-R. No total usaram 25 bases de dados publicadas e não publicadas na sua analise o que equivaleu a 14.737 ofensores do sexo feminino.

Os autores deste artigo concluíram que a relação entre o LSI-R e a reincidência criminal para o sexo feminino é estatisticamente e praticamente similar a reincidência para o sexo masculino. Os resultados deste estudo são também similares a resultados de analises anteriores que usavam amostras mistas ou dominadas pelo sexo masculino. Concluíram então, com base na sua analise que o LSI-R é praticamente o mesmo para ofensores do sexo masculino como para ofensores do sexo feminino.

2.2.4 *The Level of Service Inventory — Revised With English Women Prisoners*

Um estudo conduzido por Emma J. Palmer and Clive R. Hollin (2007)[16] também examinou o LSI-R com uma amostra de ofensores do sexo feminino. Os resultados mostraram diferenças no perfil de necessidade criminológicas para os dois géneros. Ofensores masculinos obtiveram valores superiores em algumas sub-escalas, como histórico criminal e lazer, enquanto que ofensores femininos obtiveram valores superiores em sub-escalas relacionadas com família, relações conjugais e problemas pessoais e emocionais. Apesar destas diferenças, o risco global não divergiu entre o masculino e feminino. Os autores concluíram que o valor final do LSI-R era altamente preditivo de reincidência para ambos os géneros, mas que é necessário identificar corretamente as necessidades criminológicas de ofensores do sexo feminino para que seja possível fornecer serviços e intervenções apropriadas para cada caso a ser tratado.

2.2.5 *Does The LSI-R Have Utility For Sex Offenders?*

L. M. Ragusa-Salerno, M. Ostermann, e S. S. Thomas estudaram o problema do LSI-R no que toca a abusadores sexuais depois se ser confirmado pelo *New Jersey State Parole Board (SPB)* que os resultados do LSI-R eram desconsiderados na toma de decisões sobre abusadores sexuais, pois era presumido que o instrumento sobre-classificava-os como risco baixo e não previa corretamente ofensas sexuais. Este estudo explorou a utilidade da ferramenta na população de abusadores sexuais. Foi investigada a precisão do LSI-R para reincidência sexual bem como para reincidência em geral, com o uso de uma amostra de indivíduos libertados dos estabelecimentos criminais de *New Jersey* entre 2004 e 2006. Neste período de tempo foram libertados um total de 37.298 ofensores. Como grande parte dos instrumentos de avaliação de risco para abusadores sexuais forma desenvolvidos com adultos do sexo masculino em mente e o LSI-R foi inicialmente desenvolvido e validado com amostras predominantemente masculinas, todos os ofensores do sexo feminino forma retirados da amostra dos autores. No final foi analisada uma amostra composta por 21.298 ofensores do sexo masculino.

O estudo explorou a habilidade de previsão do LSI-R para uma variedade de outros crimes para além de crimes sexuais para fornecer conhecimento de como o LSI-R se compara na previsão de crimes sexuais e crimes não sexuais. Foram considerados como crimes sexuais se durante a detenção ocorreu alguma ofensa sexual. Foram considerados como crime violentos se durante a detenção ocorreu alguma ofensa de natureza violenta. Qualquer ofensa não sexual e não violenta foi incluída em outro tipo de reincidência.

Quase 70% de todos os ofensores foram reaprendidos por qualquer ofensa no prazo de 5 anos. No que toca a amostra de abusadores sexuais, no tempo da libertação a idade media era 41 anos. Aproximadamente 60% eram negros, 22.37% eram brancos e 17.6% hispânicos. Quase 64% dos abusadores sexuais foram reaprendidos no período de 5 anos depois da libertação, 1.81% por crimes sexuais, 21.6% por um crime violento e 61.86%por outra ofensa.

O estudo concluiu que o LSI-R de facto não foi útil a prever crimes sexuais, mas teve utilidade para prever reincidência não sexual para abusadores sexuais. Os autores concluíram que o LSI-R pode ser usado em conjunto com instrumentos validados para abusadores sexuais para melhor identificar os riscos e necessidades destes ofensores fora de crimes sexuais, como mostra a investigação abusadores sexuais são mais prováveis de reincidir com um crime geral do que com crimes sexuais.

2.2.6 *Testing the Level of Service Inventory–Revised (LSI-R) for Racial/Ethnic Bias*

O uso de instrumentos de classificação de risco em estabelecimentos prisionais é cada vez mais maior, mas foram levantados algumas preocupações de que estes instrumentos sobre classificar ofensores pertencentes a minorias. Este estudo realizado, por Kevin W. Whiteacre[17], examina o LSI-R no que toca a diferenças raciais ou éticas na sua classificação.

Neste estudo foi usada uma amostra que consistia em 523 ofensores masculinos que deram entrada no estabelecimento prisional depois de 1 de janeiro de 2002 e deram saída ate 31 de

dezembro de 2003. Ofensores do sexo feminino foram excluídos pois o género pode gerar um efeito que poderá alterar os resultados para a raça e etnia.
O autor decidiu utilizar valores de corte diferentes dos recomendados. O manual do LSI-R recomenda os valores de 16 ou maior para o risco máximo, 8 a 15 para risco médio e 0 a 7 para risco mínimo. O recomendado para classificação institucional é 37 ou maior para risco máximo, 25 a 36 para risco médio e 0 a 24 para risco mínimo. O autor deste artigo uma dois valores diferentes para o corte entre risco baixo e risco elevado, 25 e 16. Numa primeira experiência valores superiores a 25 eram considerados risco elevado e valores menores que 25 eram considerados risco baixo. Numa segunda experiência valores superiores a 16 eram considerados risco elevado e valores menores que 16 eram considerados risco baixo.
Para o valor de corte de 25 foram obtidos resultados corretos para 78.9% dos Afro-Americanos, 87.0% de Caucasianos e 86.8% de Hispânicos. Afro-Americanos foram um pouco mais prováveis de serem sobre classificados do que Caucasianos e Hispânicos, 12.2%, 9.0% e 7.9% respetivamente. Porém Afro-Americanos foram também mas prováveis de serem sub classificados do que Caucasianos e Hispânicos, 9.0%, 4.0% e 5.3% respetivamente.
Para o valor de corte de 16 foram obtidos resultados corretos para 55.2% de Afro-Americanos, 70.1% de Caucasianos e 75.0% de Hispânicos. Uma diminuição em precisão era esperada para este valor de corte, mas esta mudança não afetou igualmente os 3 grupos raciais. Afro-Americanos eram altamente mais prováveis de serem sobre classificados do que Caucasianos ou Hispânicos, 42.7%, 27.7% e 25.0% respetivamente.
O autor conclui que para além do instrumento de classificação, a instituição ou programa onde está inserido tem um papel significativo nos resultados obtidos. São definidos pelo autor 3 fatores definidos localmente, os valores de corte, a medida a ser avaliada e o propósito do instrumento dentro de um maior sistema de classificação. Em termos de resultados o autor conclui que existe uma tendência em existirem mais erros de classificação para Afro-Americanos do que para Caucasianos ou Hispânico. Instituições que usem estes instrumentos tem a responsabilidade de efetua uma validação local para garantir o uso do instrumento não produzirá erros de classificação para grupos étnicos diferentes, de acordo com o autor deste estudo.

2.3 COMPAS

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) é um instrumento de apoio à decisão desenvolvido pela empresa Northpointe e é usado para avaliar a probabilidade de um ofensor reincidir em crime depois de libertado. Sistemas deste tipo são defendidos como sendo mais precisos do que humanos na altura de tomar decisões, porém o COMPAS tem sido alvo de criticas por não ser totalmente imparcial e perpetuar o desfavorecimento racial presente no sistema criminal.
A recolha de dados pode ser feita de algumas formas diferentes. O ofensor pode preencher um formulário sozinho com os seus dados, pode ser feita um entrevista estruturada onde as questões são feitas verbalmente, ou o entrevistador pode usar uma discussão guiada para recolher os dados necessários.

O COMPAS usa ambos dados estáticos e dinâmicos para gerar os resultados. O uso de medidas dinâmicas permite que mediadas tomadas sejam alteradas com o decorrer do tempo à medida que o comportamento do ofensor altera. Fatores dinâmico permitem também que sejam sobrepostos avaliações anteriores na avaliação corrente para verificar alterações nos valores de risco e necessidades.
No que toca a valores de risco o COMPAS divide o estes valores em 10 níveis iguais, sendo que o primeiro nível indica que o valor de risco está entre 0% e 10%, o segundo nível de 10% a 20% e assim sucessivamente até ao nível 10 que representa o risco máximo.
o COMPAS é considerado um instrumento de avaliação de quarta geração. Caraterizado por:

- Uma maior seleção de teorias explicativas;
- Um maior conjunto de fatores de risco e necessidades;
- Modelos estatísticos mais avançados;
- Implementação do domínio de risco e necessidades com sistemas de gestão de informação, bases de dados de justiça criminal e implementações web da tecnologia de avaliação.

Toda esta informação está presente num documento de *Frequently Asked Questions (FAQ)*[18] publicado pela Northpointe, desenvolvedora do COMPAS, também composto por informação mais detalhada e não presente nesta breve introdução à ferramenta.

2.3.1 How We Analyzed the COMPAS Recidivism Algorithm - ProPublica

Em 2016, ProPublica, uma organização focada em jornalismo de investigação, decidiu testar a validade do instrumento de avaliação de risco COMPAS[4]. Nesta investigação os autores tinham como objetivo aferir se o algoritmo tinha de facto uma inclinação para favorecer certos grupos de ofensores. Foram obtidos dados de 11.757 ofensores com entrada no estabelecimento prisional, de Broward Florida, entre 2013 e 2014. O valor final dado pelo COMPAS era contido num intervalo de 1 a 10, em que 10 era o risco mais elevado, 1 a 4 foi considerado risco baixo, 5 a 7 risco médio e 8 a 10 risco alto.
Numa primeira analise os autores verificaram o valor de risco de reincidência entre ofensores brancos e ofensores negros. Com esta analise foi possível verificar ofensores brancos tinham a inclinação para categorias de baixo risco enquanto que ofensores negros estavam distribuídos uniformemente entre todas as categorias.
Esta amostra foi composta por 3.175 ofensores negros e 2.103 ofensores brancos, 1.175 eram do sexo feminino e 4.997 do sexo masculino, 2.809 ofensores reincidiram em crime.

No que toca reincidência violenta, também foi apresentada alguma disparidade nos resultados entre ofensores brancos e ofensores negros, embora não seja tão pronunciada como nos

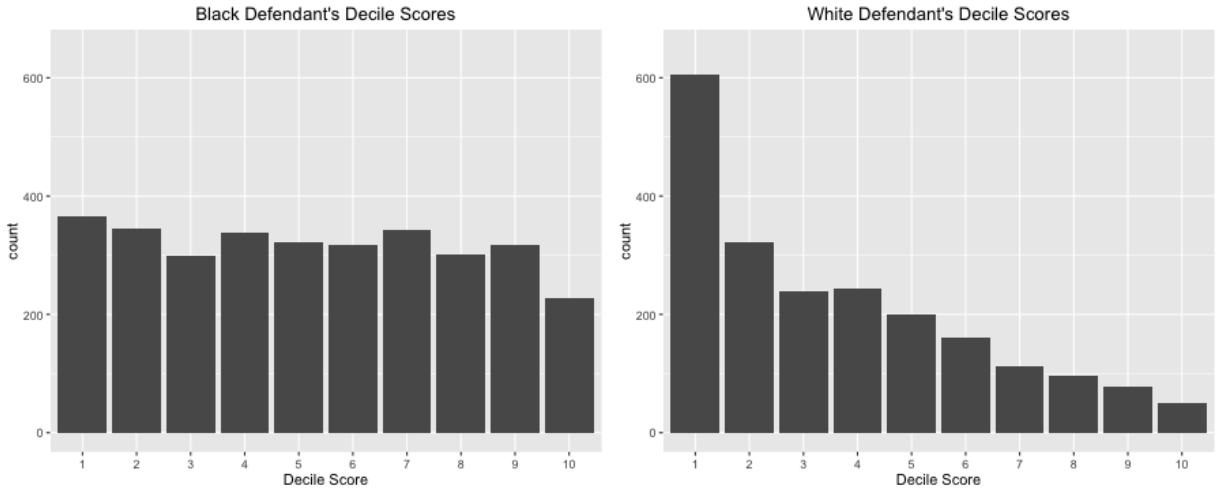

Figura 2.1: Resultados obtidos de reincidência na investigação conduzida pela organização Propublica[4]

resultados de reincidência geral, é possível verificar que são classificados em maior numero no nível de risco mais baixo.

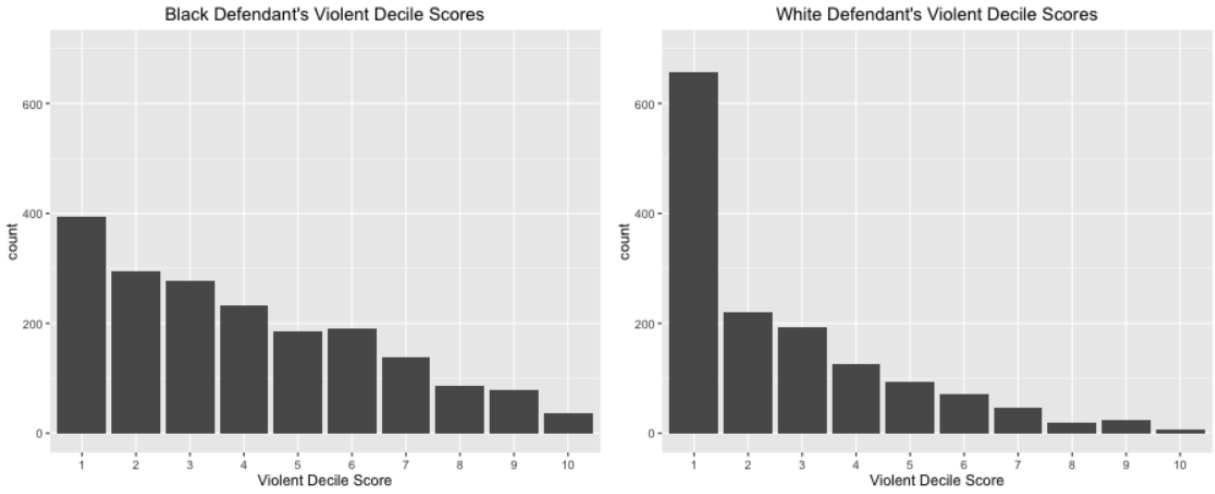

Figura 2.2: Resultados obtidos de reincidência violenta na investigação conduzida pela organização Propublica[4]

Para testar outros fatores os autores da investigação criaram um modelo de regressão que considerava outras características como raça, idade, histórico criminal, reincidência futura, grau do crime e género.

Os resultados obtidos com este modelo indicaram que o fator preditivo de maior nível de risco foi a idade. Ofensores com idades menores que 25 anos eram 2.5 vezes mais prováveis de obterem um valor de risco superior a ofensores com idades superiores. A raça do ofensor também foi um fator preditivo de valores de risco maiores. Ofensores de raça negra eram 45% mais prováveis de ter maior risco do que ofensores brancos. Ofensores de sexo feminino, mesmo com níveis baixo de criminalidade, eram 19.4% mais prováveis de serem classificados com valores de risco superiores a ofensores de sexo masculino.

Contando com reincidência violenta, a idade do ofensor foi um fator preditivo ainda mais forte. Ofensores mais novos eram 6.4 vezes mais prováveis de obterem um valor de risco

superior do que ofensores mais velhos. A raça do ofensores continuou a ser um fator preditivo forte para um valor de risco de reincidência violenta maior. Ofensores negros foram 77.3% mais prováveis de obter valores superiores de risco.
Por fim os autores testaram a ferramenta para erros de classificação, falsos positivos e falsos negativos. Os resultados desta experiência indicaram que o algoritmos era mais provável classificar erradamente ofensores negros do que ofensores brancos. Ofensores negros que não reincidiram foram quase 2 vezes mais prováveis de serem classificados pelo COMPAS com risco elevado (45%), em comparação com ofensores brancos(23%). No entanto, ofensores negros com valores de risco alto reincidiram um pouco mais que ofensores brancos (63% e 59% respetivamente).

Tabela 2.3: Resultados para ofensores negros

	Low	**High**
Não Reincidiram	990	805
Reincidiram	532	1.369
Total	1.522	2.174

Para ofensores brancos, foram observados os erros opostos. Era mais provável prever que ofensores brancos não iriam reincidir em comparação com ofensores negros. COMPAS classificou os ofensores brancos como risco baixo 70.5% mais que ofensores negros.

Tabela 2.4: Resultados para ofensores brancos

	Low	**High**
Não Reincidiram	1.139	349
Reincidiram	461	505
Total	1.600	854

Para reincidência violenta os resultados forma similar. Ofensores negros foram 2 vezes mais prováveis de serem classificados com risco mais elevado erradamente, em comparação com ofensores brancos. Estes ofensores mal classificados como risco baixo 63.2% mais do que ofensores de raça negra. Ofensores negros que foram classificados como risco elevado reincidiram um pouco mais que ofensores brancos, 21% e 17% respetivamente.

Tabela 2.5: Resultados para todos os ofensores

	Ofensores Negros		**Ofensores Brancos**	
	Low	**High**	**Low**	**High**
Não Reincidiram	1.692	1.043	1.679	380
Reincidiram	170	273	129	77
Total	1.868	1.316	1.808	457

2.3.2 THE LSI-R AND THE COMPAS - Validation Data on Two Risk-Needs Tools

No ano de 2008 foi conduzido um estudo por T. L. Fass, K. Heilbrun, D. DeMatteo, and R. Fretz[19], com o objetivo de comparar a validade de previsão de ambos os algoritmos de apoio à decisão, COMPAS e LSI-R. Os autores tinham 3 objetivos com estes estudo. Primeiramente

fornecer o primeiro, na altura, estudo empírico da ferramenta COMPAS. Em segundo descrever as variáveis criminológica mais relacionadas com reincidência num período de 1 ano após a libertação. Por ultimo, comparar a validade do COMPAS e o LSI-R de previsão de reincidência num período de 1 ano após a libertação.
A amostra utilizada pelos autores foi composta por 975 ofensores masculinos (COMPAS 276, LSI-R 696). O intervalo idades dos participantes era dos 18 aos 63 anos com uma media de 32.5 anos. No que toca a etnias, 71.4% eram Afro-Americanos, 15% Hispânicos ou Latinos e 13.6% Caucasianos.
Os resultados deste estudo mostram que o algoritmo LSI-R previu corretamente no total 48.4% dos resultados, 80.4% da população de Caucasianos, 43.3% da população de Afro-Americanos, e 82.4% da população de Hispânicos. Ofensores Afro-Americanos (51.8%) foram mais prováveis de serem classificado como falsos positivos do que Caucasianos (7.6%) ou Hispânicos (0%). Hispânicos (17.7%) e Caucasianos (12%) foram mais prováveis de serem falsos negativos do que Afro-Americanos (4.78%).
Como algoritmo COMPAS foram previstos corretamente 85% dos resultados no total, 97.6% para a população de Caucasianos, 90.9% da população de Hispânicos e 76.4% para a população de Afro-Americanos. Afro-Americanos (7.32%) foram mais prováveis de serem falsos positivos do que Caucasianos (0%) ou Hispânicos (0%). Em semelhança ofensores Afro-Americanos (16.2%) também foram mais prováveis de serem falsos negativos em comparação com Hispânicos (9.1%) ou Caucasianos (2.4%).
Os autores deste estudo concluem que estes resultados indicam que existe um falha de precisão de previsão entre grupos étnicos diferentes. Afro-Americanos foram mais prováveis de serem sobre classificados em comparação com Hispânicos e Caucasianos, previsão indica que vão reincidir quando na verdade não reincidiram. Para o LSI-R Hispânicos e Caucasianos foram mais prováveis de serem sub classificados do que Afro-Americanos, previsão indica que não haverá reincidência quando na realidade houve reincidência. Para o algoritmo COMPAS o grupo de Afro-Americanos foram em maior probabilidade sub classificados.
Em conclusão estes resultados indicam uma tendência para sobre classificar ou sub classificar ofensores dependendo da sua etnia. Os autores apontam a preocupação sobre o uso de instrumentos de avaliação de risco em certas populações. Finalmente os autores acrescentam que o estudo foi realizado com dados de de ofensores masculinos maioritariamente Caucasianos, Afro-Americanos e Hispânicos, por esta razão os resultados obtidos não podem ser generalizados para ofensores femininos ou de raça e etnias diferentes.

2.4 Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia

Num artigo publicado em 2019 intitulado *Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia*[20], os autores têm como objetivo estudar as limitações de *machine learning* para a previsão de reincidência juvenil. Para este estudo foram utilizados modelos de *machine learning* (ML) em conjunto com o SAVRY, uma ferramenta estruturada de avaliação de risco, numa base de dados de origem

na Catalunha.
SAVRY (*Structured Assessment of Violence Risk in Youth*) é uma ferramenta de avaliação de risco usada para ofensores jovens, idades entre os 12 e 18 anos. SAVRY é composta por 24 itens separados em 3 domínios de risco, fatores de risco históricos, fatores de risco sociais/contextuais e fatores de risco individuais/clínicos. Cada um destes fatores é classificado num de três níveis de risco, baixo, moderado e alto. Para além destes 24 fatores de risco, SAVRY, também inclui 6 fatores protetivos que são classificados como presentes ou ausentes.
O método neste artigo começa pela recolha de dados. Foi utilizado um *dataset* de ofensores que terminaram a sua sentença em 2010 no sistema de justiça juvenil de Catalonia, para um total de 4753 ofensores. Os crimes terão sido cometidos entre 2002 e 2010 quando os ofensores tinham entre 12 e 17 anos. Para observação de reincidência estes ofensores foram novamente avaliados em 31 de dezembro de 2013 e 2015. Os autores optaram por apenas utilizar uma sub amostra de ofensores que forma submetidos a uma avaliação com o SAVRY, um total de 855 jovens ofensores.
Os dados recolhidos não eram compostos por um conjunto para teste. Para dividir os dados em conjuntos de treino e de teste os autores utilizaram *k-fold cross validation*. Os dados de validação foram escolhidos do conjunto de treino. Foram utilizados vários algoritmos de *machine learning* como:

- *Logistic Regression (logit)*;
- *Multi-Layer Perceptron (mlp)*;
- *Suport Vector Machine com Kernel Linear (lsvm)*;
- *Suport Vector Machine com Kernel Radial (rsvm)*;
- *K-Nearest Neighbors (knn)*;
- *Random Forest (rf)*;
- *Naive Bayes (nb)*.

Sendo que destes apenas foram apenas reportados os algorimtos com melhores resultados, que correspondem ao *Logistic Regression (logit)* e *Multi-Layer Perceptron (mlp)*
Dependendo das *features* escolhidas, foram conduzidas 4 expriencias, a primeira denominada "*Static ML*", correspondia a *features* estáticas como histórico criminal, género, nacionalidade, número de crimes anteriores e tipo de crime.
A segunda experiência, "*SAVRY ML*", correspondia a todas as *features* do SAVRY nomeadamente, a avaliação final, os 24 fatores de risco, as 5 medias das características do individuo bem como o programa no qual estava inserido. durante a avaliação do SAVRY.
A terceira experiência, "*Static + SAVRY ML*" correspondia como o nome indica, o uso em conjunto das *features* estáticas e do SAVRY.
Como base, os autores, somaram todos as pontuações de todos os fatores de risco do SAVRY sem uso de *machine learning* denotada por "*SAVRY SUM*" em adição com a avaliação obtida por especialistas, "*Expert*".

No que toca a resultados, os autores compararam o *"SAVRY ML"* com as outras experiências com e sem *features* do SAVRY. Foi observado que ao não incluir *features* demográficas e de histórico criminal a precisão diminui em todos os métodos. Combinar *features* do SAVRY com demográficos estáticos, ou aumentar o conjunto de treino fez com que fossem obtidos melhores resultados.

Para além do desempenho os autores também estavam interessados em medir o nivel de discriminação dos métodos de *machine learning*. Foi analisada a equidade em termos de genero e nacionalidade. Foram selecionados os modelos de *machine learning* com melhores resultados e comparados com os resultados do *"SAVRY SUM"* e *"Expert"*. Em termos de justiça em relação ao género, o *"SAVRY SUM"* está dentro dos limites de equidade em termos da avaliação especialista. Os modelos *mlp* e *logit* foram menos prováveis de classificar erradamente ofensores do género feminino como reincidentes do que ofensores masculinos. Com *"SAVRY ML"* as mulheres tem menos probabilidade de serem classificadas erradamente como não reincidentes. Os métodos de *machine learning*, enquanto que estão dentro do intervalo aceitável quando usam *features* do SAVRY, tornam-se discriminatórios quando usam *features* demográficas, com as mulheres serem classificadas como não reincidentes em maior probabilidade.

No que toca a nacionalidade, foi observado que os métodos de *machine learning* obtiveram uma disparidade mais elevada em comparação com o *"SAVRY Sum"* e a avaliação especialista. Estrangeiros têm maior probabilidade de de serem falsamente classificados como reincidentes e são menos prováveis de serem classificados como não reincidentes. Os dados referentes a estrangeiros foram consequentemente divididos em sub grupos, *"Maghrebi"*, *"Latin American"*, *"European"* e *"Other"*. Um sistema pode ser justo para um grupo mas pode descriminar para outro certo sub grupo. O grupo *"European"* e *"Other"* como são mais pequenos, 37 e 13 respetivamente, foi excluídos da analise, de notar que estes grupos são mais prováveis de serem descriminados positivamente. Métodos de *machine learning* obtiveram mais disparidade em relação a *"Maghrebi"* e *"Latin Americans"* para todas as métricas quando incluíam *features* não SAVRY no treino. Treino com itens do SAVRY tem uma disparidade um pouco mais alta do que *"SAVRY Sum"*. Para o grupo *"Latin Americas"* a disparidade está dentro dos níveis aceitáveis mas passa estes níveis para o grupo *"Maghrebi"*. Em todas as experiências o *logit* é mais injusto do que o *mlp* quando treinado com *features* não SAVRY.

Em conclusão, os métodos de *machine learning* obtém um melhore desempenho preditivo, contudo estes métodos, *logit* e *mlp* são discriminatórios no que toca a género e nacionalidade. A analise presente neste estudo da importância de *features* mostra que quando fatores estáticos estão em combinação com SAVRY os métodos optam por depender mais nas *features* do SAVRY. Os autores retiram deste estudo que o SAVRY é em geral justo, enquanto que modelos de *machine learning* tendem a discriminar contra ofensores masculinos, estrangeiros ou ofensores pertencentes a grupos nacionais específicos.

2.5 Algorithmic Decision Making and the Cost of Fairness

Um artigo realizado por Marzieh Karimi-Haghighi e Carlos Castillo em 2021 intitulado de *Enhancing a Recidivism Prediction Tool with Machine Learning: Effectiveness and Algorithmic Fairness*[21] aborda como a aplicação de *machine learning* pode ser usada para aumentar a eficácia de ferramentas de avaliação de risco em sistemas criminais sem introduzir viés indesejado. Neste estudo é usada a ferramenta RisCanvi e a as duas dimensões em estudo são a precisão preditiva e a equidade destes algoritmos.
RisCanvi é uma ferramenta de avaliação de risco introduzida inicialmente no sistema prisional da Catalunha em 2009. Este é aplicado varias vezes durante o período em que o ofensor está inserido no estabelecimento prisional. Dados para o RisCanvi ao contrário de outras ferramentas semelhantes, não são obtidos através de um questionário mas sim a partir de uma entrevista conduzida por profissionais. Existem duas versões desta ferramenta, o RisCanvi-S, uma versão abreviada com apenas 10 itens de risco e o RisCanvi-C, a versão completa com todos os 43 itens de risco. Estes itens de risco podem ser categorizados em 5 categorias diferentes:

- Criminal;
- Biográfica;
- Família/Social;
- Clínica;
- Atitudes/Personalidade.

Estes itens podem também ser divididos em fatores estáticos como histórico criminal ou fatores dinâmicos como comportamentos anti-sociais e atitudes em geral.
Como *dataset* os autores primeiramente obtiveram um conjunto de 7.239 ofensores que entraram no sistema prisional entre 1989 e 2012 e que forma avaliados com o RisCanvi entre 2010 e 2013. O numero de ofensores no conjunto diminui para 2.634 pois os autores apenas estavam interessados em ofensores com informação sobre a nacionalidade registada. A população foi depois filtrada em termos de reincidência violenta e geral, liberdade e em termos da data da ultima avaliação RisCanvi. A amostra final era composta por 2.067 ofensores dos quais 146 reincidiram em crimes violentos e 310 reincidiram em crime geral.
O estudo foi focado na versão RisCanvi completa (RisCanvi-C) que consiste em mais fatores de risco e os resultados são divididos em 3 níveis de risco, baixo, médio e alto.
Vários métodos de *machine learning* foram usados como *logistic regression, multi-layer perceptron (MLP)* e *support vector machines (SVM)*. Foram usados sub conjuntos de *features* como entrada para os modelos de *machine learning* como os 43 fatores de risco do RisCanvi-C, fatores de risco para reincidência violenta/geral e um conjunto de *features* selecionadas dos 43 fatores de risco. Adicionalmente foram usada 3 *features* demográficas, género, nacionalidade e idade.
Com este estudo os autores compararam a eficácia e equidade de métodos de *machine learning* e a ferramenta de avaliação de risco RisCanvi. Em termos de eficácia os métodos

de *machine learning* obtiveram resultados um pouco melhores. No entanto o uso destes métodos pode introduzir discriminação entre grupos como ofensores estrangeiros e ofensores nacionais, ofensores jovens e ofensores mais velhos. Resultados obtidos mostram que procedimentos que mitigação do viés podem, de forma substancial, reduzir taxas de falsos positivos entre múltiplos grupos. Com base nestes resultados os autores propõem que métodos de *machine learning* não devem ser introduzidos como ferramentas de previsão de reincidência sem a aplicação de procedimentos de mitigação de viés. Apesar de produzirem um desempenho um pouco melhor em comparação com a ferramenta de avaliação de risco, é possível introduzirem níveis de discriminação na sua avaliação.

2.6 Conclusão

Neste capitulo foram apresentados vários sistemas de avaliação de risco usados em sistemas de gestão de ofensores. Em primeiro lugar foi apresentado o LSI-R, uma ferramenta de terceira geração desenvolvida nos anos 80 e posteriormente atualizada nos anos 90. Foi apresentado o seu método de avaliação de risco e seguidamente alguns artigos publicados sobre referentes à justiça do mesmo. Estes artigos incluem o tema de desfavorecimento racial, o seu uso para ofensores do sexo feminino e ofensores sexuais.
De seguida foi apresentado a ferramenta COMPAS uma das mais investigadas atualmente. Foi feita uma breve apresentação inicial seguida de uma investigação conduzida pela organização ProPublica que analisava a validade do instrumento para diferentes grupos étnicos. Foi também apresentado um artigo que tinha como objetivo comparar o COMPAS com o LSI-R.
Após esta secção foram analisados mais artigos que utilizavam instrumentos diferentes também para verificar a sua justiça perante diferentes etnias e grupos de ofensores.
Depois deste estudo do estado da arte e, tendo em conta os resultados apresentados, é possível concluir que poderá existir uma potencial injustiça racial em alguns destes instrumentos, nomeadamente no instrumento COMPAS. Por vezes esta disparidade pode ser eliminada ao identificar as diferentes necessidade presentes no ambiente e adaptar o instrumentos para estas necessidades. É possível também concluir que métodos de *machine learning* usados no mesmo contexto por vezes podem também em si ser discriminatórios e devem ser usados em conjunto com métodos de mitigação de viés.

Capítulo 3

Planeamento e Trabalho a Realizar

3.1 Objetivos

O trabalho a ser desenvolvido nesta dissertação pode ser dividido em duas partes. Em primeiro lugar irá ser feita uma análise e comparação de *datasets* de ofensores com o seu nível de risco para diferentes instrumentos de avaliação de risco.
Em segundo lugar irá ser feita o calculo de nível de reincidência com o uso de modelos de *machine learning* com o uso de vários métodos.
Por fim os resultados obtidos em ambas as partes, a analise de instrumentos de avaliação de risco e o uso de modelos de *machine learning*, serão analisados em detalhe e comparados para averiguar qual o método mais justo e menos imparcial e quais as caraterísticas e detalhe a ter em atenção quando qualquer um destes métodos é usado.

3.2 Instrumentos de Avaliação de Risco

O trabalho a ser desenvolvido nesta dissertação irá envolver, numa primeira parte, a análise de *datasets* compostos por informação referente a ofensores que tenham sido avaliados com instrumentos de avaliação de risco, como por exemplo LSI-R ou COMPAS. Em primeira instância irá ser retirada uma amostra representativa dos dados presentes nos *datasets*. Com esta amostra será feita uma comparação de níveis de risco tendo em conta diferentes caraterísticas como a idade, género ou etnia. Poderá ser feita uma comparação entre diferentes amostras e entre diferentes instrumentos de avaliação de risco para aferir se de facto existe uma imparcialidade nestes instrumentos.

3.3 Modelos de *Machine Learning*

Numa segunda parte vão ser usados modelos de *machine learning*, treinados em *datasets* públicos, para aferir o nível de risco de reincidência de ofensores. Neste passo vão ser implementados vários métodos de *machine learning*, para descobrir qual o método que obtém melhores resultados.
Este passo tem como objetivo verificar se a utilização de inteligência artificial resulta em predições mais imparciais e corretas do que com apenas o uso de métodos de avaliação de risco testados no passo anterior.
Com base no estudo feito para o estado da arte estes métodos de *machine learning* ainda que apresentem um desempenho superior a instrumentos de avaliação de risco, é possível que seja introduzido um nível de discriminação na sua avaliação. É recomendado que este

nível de discriminação seja removido com métodos de mitigação de viés em conjunto com os modelos de *machine learning*.

3.4 Análise Final

Numa ultima análise deste trabalho irá ser feita a comparação de resultados obtidos dos instrumentos de avaliação de risco com os resultados obtidos dos modelos de *machine learning*. Irá ser feita a comparação de desempenho em geral, se os resultados são ou não corretos e se existe alguma imparcialidade em cada um dos métodos.
Também será feita uma análise de possíveis alterações e aspetos a ter em conta quando é feita esta avaliação de risco em ofensores reais.

Capítulo 4

Tecnologias Utilizadas

4.1 Introdução

Neste capitulo intitulado como *Tecnologias Utilizadas*, serão apresentadas todas as tecnologias usadas para o desenvolvimento desta dissertação, bem como uma breve descrição e de que forma foram utilizadas. As tecnologias apresentadas são Python, Sklearn, Pycharm, Swagger e Postman

4.2 Python

Python[22] é um linguagem de programação de alto nível. Criada por Guido van Rossum[23] e com a sua primeira iteração em 1991. O seu *design* realça a leitura de código com o uso de *whitespace* significativo. A sua construção e abordagem a orientação a objetos aponta para ajudar programadores a escrever código claro e lógico para projetos de todas as escalas.
Esta linguagem foi utilizada para a construção dos modelos de *machine learning* e analise de resultados dos mesmos. No capitulo de Desenvolvimento [5] estes modelos serão descridos em maior detalhe bem como os resultados obtidos.

4.3 Sklearn

Sklearn[24] faz parte de um modulo Python, *scikit-learn* usado para *machine learning* construido a partir do SciPy[25], uma biblioteca Python *open-source* usada para computação cientifica e técnica. O projeto *scikit-learn* foi iniciado por David Cournapeau[26] em 2007 como um projeto inserido no *Google Summer of Code*[27], desde então vários voluntários contribuíram para o seu desenvolvimento.
Este modulo foi usado para a criação, treino e teste de todos os modelos presentes nesta dissertação. Foi também este modulo que possibilitou o uso de diferentes métodos de *machine learning*, a analise da importância de cada uma das *features* selecionadas para a previsão e toda a analise de desempenho dos ditos modelos.

4.4 Pycharm

Pycharm é um IDE (*Integrated Development Environment*) usados para programar em Python. Fornece analise de código, um *debugger* gráfico, testes integrados e suporta desenvolvimento Web. É *cross-platform*, o que significa que funciona em sistemas operativos Windows, macOS e Linux. Foi desenvolvido em 2010 pela empresa JetBrains[28] e tornou-se *open-source* a 22 de outubro de 2013.

Este IDE foi utilizado para toda a programação do código desenvolvido no decorrer desta dissertação.

4.5 Swagger

Swagger[29] foi criado em 2011 por Tony Tam[30], co-fundador do site *Wordnik*. Esta ferramenta foi desenvolvida para combater a necessidade para a automação da criação de documentação de APIs.
Swagger pode ser utilizado para desenvolver APIs de raiz, interagir com APIs já construidas e documentar essa mesmas APIs.
Swagger foi utilizado para a construção da API a ser discutida na secção de *Trabalho Adicional*6.4.

4.6 Postman

Postman[31] é uma plataforma que permite que desenvolvedores desenhem, construam e testem as suas APIs. Esta plataforma começou em 2012 como projeto secundário de Abhinav Asthana[32] que queria simplificar o teste de APIs.
Swagger foi utilizado para os testes da API construida com o *swagger*, a ser discutida na secção de *Trabalho Adicional*6.4.

Capítulo 5

Desenvolvimento

5.1 Introdução

Neste capitulo intitulado como *Desenvolvimento* vai ser apresentado todo o desenvolvimento pratico desta dissertação. Vão ser apresentados os *datasets* utilizados, como foi feito o tratamento dos dados, as *features* utilizadas e o *target* de cada um. Vão ser também apresentados os modelos de *machine learning* utilizados, como foram implementados e os resultados obtidos. Será também apresentada a analise feita a cada um dos *datasets* e os resultados obtidos em forma de gráficos e tabelas.

5.2 *Datasets*

Nesta primeira secção intitulada *Datasets* vão ser apresentados todos os *datasets* utilizados, com uma breve descrição, as colunas existentes e as colunas selecionadas para a construção dos modelos de *machine learning*.

5.2.1 *NIJ Recidivism Challenge Dataset*

O primeiro *dataset* é intitulado como ”NIJ Recidivism Challenge Dataset” [33]. Os dados foram fornecidos pelo *Georgia Department of Community Supervision*.
Este *dataset* é constituído por 54 colunas e 25835 linhas. As colunas incluem características como o género, idade na altura de saída, informação sobre uso de drogas e detenções anteriores e a reincidência no primeiro ano, segundo e terceiro.

- ID
- Gender
- Race
- Age_at_Release
- Residence_PUMA
- Gang_Affiliated
- Supervision_Risk_Score_First
- Supervision_Level_First
- Education_Level
- Dependents
- Prison_Offense
- Prison_Years
- Prior_Arrest_Episodes_Felony
- Prior_Arrest_Episodes_Misd
- Prior_Arrest_Episodes_Violent
- Prior_Arrest_Episodes_Property
- Prior_Arrest_Episodes_Drug
- Prior_Arrest_Episodes_PPViolationCharges
- Prior_Arrest_Episodes_DVCharges

- Prior_Arrest_Episodes_GunCharges
- Prior_Conviction_Episodes_Felony
- Prior_Conviction_Episodes_Misd
- Prior_Conviction_Episodes_Viol
- Prior_Conviction_Episodes_Prop
- Prior_Conviction_Episodes_Drug
- Prior_Conviction_Episodes_GunCharges
- Prior_Revocations_Parole
- Prior_Revocations_Probation
- Condition_MH_SA
- Condition_Cog_Ed
- Condition_Other
- Violations_ElectronicMonitoring
- Violations_Instruction
- Violations_FailToReport
- Violations_MoveWithoutPermission
- Delinquency_Reports
- Program_Attendances
- Program_UnexcusedAbsences
- Residence_Changes
- Avg_Days_per_DrugTest
- DrugTests_THC_Positive
- DrugTests_Cocaine_Positive
- DrugTests_Meth_Positive
- DrugTests_Other_Positive
- Percent_Days_Employed
- Jobs_Per_Year
- Employment_Exempt
- Prior_Conviction_Episodes_PPViolationCharges
- Prior_Conviction_Episodes_DomesticViolenceCharges
- Recidivism_Within_3years
- Recidivism_Arrest_Year1
- Recidivism_Arrest_Year2
- Recidivism_Arrest_Year3
- Training_Sample

Para a construção do modelo de *machine learning* foram retirados todas as colunas com informação em falta, a coluna do ID e as colunas "*Recidivism_Arrest_Year1, Recidivism_Arrest_Year2, Recidivism_Arrest_Year3 e Recidivism_Within_3year*". Esta ultima coluna será utilizada como *target*, ou seja, a característica que vai ser prevista pelo modelo.

Este *dataset* é composto por caraterísticas numéricas e caraterísticas categóricas, como por exemplo a raça pode ser "WHITE" para ofensores caucasianos ou "BLACK" para ofensores negros. No entanto estas caraterísticas categóricas não podem ser utilizadas para a construção do modelo e por isso foi feita um tratamento dos dados previamente. Este tratamentos foi feito com uso de "*LableEnconder*" que encontra todas estas caraterísticas e substitui por caraterísticas numéricas. Voltando ao exemplo da raça, a caraterística "WHITE" passaria a ser representada por "0" e a caraterística "BLACK" passaria a ser representada por "1", o mesmo foi feito para todas as outras caraterísticas categóricas.

5.2.2 *Compas Scores Dataset*

O segundo *dataset* é intitulado como "Compas Scores Dataset"[34]. Este *dataset* foi utilizado na investigação conduzida pela empresa *ProPublica* descrita no capitulo Estado da Arte [2] na secção "*How We Analyzed the COMPAS Recidivism Algorithm - ProPublica*" [2.3.1].
A versão original deste *dataset* é constituída por 47 colunas e 11757 linhas. Esta versão tem imensos campos em falta e por tanto para o desenvolvimento foi usada uma versão já pré-processada e com linhas e colunas, com dados em falta retiradas. Esta nova versão é constituída por 21 colunas e 6172 linhas [35]. As colunas incluem características como o género, idade, etnia e informação sobre detenções anteriores, muito em semelhança com o *dataset* descrito na secção anterior [5.2.1].

- age
- priors_count
- days_b_screening_arrest
- c_jail_time
- juv_fel_count
- juv_other_count
- juv_misd_count
- c_charge_degree:F
- c_charge_degree:M
- race:African-American
- race:Asian
- race:Caucasian
- race:Hispanic
- race:Native_American
- race:Other
- age_cat:25_–_45
- age_cat:Greater_than_45
- age_cat:Less_than_25
- sex:Female
- sex:Male
- is_recid

Este *dataset* tem como objetivo prever se certo ofensor é reincidente ou não, sendo o *target* a coluna ”is_recid”.
Foi também usada uma versão construida a partir do *dataset* original em que apenas foram retiradas as colunas onde foram encontradas campos com dados em falta e nenhuma linha foi retirada. Assim esta versão acabou com 13 colunas e as originais 11757 linhas.

- sex
- age
- race
- juv_fel_count
- juv_misd_count
- juv_other_count
- priors_count
- c_charge_degree
- r_charge_degree
- is_violent_recid
- v_type_of_assessment
- type_of_assessment
- decile_score

Esta versão do *dataset* tem como objetivo prever o nível de risco de reincidência de certo ofensor e não apenas se é reincidente ou não. Para este propósito foi usada a coluna ”decile_score” como *target* do modelo de *machine learning*

5.3 Analise de *Datasets*

5.3.1 COMPAS

Como já foi referido este *dataset* foi utilizado pela empresa Propublica na sua investigação da utilização da ferramenta COMPAS[4]. A estrutura deste do mesmo já foi apresentada na secção [5.2.2], na secção presente irá ser apresentada toda a analise de dados dos ofensores incluídos neste *dataset*. Vão ser analisadas características como a idade, género, etnia e o risco de reincidência com base na etnia de cada grupo presente nos dados.

Em primeiro lugar foi analisada a idade dos ofensores. Como pode ser verificado no seguinte gráfico de barras5.1, o intervalo de idades mais comum é o intervalo dos 25 a 45 anos com

6649 ofensores, de seguida o intervalo de idades superiores a 45 com 2668 ofensores e por fim 2440 para os ofensores com idade inferior a 25 anos.

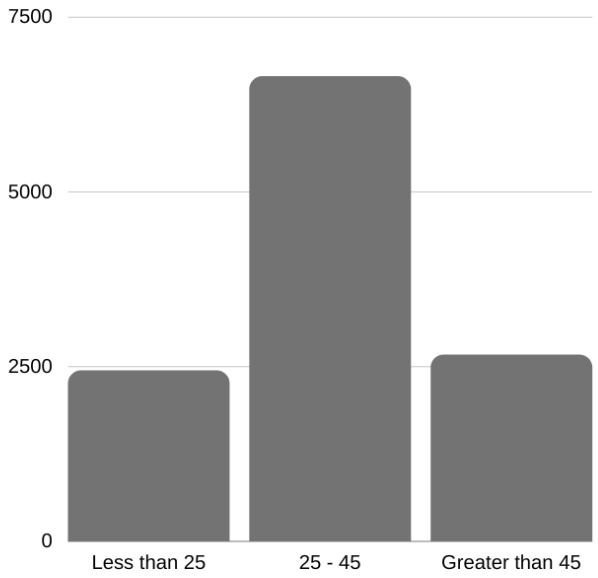

Figura 5.1: Numero de ofensores com base no intervalo de idades - COMPAS *dataset*

De seguida foi analisado o género dos ofensores presentes neste *dataset*. Esta analise é apresentada num gráfico circular com a percentagem de cada um dos géneros5.2. Como pode ser verificado nesse mesmo gráfico a população de ofensores masculinos, 79.4% é muito superior à população feminina 20.6%.

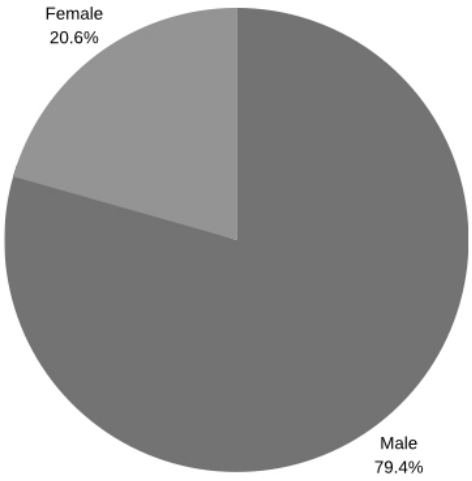

Figura 5.2: Numero de ofensores com base no género - COMPAS *dataset*

Foi também analisado o numero de ofensores com base nas etnias presente no *dataset*5.3. Estas etnias são divididas em 6 grupo, *African American, Asian, Caucasian, Hispanic, Native American* e *Other*.
Em primeiro lugar com 9336 ofensores está o grupo de *African Americans*, de seguida com 4085 ofensores *Caucasians*, 1100 para ofensores de etnia hispânica, 661 para ofensores inseridos na categoria *Other* e por fim *Asian* e *Native American* com 58 e 40 ofensores respetivamente.

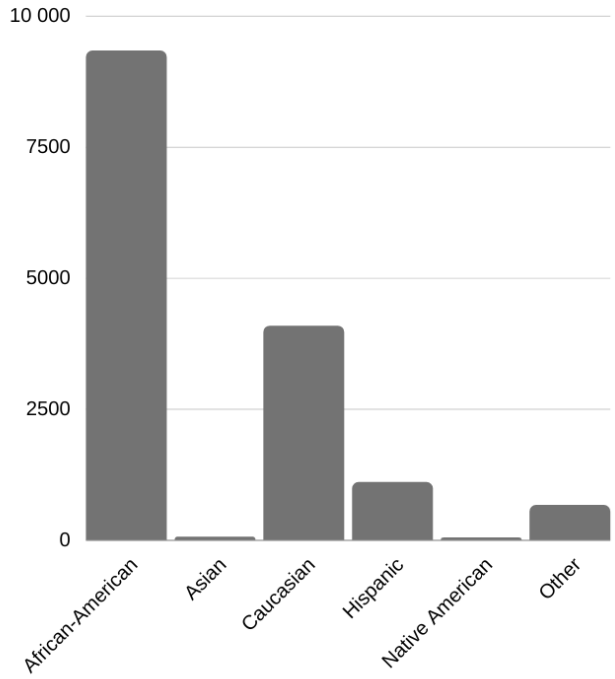

Figura 5.3: Numero de ofensores com base na etnia - COMPAS *dataset*

De seguida foi feita a analise de idades em conjunto com a étnica no gráfico5.4. No que toca a *African Americans* maior parte dos ofensores tem idade comprimida entre 25 e 45 anos com 3398 ofensores, seguido do intervalo de inferior a 25 com 1449 ofensores e por fim com idade superior a 45 com 966 ofensores. *Caucasians* em semelhança com os *African Americans* o maior numero de ofensores está entre os 25 e os 45 anos com 2196 ofensores, seguidos dos ofensores com idade superior a 45 com 1254 ofensores e por fim a idade inferior a 25 com 635 ofensores. Os ofensores do grupo *Hispanic* também em semelhança com os grupos anteriores estão presentes em maior numero no intervalo de idades de 25 a 45 anos com 636 ofensores, seguido de idades superiores a 45 com 256 e idades inferiores a 25 com 208. Ofensores classificados como *Other* seguem a mesma tendência e aparecem em maior numero entre os 25 e 45 anos com 366 ofensores, seguidos dos ofensores com idades superiores a 45 com 164 ofensores e idades inferiores a 25 com 131 ofensores. Por fim os grupos étnicos *Asian* e *Native American* tem números muito parecidos sendo que ambos tem uma população reduzida neste *dataset*. No intervalo 25 a 45 anos a população asiática tem 29 ofensores, e a população americana nativa é composta por 24, no intervalo superior a 45 estas etnias são compostas por 18 e 10 ofensores respetivamente e por fim os ofensores com

idade inferiores a 25 são 11 e 6 também respetivamente.

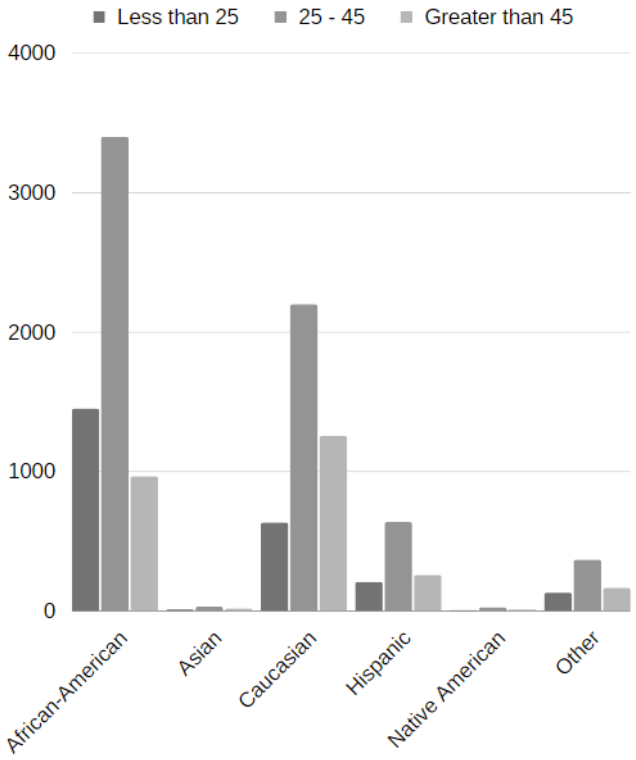

Figura 5.4: Numero de ofensores com base na etnia e idade - COMPAS *dataset*

Antes da analise de risco de reincidência foi feita também a analise de idade em conjunto com o género no gráfico 5.5. Nesta analise é possível verificar que o maior numero de ofensores, tanto masculinos com femininos, tem idades entre os 25 e 45 anos com 5230 ofensores masculinos e 1419 ofensores femininos. De seguida o maior numero de ofensores tem idade superior a 45 anos com 2154 ofensores masculinos e 514 ofensores femininos. Por fim com idade inferior a 25 anos estão inseridos no *dataset* 1952 ofensores masculinos e 488 ofensores femininos.

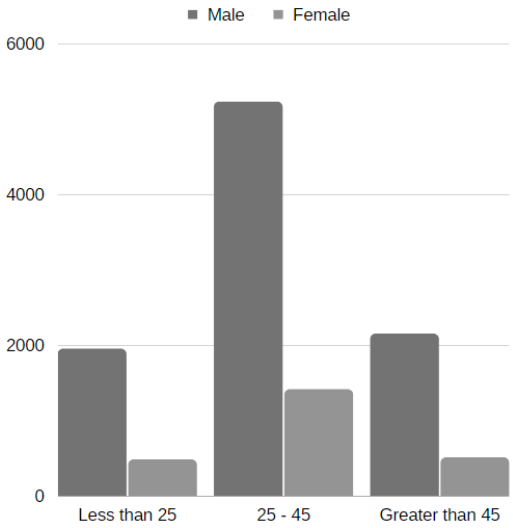

Figura 5.5: Numero de ofensores com base no género e idade - COMPAS *dataset*

Para a reincidência foi em primeiro lugar feita a analise em conjunto com género dos ofensores. Tanto para os ofensores de género masculino como para os de género feminino, as distribuições seguiram o mesmo padrão, ambas podem ser descritas como decrescentes em que estão concentradas em maior numero no nível 1, nível mais baixo de risco do que no nível 10, o nível mais alto. Em termos de valores para o nível 1 os ofensores masculinos chegam aos 2049 e feminino aos 528, para o nível máximo os valores chegam a 520 e 90, respetivamente, masculino e feminino.

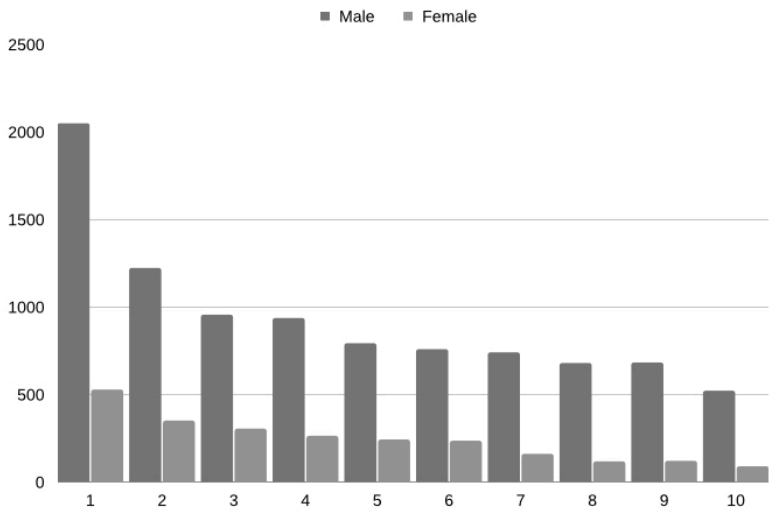

Figura 5.6: Distribuição de ofensores pelos níveis de risco com base no género - COMPAS *dataset*

A analise de risco de reincidência em conjunto com o grupo étnico, foi divida em 6 gráficos de barras cada um para uma etnia diferente presente no *dataset* em que cada barra corresponde a um nível de risco, de 1 a 10.

Em primeiro foi analisado a etnia *African American* na figura 5.7. Este grupo étnico é distribuído por todos os níveis de risco de uma forma uniforme, sendo que estão em maior numero no nível 1 com 694 ofensores do que no nível 10 com 470, ainda assim a diferença não é muito significativa.

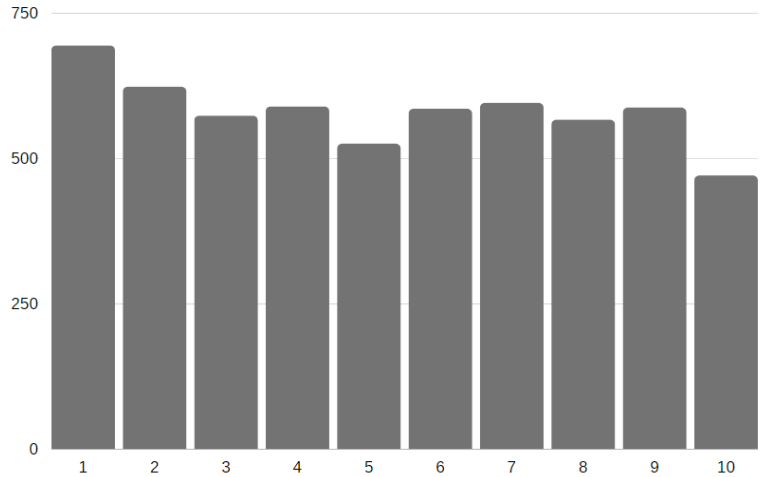

Figura 5.7: Distribuição de ofensores de etnia *African American* pelos níveis de risco - COMPAS *dataset*

No que toca a ofensores pertencentes ao grupo *Caucasians*, na figura5.8, estão distribuídos pelos níveis de risco de uma forma decrescente, significando que estão em maior numero no nível 1 e que o seu numero vai decrescendo à medida que o nível vai aumentando. No nível 1 estão presentes 1192 ofensores enquanto que no nível 10 estão presentes 91 ofensores.

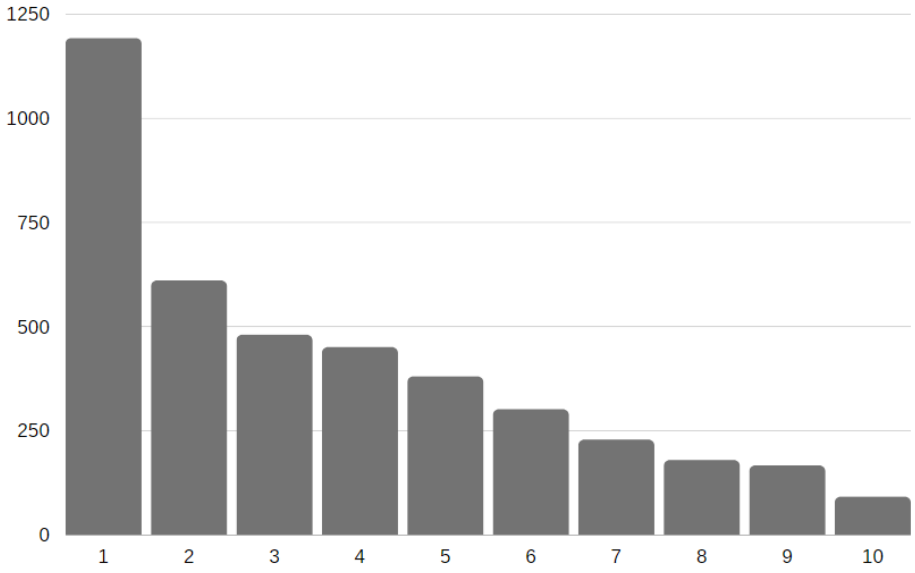

Figura 5.8: Distribuição de ofensores de etnia *Caucasian* pelos níveis de risco - COMPAS *dataset*

A figura5.9 apresenta a analise para a etnia asiática. Em semelhança com o grupo anterior este grupo está representado em maior numero no nível 1 de risco com 27 ofensores e em menor numero nos níveis 7 e 10 com apenas 1 ofensor em cada. É de referir que o numero total de asiáticos neste *dataset* é muito reduzido pelo que não é possível tirar conclusões definitivas sobre a relação entre a etnia e nível de risco.

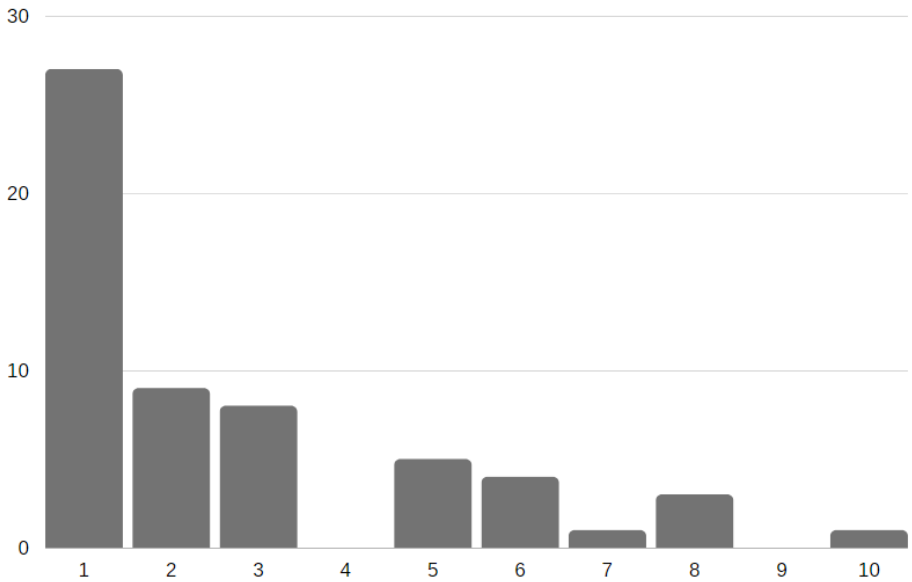

Figura 5.9: Distribuição de ofensores de etnia *Asian* pelos níveis de risco - COMPAS *dataset*

Na figura5.10, está representada a etnia *Hispanic*. Esta é muito parecida á *Caucasian* em

termos da distribuição pelos diferentes níveis de risco, apresenta também uma distribuição decrescente, sendo que no nível 1 estão presentes 376 ofensores e no nível 10 32 ofensores.

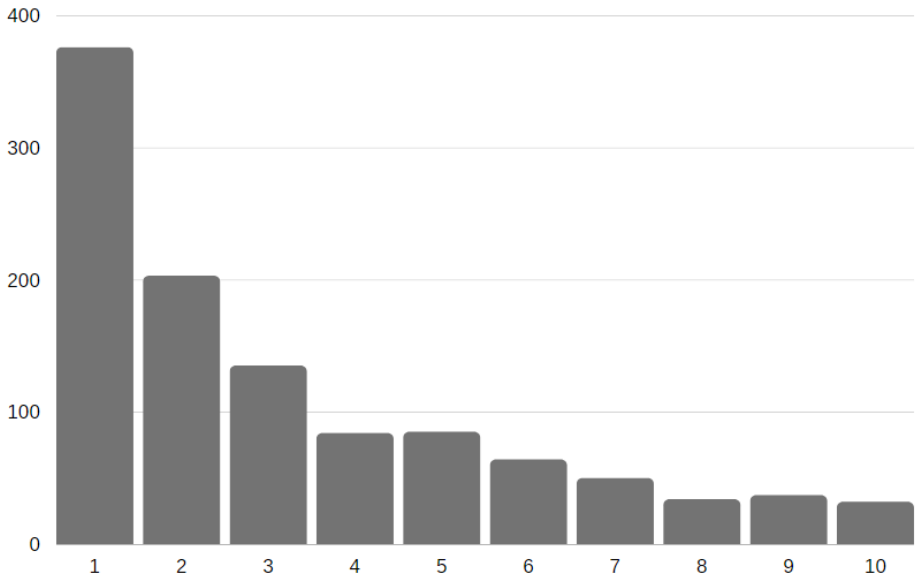

Figura 5.10: Distribuição de ofensores de etnia *Hispanic* pelos níveis de risco - COMPAS *dataset*

No que toca a *Native-Americans*, representados na figura5.11, a sua distribuição é um pouco irregular em que os níveis 1 e 2 são compostos por 8 ofensores e o nível 10 por 12 ofensores. Em semelhança com o grupo étnico *Asian* este grupo também sofre por ser reduzido em numero total de ofensores e por isso não é possível tirar conclusões no que toca à relação entre etnia e nível de risco.

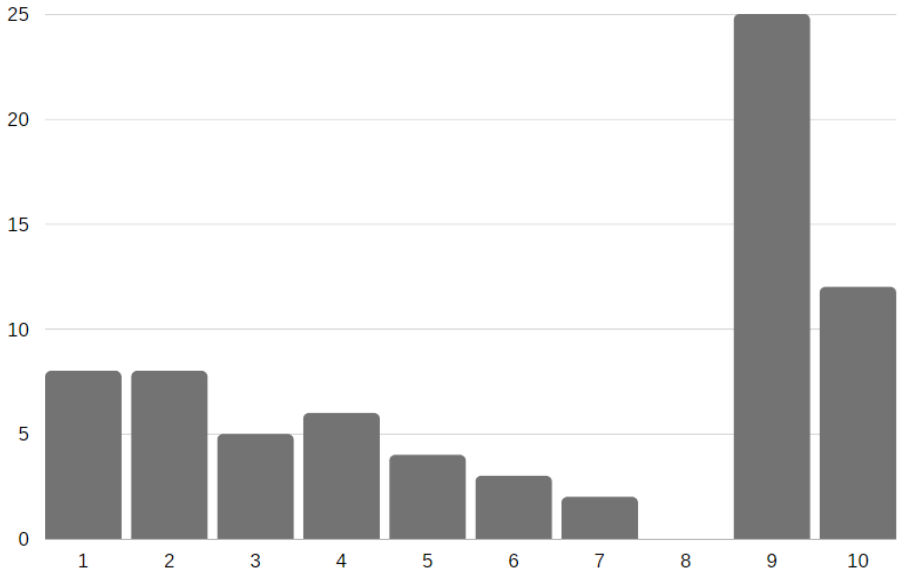

Figura 5.11: Distribuição de ofensores de etnia *Native-American* pelos níveis de risco - COMPAS *dataset*

Por fim, a analise do grupo *Others*, este grupo por não representar uma etnia por si torna também difícil retirar uma conclusão a partir destes dados, ainda assim este grupo segue uma distribuição também decrescente de certa forma em que o nível 1 é composto por 284

ofensores e o nível 10 por 12 ofensores.

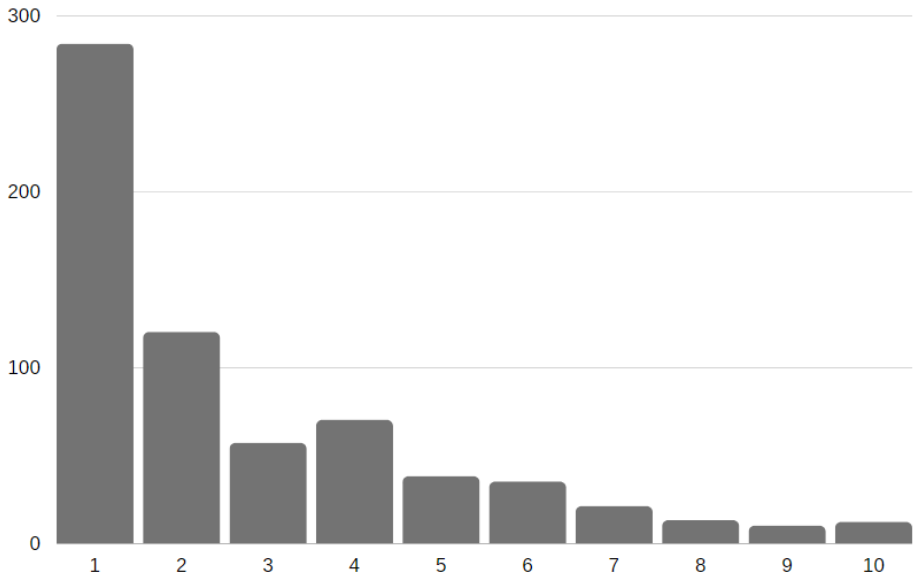

Figura 5.12: Distribuição de ofensores do grupo *Other* pelos níveis de risco - COMPAS *dataset*

Desta forma é terminada a analise de dados do *dataset COMPAS*. Todas as conclusões retiradas vão ser apresentadas no fim do documento.

5.3.2 NIJ Recidivism Challenge

Nesta secção irá ser feita uma analise semelhante à analise feita ao *dataset* do COMPAS utilizado. Para esta analise vão ser utilizadas características como a idade, o género e a etnia bem como o conjunto de idade e género, idade e etnia e por fim a reincidência com a idade e etnia. Este *dataset* difere do *dataset* do COMPAS principalmente na forma como prevê a reincidência, enquanto que o *dataset* do COMPAS divide os ofensores em níveis de risco, este apenas classifica os ofensores como reincidentes ou não reincidentes.

Em primeiro lugar foi feita a analise no que toca ao género de cada ofensor. Para isto foi criado o gráfico circular seguinte5.13. Em semelhança com o primeiro *dataset* analisado, a população é maioritariamente masculina, com 22668 ofensores do sexo masculino e apenas 3167 do sexo feminino.

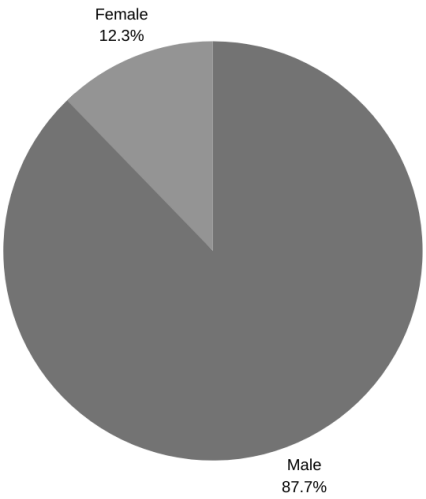

Figura 5.13: Distribuição de ofensores pelo seu género - NIJ *dataset*

Em segundo lugar foi analisado o *dataset* tendo em conta a etnia de cada ofensor, neste apenas são apresentadas duas etnias, *White* e *Black*. Esta analise é também apresentada com o uso de um gráfico circular 5.14. Neste gráfico é possível verificar que a diferença entre etnia não é tão significativa como a diferença entre géneros. Neste *dataset* há uma maior quantidade de ofensores classificados como *Black*, 14847 do que ofensores classificados como *White*, 10988.

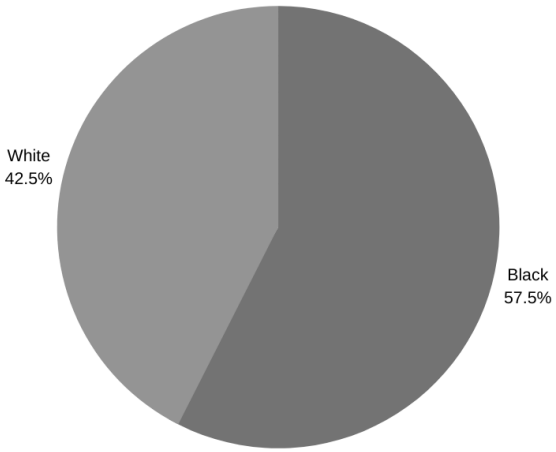

Figura 5.14: Distribuição de ofensores pela sua etnia - NIJ *dataset*

Foi também feita uma analise das idades dos ofensores presentes neste *dataset*, representada na figura 5.15. Neste conjunto de dados os ofensores são divididos em 7 intervalos de idade diferentes, 18-22, 23-27, 28-32, 33-37, 38-42, 43-47 e 48+. No primeiro intervalo, com idades entre os 18 e 22 anos estão inseridos 2066 ofensores, dos 23 aos 27 anos 5176

ofensores, dos 28 aos 32 4982 ofensores, dos 33 aos 37 4271 ofensores, dos 38 aos 42 2993 ofensores, dos 43 aos 47 2620 ofensores e finalmente com idades superiores a 48 existem 3727.

É possível verificar que maior parte dos ofensores tem idades contidas entre os 23 e 37 anos e os intervalos de 18 a 22 anos e 43 a 47 anos tem o menor numero de ofensores.

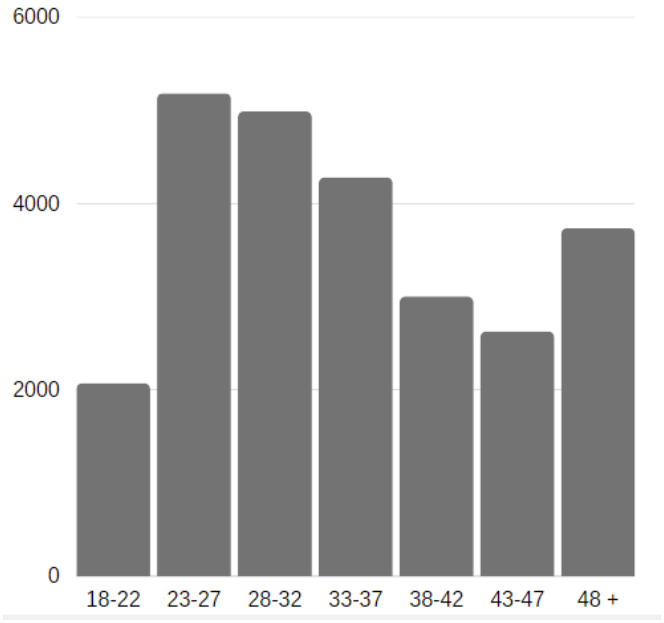

Figura 5.15: Distribuição de ofensores pela sua idade - NIJ *dataset*

De seguida foi feita a analise da distribuição de idades em conjunto com o género de cada ofensor 5.16. Podemos retirar do seguinte gráfico de barras que no intervalo de idades de 18 a 22 anos é onde existe menor concentração de ofensores tanto masculinos, 1938 como femininos 128. No intervalo de 23 a 27 anos estão inseridos 4656 ofensores masculinos e 520 ofensores femininos, este intervalo é o máximo para os ofensores masculinos. No intervalo de 28 a 32 anos existem 4376 ofensores masculinos e 606 ofensores femininos, sendo este o valor máximo para os ofensores femininos. No intervalo de 33 a 37 anos estão representados 3693 ofensores masculinos e 578 ofensores femininos. No intervalo de 38 a 42 anos existem 2545 ofensores masculinos e 448 ofensores femininos. Dos 43 as 47 anos estão inseridos 2204 ofensores masculinos e 416 ofensores femininos. Por fim com idades superiores a 48 anos estão neste *dataset* 3256 ofensores masculinos e 471 ofensores femininos.

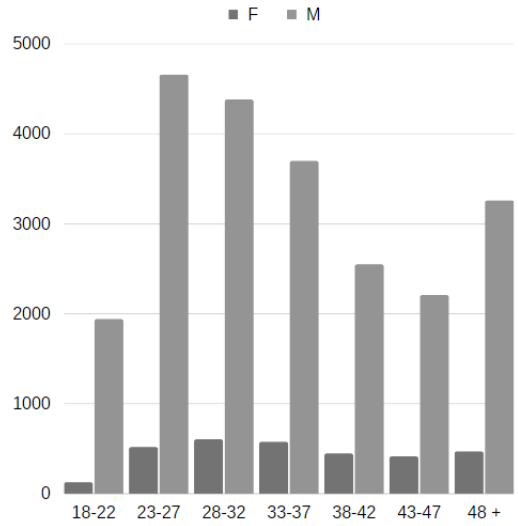

Figura 5.16: Distribuição de ofensores pela sua idade e género - NIJ *dataset*

Antes da analise da reincidência foi ainda analisadas as idades dos ofensores em conjunto com a sua etnia 5.17. Do seguinte gráfico podemos retirar que ofensores classificados com a etnia *White* no intervalo de 18 a 22 estão em menor numero do que noutros intervalos de idades com apenas 600 ofensores, enquanto que os ofensores *Black* são 1466. No intervalo de 23 a 27 anos é encontrado o máximo de ofensores *Black* com 3388 ofensores, enquanto que ofensores de etnia *White* são 1788. No intervalo de 28 a 32 anos é encontrado o numero máximo de ofensores *White* com 2103 ofensores, e 2879 ofensores de etnia *Black*. No intervalo 33 a 37 anos estão inseridos 1910 ofensores de etnia *White* e 2361 de etnia *Black*. O intervalo de 38 a 42 anos é composto por 1479 ofensores de etnia *White* e 1514 de etnia *Black*. No intervalo de 43 a 47 anos é encontrado valor mínimo para ofensores de etnia *Black* com 1275 ofensores e 1345 ofensores de etnia *White*. Por fim no que toca a ofensores com idades superiores a 48 anos existem 1763 ofensores de etnia *White* e 1964 de etnia *Black*.

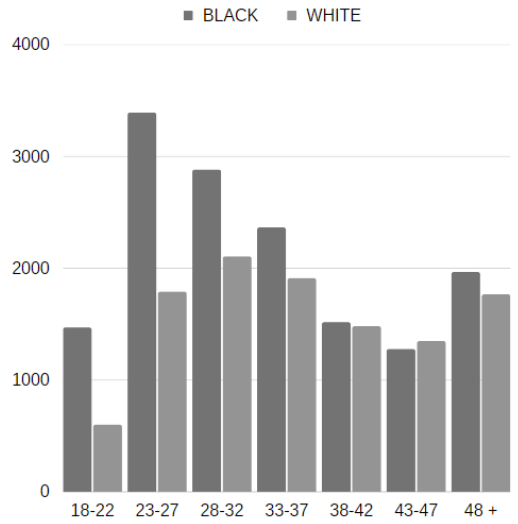

Figura 5.17: Distribuição de ofensores pela sua idade e etnia - NIJ *dataset*

No que toca a reincidência foi em primeiro lugar analisada em conjunto com o género de

cada ofensor 5.18. No seguinte gráfico pode se verificar que quando se trata de ofensores de género masculino são em maior numero classificados como reincidentes, *True*, 13462 ofensores do que classificados como não reincidentes, *False*, 9206 ofensores. Quando se trata de ofensores de género feminino, as classificações como reincidente e como não reincidente tem um numero bastante parecido, 1725 ofensores para reincidentes e 1442 ofensores para não reincidentes.

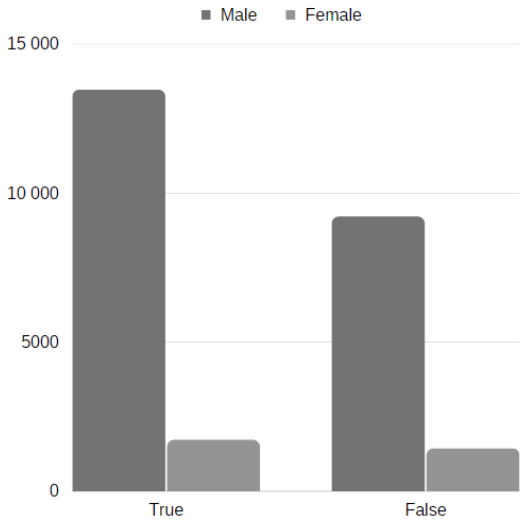

Figura 5.18: Distribuição da reincidência de ofensores com base no seu género - NIJ *dataset*

Por fim, a analise da reincidência com base na etnia de cada ofensor 5.20. Neste gráfico podemos retirar conclusões que ofensores de etnia *Black* são classificados como reincidentes e não reincidentes em maior numero, 8713 e 6134 respetivamente. Os ofensores de etnia *White* são classificados como reincidentes 6191 ofensores e 4797 como não reincidentes. Estas diferenças podem ser explicadas por os ofensores de etnia *Black* aparecerem em maior numero neste *dataset*, a diferença entre as duas etnias no que toca à não reincidência não é muito significativa. Já quando se trata da reincidência os valores para as duas etnias já tem uma diferença significativa.

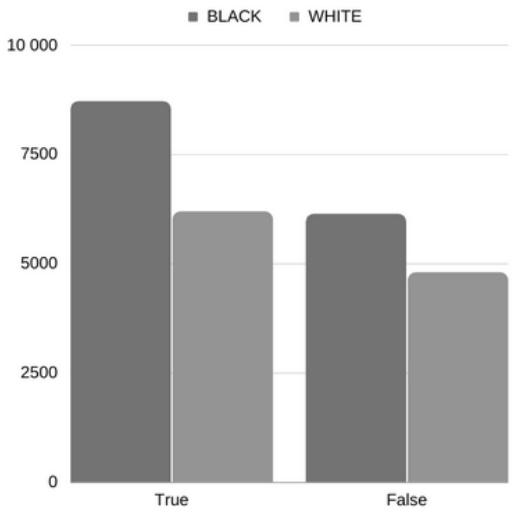

Figura 5.19: Distribuição da reincidência de ofensores com base na sua etnia - NIJ *dataset*

5.4 Métodos de *Machine Learning*

Para que os *dataset* possam ser usados para a construção de um modelo têm que em primeiro lugar pré processados e adaptados para a previsão pretendida. As variáveis de entrada, *features* para o *dataset* NIJ foram as seguintes 35 colunas:

- Gender
- Race
- Age_at_Release
- Residence_PUMA
- Education_Level
- Dependents
- Prison_Years
- Prior_Arrest_Episodes_Felony
- Prior_Arrest_Episodes_Misd
- Prior_Arrest_Episodes_Violent
- Prior_Arrest_Episodes_Property
- Prior_Arrest_Episodes_Drug
- Prior_Arrest_Episodes_PPViolationCharges
- Prior_Arrest_Episodes_DVCharges
- Prior_Arrest_Episodes_GunCharges
- Prior_Conviction_Episodes_Felony
- Prior_Conviction_Episodes_Misd
- Prior_Conviction_Episodes_Viol
- Prior_Conviction_Episodes_Prop
- Prior_Conviction_Episodes_Drug
- Prior_Conviction_Episodes_GunCharges
- Prior_Revocations_Parole
- Prior_Revocations_Probation
- Condition_MH_SA
- Condition_Cog_Ed
- Condition_Other
- Violations_ElectronicMonitoring
- Violations_Instruction
- Violations_FailToReport
- Violations_MoveWithoutPermission
- Delinquency_Reports
- Program_Attendances
- Program_UnexcusedAbsences
- Residence_Changes
- Employment_Exempt

Para o *dataset* COMPAS foi também efetuada uma tarefa de pré processamento, ainda que tenha sido eliminada um coluna, *is_recid*, sendo esta característica a variável de saída do modelo. As variáveis de entrada, *features* para o *dataset* COMPAS foram as seguintes 20 colunas:

- age
- priors_count
- days_b_screening_arrest
- c_jail_time

- juv_fel_count
- juv_other_count
- juv_misd_count
- c_charge_degree:F
- c_charge_degree:M
- race:African-American
- race:Asian
- race:Caucasian
- race:Hispanic
- race:Native_American
- race:Other
- age_cat:25_–_45
- age_cat:Greater_than_45
- age_cat:Less_than_25
- sex:Female
- sex:Male

5.4.1 AdaBoost Classifier

O algoritmo de *machine learning* AdaBoost[36][5] ou *Adaptive Boosting*, é uma técnica de *boosting* usada em *machine learning*. Este *boosting* funciona da seguinte forma, cria n árvores de decisão durante o período de treino. Quando a primeira árvore/modelo é criada, o registo classificado incorretamente no primeiro modelo é lhe dado prioridade. Só estes registos são enviados para o segundo modelo. Este processo continua até ser atingida uma condição especificada. Quando o primeiro modelo é criado o algoritmo "aponta" o erros deste primeiro modelo. Este registo é depois usado como *input* para o próximo modelo. Os modelos 1,2,3,...,N são individuais que podem ser chamados de árvores de decisão. Todos os métodos de *boosting* funcionam pelo mesmo principio.

No que toca ao *AdaBoost* o seu funcionamento é parecido. O estimador mais comum de ser usado com o *adaboost* são árvores de decisão com apenas um nível, o que significa que são compostas por apenas um nodo e duas folhas. Estas árvores são também conhecidas por *Decision Stumps*5.23.
O algoritmo funciona da seguinte forma, é em primeiro lugar criado um modelo com com todos os pontos com o mesmo peso. São depois atribuídos pesos superiores aos pontos que foram classificados de uma forma errada e assim sucessivamente.

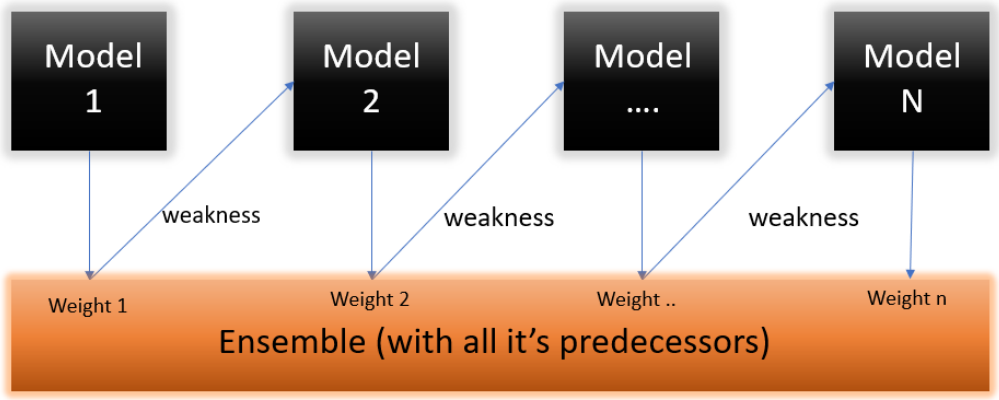

Figura 5.20: Estrutura do principio de *boosting* [5]

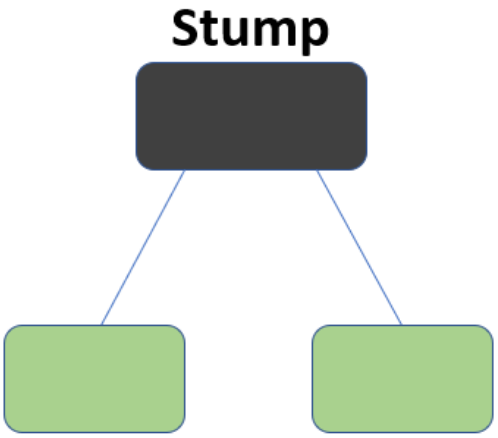

Figura 5.21: Estrutura de um *Decision Stump* [5]

5.4.1.1 Resultados - *AdaBoost Classifier*

A secção de resultados começa com a apresentação dos resultados das métrica de avaliação, *Score*, *Precision*, *Recall* e *F1 Score*, relativos ao modelo construido com o uso do método *Adaboost Classifier*, de cada um dos *datasets*. Em primeiro lugar vão ser apresentados os resultados relativos ao *dataset* COMPAS. Para a métrica *Score* o modelo obteve um resultado de 0.687449, o resultado para a métrica *Recall* o modelo foi classificado com 0.687796 por fim as métricas de *Precision* e *F1 Score* tiveram um resultado de 0.687449 e 0.686899 respetivamente.

No que toca ao *dataset NIJ Recidivism Challenge* os resultados foram, para a métrica *Score* o modelo obteve um resultado de 0.676602, na métrica *Precision* o modelo foi avaliado com 0.675006, da métrica de *Recall* resultou um valor de 0.676601 e por fim para a métrica *F1 Score* foi obtido um resultado de 0.666247.

Todos estes valores podem ser consultado na seguinte tabela 5.1 e gráfico de barras 5.22:

Tabela 5.1: Resultados obtidos com o uso de *Adaboost Classifier*

	Score	**Precision**	**Recall**	**F1 Score**
COMPAS	0.687449	0.687796	0.687449	0.686899
NIJ	0.676602	0.675006	0.676602	0.666247

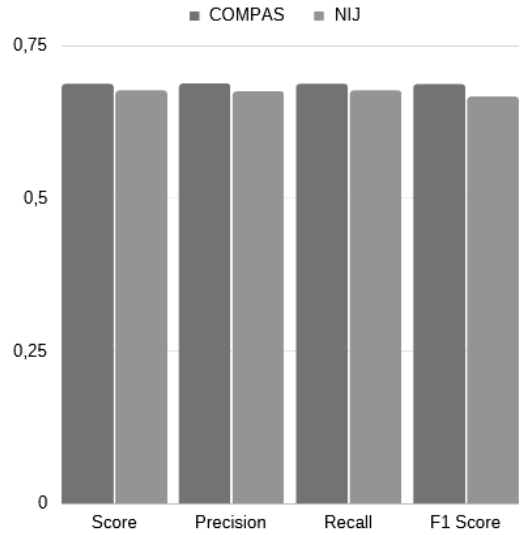

Figura 5.22: Resultados das métricas para o método *Adaboost Classifier*.

Quando é feita uma avaliação a um modelo de *machine learning* é também importante verificar quais as variáveis de entrada, *features* contribuíram mais para previsão.

As *Features* mais importantes na construção do modelo com o uso de *Adaboost Classifier* foram:

- *Dataset* COMPAS
  - *age*
  - *priors_count*
  - *c_jail_time*
- *NIJ Recidivism Challenge*
  - *Age_at_Release*
  - *Prior_Arrest_Episodes_Felony*
  - *Delinquency_Reports*

5.4.2 Logistic Regression

Este tipo de modelo estatístico, também conhecido por *logit model*, é muito usado para classificações e analises preditivas. *Logistic Regression*[37] estima a probabilidade de um evento ocorrer com base num *dataset* com variáveis independentes. Numa *Logistic Regresssion*, é aplicada uma transformação *logit* na probabilidade de sucesso a dividir pela probabilidade

de insucesso. Esta função é representada pelas seguinte formulas:

Logit(pi) = 1/(1+ exp(-pi))

ln(pi/(1-pi)) = Beta_0 + Beta_1*X_1 + ... + B_k*K_k

Nestas formulas o *logit(pi)* é a variável dependente ou de resposta e a variável X é a variável dependente. O parâmetro *Beta* é comum ser estimado por uma MLE (*Maximum Likelihood Estimation*). Este método testa valores diferentes para o *Beta* através de múltiplas iterações para otimizar o *Fit* das probabilidades. Todas estas iterações produzem a função *log likelihood* e a regressão procura maximizar esta função para encontrar a melhor estimação. Assim que o coeficiente ótimo é encontrado, as probabilidades condicionais para cada observação podem ser calculadas, registadas e somadas para em conjunto produzir a probabilidade preditiva. Para um classificação binária uma probabilidade menor que 0.5 vai prever 0 enquanto que uma probabilidade maior que 0.5 vai prever 1.

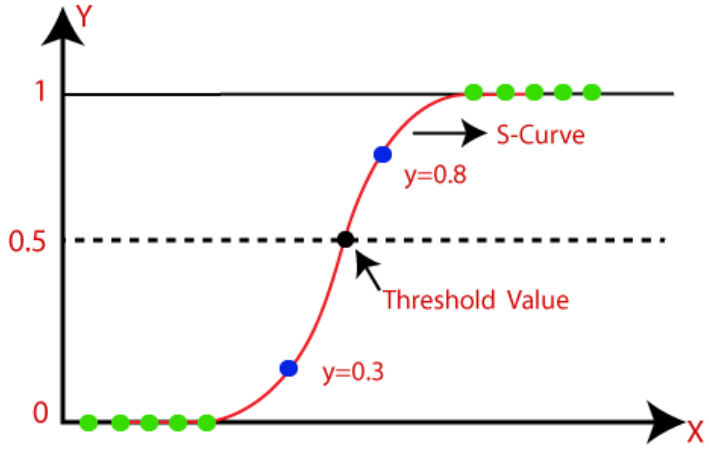

Figura 5.23: Estrutura do método *logistic regression* [6]

5.4.2.1 Resultados *Logistic Regression*

A secção de resultados começa com a apresentação dos resultados das métrica de avaliação, *Score*, *Precision*, *Recall* e *F1 Score*, relativos ao modelo construido com o uso do método *Logistic Regression*, de cada um dos *datasets*. Em primeiro lugar vão ser apresentados os resultados relativos ao *dataset* COMPAS. Para a métrica *Score* o modelo obteve um resultado de 0.686640, o resultado para a métrica *Recall* o modelo foi classificado com 0.694523 por fim as métricas de *Precision* e *F1 Score* tiveram um resultado de 0.686640 e 0.685750 respetivamente.

No que toca ao *dataset NIJ Recidivism Challenge* os resultados foram, para a métrica *Score* o modelo obteve um resultado de 0.667118, na métrica *Precision* o modelo foi avaliado com 0.663971, da métrica de *Recall* resultou um valor de 0.667118 e por fim para a métrica *F1 Score* foi obtido um resultado de 0.654944.

Todos estes valores podem ser consultado na tabela 5.2 e no gráfico 5.24:

Tabela 5.2: Resultados obtidos com o uso de *Logistic Regression*

	Score	**Precision**	**Recall**	**F1 Score**
COMPAS	0.686640	0.694523	0.686640	0.685750
NIJ	0.667118	0.663971	0.667118	0.654944

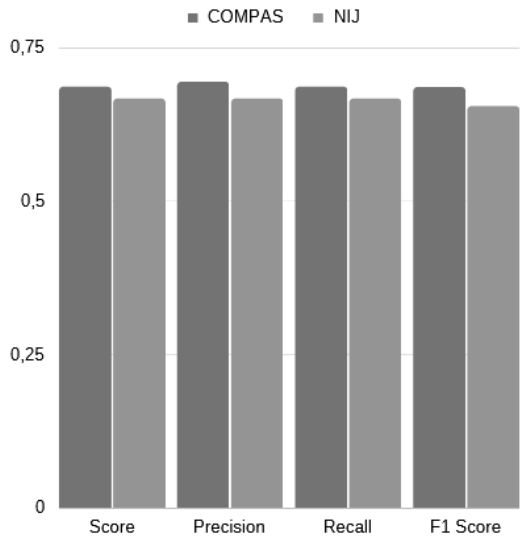

Figura 5.24: Resultados das métricas para o método *Logistic Regression*.

5.4.3 Random Forest Classifier

Uma árvore de decisão[7] é um algoritmo de aprendizagem supervisionada que tanto pode ser utilizado para tarefas de classificação e de regressão. Tem uma estrutura hierárquica de árvore que consiste num nodo raiz, *root node*, ramos, nodos internos e nodos folha.
Como mostra a figura a cima 5.26 uma árvore de decisão é iniciada por um nodo de raiz *root node*, que não recebe ramos. Os ramos saídos do *root node* conectam a nodos internos *internal nodes*, também conhecidos por *decision nodes*. O nodos folha, *leaf nodes* ou *terminal nodes* representam todos os possíveis resultados.
Uma *Random Forest*[8], tal como nome indica, consiste num numero grande de árvores de decisão individuais que operam em conjunto. Cada árvore individual na *Random Forest* dá como *output* uma classe de predição e a classe com maior numero de votos é escolhida como a predição do modelo.
O conceito fundamental por detrás de uma *Random Forest* é simples mas eficaz, a sabedoria de multidões, *the wisdom of crowds*. Um grande numero de modelos não correlacionados que operam como um comité vão ter um desempenho muito superior a qualquer modelo individual.
A baixa correlação entre modelos é a chave para este método. Modelos não relacionados podem produzir previsões em conjunto que são mais precisas do que as previsões individuais.

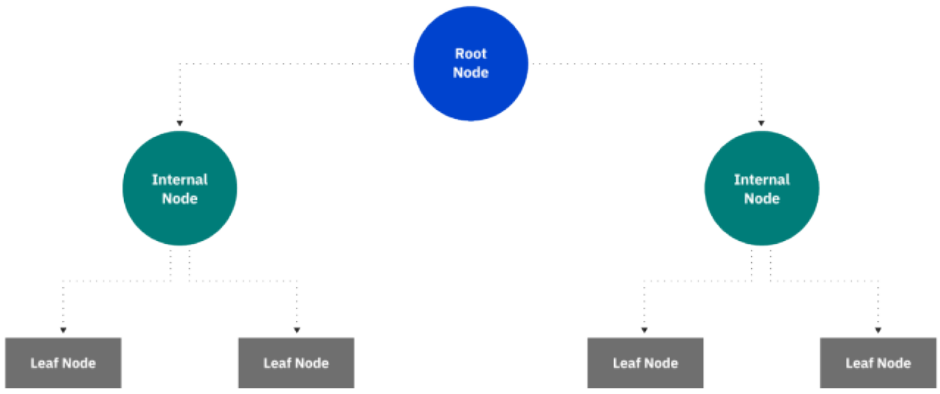

Figura 5.25: Estrutura de uma *Decision Tree* [7]

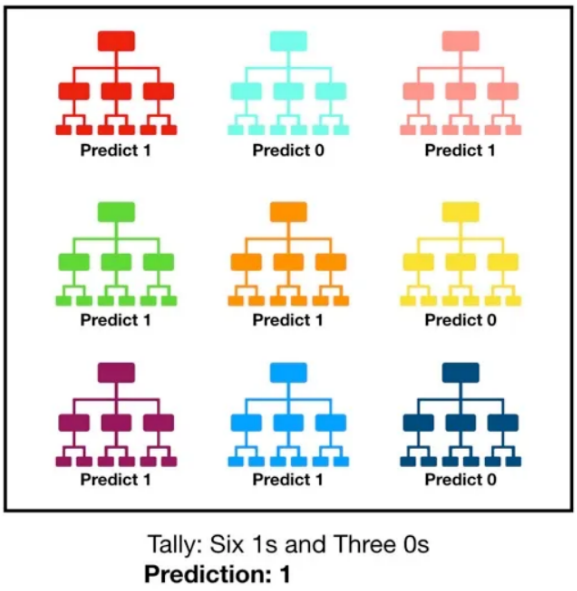

Figura 5.26: Estrutura de uma *Random Forest*[8]

As árvores pertencentes a uma *Random Forest* protegem se de erros individuais, enquanto que algumas árvores podem estar erradas, muitas outras vão estar corretas, e assim em grupo a *Random Forest* vai incidir na direção correta.

5.4.3.1 Resultados *Random Forest Classifier*

A secção de resultados começa com a apresentação dos resultados das métrica de avaliação, *Score*, *Precision*, *Recall* e *F1 Score*, relativos ao modelo construido com o uso do método *Random Forest Classifier*, de cada um dos *datasets*. Em primeiro lugar vão ser apresentados os resultados relativos ao *dataset* COMPAS. Para a métrica *Score* o modelo obteve um resultado de 0.616194, o resultado para a métrica *Recall* o modelo foi classificado com 0.615811 por fim as métricas de *Precision* e *F1 Score* tiveram um resultado de 0.616194 e 0.615872 respetivamente.

No que toca ao *dataset NIJ Recidivism Challenge* os resultados foram, para a métrica *Score* o modelo obteve um resultado de 0.676795, na métrica *Precision* o modelo foi avaliado com 0.673745, da métrica de *Recall* resultou um valor de 0.676795 e por fim para a métrica *F1 Score* foi obtido um resultado de 0.666899.

Todos estes valores podem ser consultado na tabela 5.3 e na figura 5.27:

Tabela 5.3: Resultados obtidos com o uso de *Random Forest Classifier*

	Score	**Precision**	**Recall**	**F1 Score**
COMPAS	0.616194	0.615811	0.616194	0.615872
NIJ	0.676795	0.673745	0.676795	0.666899

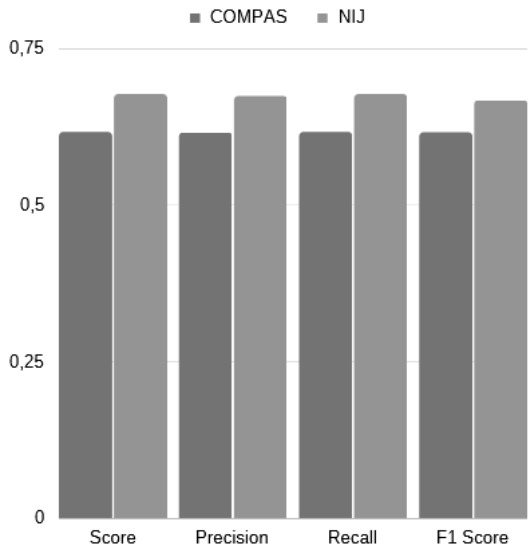

Figura 5.27: Resultados das métricas para o método *Random Forest Classifier*.

Para o modelo de *Random Forest Classifier* foi também feita a analise de quais variáveis de entrada, *features* foram mais importantes para a construção do modelo. No que toca ao *dataset* do COMPAS, as *features* mais importantes foram as mesmas que no modelo desenvolvido com *AdaBoosst*. Já para o *dataset* NIJ apenas a primeira *feature*, *Age_at_Release* foi comum ao algoritmo anterior.

As *Features* mais importantes na construção do modelo com o uso de *Random Forest Classifier* foram:

- *Dataset* COMPAS
  - *age*
  - *priors_count*
  - *c_jail_time*
- *NIJ Recidivism Challange*

- *Age_at_Release*
- *Residence_PUMA*
- *Prior_Arrest_Episodes_PPViolationCharges*

5.5 Conclusão

Neste capitulo de Desenvolvimento, foram em primeiro lugar apresentados os dois *datasets* usados, o *NIJ Recidivism Challenge Dataset* e o COMPAS *dataset*, com uma breve descrição e a lista de todos as colunas que constituem o *dataset*. Foi depois demonstrada a analise de dados de ambos os *datasets*. Esta analise foi mais focada nos atributos de idade, género e etnia dos ofensores. Foram depois apresentadas todas as variáveis de entrada para cada um dos *datasets*, e cada um dos métodos de *machine learning* usados, *Adaboost Classifier*, *Logistic Regression* e *Random Forest Classifier*. Cada uma das secções dedicadas aos métodos utilizados apresentou também todos os resultados obtidos nas métricas *Score, Precision, Recall* e *F1 Score* e as variáveis de entrada mais importantes para cada um. Para o *dataset* COMPAS as variáveis mais importantes coincidiram nos métodos em que estas foram analisadas, no *dataset* NIJ apenas a primeira variável foi a mesma nos modelos construidos.

Capítulo 6

Discussão de Resultados e Conclusões Finais

6.1 Introdução

Neste capitulo intitulado de Discussão de Resultados e Conclusões Finais vão ser, em primeiro lugar apresentados todos os resultados obtidos, quer na analise dos *datasets* quer com os métodos de *machine learning*. A partir destes resultados vai ser iniciada um discussão sobre os mesmos. Posteriormente a esta discussão vão apresentadas as conclusões finais retiradas desta analise de resultados. Por fim vai ser apresentado uma secção de trabalho adicional que foi desenvolvido em paralelo com a dissertação num projeto relacionado com o tema.

6.2 Discussão de Resultados

Esta secção de discussão de resultados vai ter a seguinte estrutura. Em primeiro lugar vão ser analisados os resultados referentes ao estudo dos *datasets* escolhidos comparando os resultados de ambos. Posteriormente irá ser apresentada a analise de resultados obtidos com os métodos de *machine learning*.

A analise de resultados provenientes dos *datasets* vai ser inicializada com a característica de género. A população do *dataset* do COMPAS, como mostra a figura 5.2, é em maioria composta por ofensores do género masculino, 79.4%, enquanto que ofensores do género feminino são apenas 20.6%. Para o *dataset* NIJ também é possível verificar a mesma situação, a população masculina é bastante superior que a população feminina, 87.7% e 12.3% 5.13.

Quando se trata de idade dos ofensores ambos os *datasets* têm, novamente, valores semelhantes. O COMPAS divide a idade em apenas 3 categorias, menos de 25 anos, de 25 a 45 anos e mais de 45 anos, onde a maior concentração de ofensores está localizada na categoria intermédia, 25-45 5.1. O *dataset* NIJ divide a idade dos ofensores em 7 categorias diferentes mas a maior concentração de ofensores está localizada entre os 23 e 37 anos 5.15.

Passado para a analise dos resultados referente à etnia dos ofensores, o *dataset* COMPAS divide a população em 6 categorias de etnia, *African American, Asian, Caucasian, Hispanic, Native American* e *Other*. Já o *dataset* NIJ divide apenas em 2 categorias, *White* e *Black*, por essa razão apenas estas duas categorias vão ser comparadas. No *dataset* COMPAS a população de *African Americans* é bastante superior à população de ofensores *Caucasian*, 9336 e 4085 respetivamente5.3. Para o *dataset* NIJ a população classificada como *Black* é também superior ainda que por uma fração menor, 14847 (57.5%) e 10988 (42.5%) 5.14.

De seguida vão ser analisados os resultados da reincidência em conjunto com o género do ofensor. Os *datasets* escolhidos classificam a reincidência de formas um pouco diferentes, enquanto que o *dataset* COMPAS divide os ofensores por níveis de risco, *dataset* NIJ apenas classifica os ofensores com reincidentes ou não reincidentes. A distribuição masculina pelos níveis de reincidência no COMPAS pode ser descrita como decrescente, existem mais ofensores no nível 1 de risco do que no seguinte e assim sucessivamente. Para a população feminina a distribuição é novamente decrescente, ainda que em valores menores por a população feminina ser bastante inferior à masculina 5.6. Para o *dataset* NIJ os ofensores masculinos são classificados em maior numero como reincidentes do que não reincidentes, 13468 e 9206 respetivamente. Enquanto que a população feminina tem valores mais semelhantes de reincidência e não incidência, 1725 e 1442 respetivamente 5.18.

Por fim a analise de reincidência em conjunto com a etnia do ofensor. Como foi referido anteriormente o *dataset* NIJ apenas é composto por ofensores classificados com *White* e *Black*, por essa razão no *dataset* COMPAS apenas vão ser analisadas estas duas etnias. No COMPAS para os ofensores classificados como *African American* a distribuição pelos 10 níveis de risco de reincidências segue um padrão muito uniforme, em que todos os níveis são compostos por aproximadamente o mesmo numero de ofensores 5.7. Para os ofensores classificados como *Caucasian* a distribuição segue um padrão decrescente em que o nível 1, nível mais baixo, é composto por maior parte dos ofensores e este numero decresce até ao nível 10 onde se encontra a menor concentração de de ofensores 5.8. No que toca ao *dataset* NIJ os ofensores de etnia *Black* são também classificados em maior numero como reincidentes, mas são também classificados em maior numero como não reincidentes quando comparados com ofensores de etnia *White* 5.20. Este resultado vai de encontro ao resultado obtido no *dataset* COMPAS, em que os ofensores de etnia negra, afro-americana são classificados em maior numero tanto num nível de risco baixo como num nível de risco alto.

Para terminar a discussão de resultados vão ser apresentados os dados obtidos com os métodos de *machine learning Adaboost Classifier*, *Logistic Regression* e *Random Forest Classifier*. Em primeiro lugar vão ser analisados e comparados os valores obtidos em cada uma das métricas usadas, para cada um dos *datasets*.

Na tabela seguinte 6.1 e nas figuras 6.1 e 6.2 para o *dataset* COMPAS e NIJ respetivamente, é possível consultar todos os resultados obtidos com as métricas escolhidas para cada um dos métodos usados em ambos *datasets*.

A partir da analise da tabela anterior é possível verificar que os resultados para o *dataset* COMPAS foram melhores, ainda por pouca margem, que os resultados do *dataset* NIJ, exceto para o método *Random Forest Classifier*. Em relação ao *dataset* COMPAS, o método que obteve melhor *Score* foi o *Adaboost Classifier*, este método também obteve melhores valores para *Recall* e *F1 Score*. Para a métrica *Precision* o método *Logistic Regression* foi o que obteve o melhor valor.

Tabela 6.1: Tabela representativa todos os resultados obtidos com métodos de *machine learning*

	COMPAS				NIJ			
	Score	**Precision**	**Recall**	**F1 Score**	**Score**	**Precision**	**Recall**	**F1 Score**
Adaboost Classifier	0.687449	0.687796	0.687449	0.686899	0.676602	0.675006	0.676602	0.666247
Logistic Regression	0.686640	0.694523	0.686640	0.685750	0.667118	0.663971	0.667118	0.654944
Random Forest Classifier	0.616194	0.615811	0.616194	0.615872	0.676798	0.673745	0.676795	0.666899

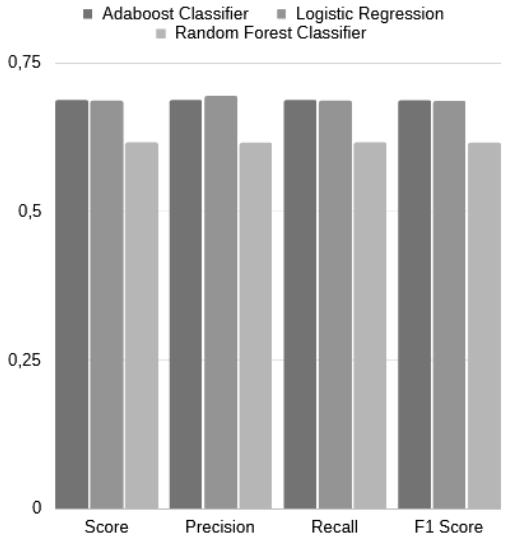

Figura 6.1: Resultados das métricas para o *dataset* COMPAS.

No que toca ao *dataset* NIJ o método *Adaboost Classifier* e o *Random Forest Classifier* obtiveram valores parecidos sendo estes os melhores métodos para o *dataset* em causa. *Logistic Regression* obteve os valores mais baixos entre os 3 métodos. Em termos de *Score, Recall* e *F1 Score* o *Random Forest Classifier* foi o melhor método, em termos de *Precision* o método que obteve melhores resultados foi o *Adaboost Classifier*.

Foram também analisadas as variáveis de entrada com mais importância para cada um dos métodos. Para o *dataset* COMPAS as variáveis, tanto para o método *Adaboost Classifier* como para o *Random Forest Classifier* foram as mesmas, *age, priors_count* e *c_jail_time*. Já para o *dataset* NIJ as variáveis com mais importância no metodo *Adaboost Classifier* foram *Age_at_Release, Prior_Arrest_Episodes_Felony* e *Delinquency_Reports*. Para o método *Random Forest Classifier* as variáveis classificadas com mais importância foram *Age_at_Release, Residence_PUMA*, *Pior_Arrest_Episodes_PPViolationChargers*.

Destes resultados podemos retirar que a idade é das características que mais influenciam os resultados de previsão em conjunto com o numero de crimes prévios. É de realçar que as características de género e etnia não estiveram entre as mais importantes para a previsão de reincidência na construção dos modelos descritos.

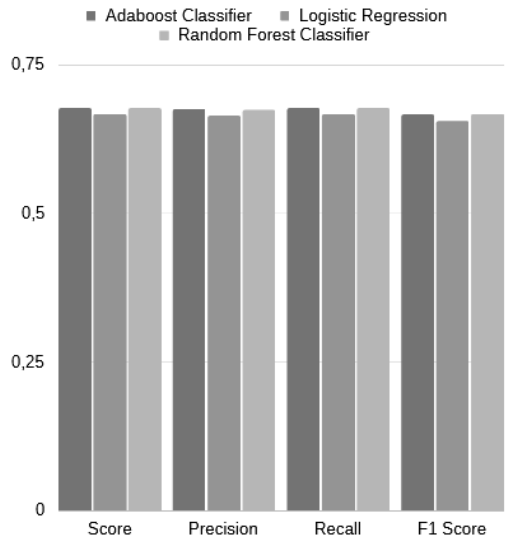

Figura 6.2: Resultados das métricas para o *dataset* NIJ.

6.3 Conclusões Finais

Nos dias que correm estabelecimentos prisionais inteligentes já são uma realidade em varias partes do mundo. Estes sistemas contribuem tanto para a monitorização de ofensores como para a ajuda de tomada de decisões. Estes sistemas são chamados sistemas de apoio à decisão, e com a implementação destes sistemas técnicos e responsáveis podem tomar decisões melhores tanto para os ofensores em si como para o resto da comunidade prisional. No entanto estes sistemas são alvo de investigações e artigos referentes à justiça e ao viés presente nos mesmos na tarefa de indicar valor de risco de reincidência para um ofensor, podendo estes classificar erradamente certo ofensor por características como género ou etnia.

O tema desta dissertação incide exatamente neste tópico. Este documento começa com uma breve introdução [1] ao tema e ao uso de inteligência artificial em sistemas prisionais em geral. É também apresentado o referido possível problema do viés e justiça nestes sistemas. De seguida é apresentado o estado da arte [2] relevante ao tema. Neste estado da arte são apresentados em detalhe dois algoritmos de apoio à decisão, o LSI-R e o COMPAS. Referente aos algoritmos são estudados artigos em que estão presentes. Ainda neste capitulo de estado da arte são estudados mais um numero de artigos com ferramentas e populações variadas. Depois deste estudo do estado da arte é foi concluído que poderá existir uma potencial injustiça ou viés em alguns destes instrumentos, nomeadamente no instrumento COMPAS. Por vezes esta disparidade pode ser eliminada ao identificar as diferentes necessidade presentes no ambiente e adaptar o instrumentos para estas necessidades. É possível também concluir que métodos de *machine learning* usados no mesmo contexto por vezes podem também em si ser discriminatórios e devem ser usados em conjunto com métodos de mitigação de viés.

No capitulo seguinte intitulado de *Planeamento e Trabalho a Realizar* [3], foi identificado o trabalho que que se viria a desenvolver no decorrer desta dissertação. Este trabalho dividiu-se em duas partes, em primeiro lugar uma analise de *datasets* com informação referentes à reincidência de um ofensor, e numa segunda parte o uso de métodos de *machine learning*

para a construção de modelos capazes de prever a reincidência de um dado ofensor.

No capitulo 4 [4], *Tecnologias Utilizadas* foram apresentadas todas as ferramentas e tecnologias utilizadas nesta dissertação. Estas incluem *Python*, a linguagem escolhida para a construção dos modelos e analise de dados dos *datasets*, *Sklearn*, um modulo *Python* que fornece ferramentas para a criação, treino e teste dos modelos. Por fim o IDE escolhido para a programação foi o *Pycharm*. Ainda neste capitulo foram apresentadas duas outras ferramentas que foram utilizadas em trabalho adicional descrito na secção 6.4, estas foram o *Swagger* e *Postman*.

No capitulo 5 [5], *Desenvolvimento*, foram analisados dois *datasets* escolhidos, *NIJ Recidivism Challenge Dataset* e *Compas Scores Dataset*. Nesta analise foram apresentadas todas as colunas de ambos os *datasets* e só depois foi iniciada a analise de dados que constituem os *datasets*, esta analise focou-se nas características de idade, género e etnia. Ainda neste capitulo forma apresentados os métodos de *machine learning* utilizados, *Adaboost Classifier, Logistic Regression* e *Random Forest Classifier*, e foram apresentados todos os resultados obtidos pelas métricas de avaliação, *Score, Precision, Recall* e *F1 Score*, e ainda as variáveis de entrada com mais importância para cada um dos métodos.

Com o desenvolvimento desta dissertação e todas as tarefas envolventes foi possível concluir que grande parte da população prisional é constituída por ofensores masculinos de etnia negra e de idades entre os 20 e 40 anos, aproximadamente. No que toca a avaliação de reincidência ofensores masculinos são classificados como reincidentes em maior numero do que ofensores femininos, é de referir que a população masculina é também em geral maior que a população feminina. Já quando é feita a avaliação da reincidência dos ofensores com base na sua etnia, ofensores brancos tendem a ser classificados de forma decrescente no sistema COMPAS, ou seja ocupam em maior numero os níveis de menor risco enquanto que os ofensores de etnia negra tendem a ser classificados de uma forma uniforme tanto nos níveis de menor risco como de risco mais elevado. O *dataset* NIJ também vai de encontro a está conclusão em que ofensores negros são classificados em maior numero como reincidentes como não reincidentes do que ofensores brancos. Com a analise dos métodos de *machine learning* foi possível retirar que o método *Adaboost Classifier* foi o que obteve bons resultados em ambos os *datasets* ainda que no NIJ o método *Random Forest Classifier* teve melhores resultados nas métricas usadas. É também possível concluir que as características que mais influenciam a previsão de reincidência são a idade e crimes anteriores e que a etnia e género, em nenhum dos métodos foram as mais importantes.

Para finalizar grande parte dos instrumentos descritos e modelos de *machine learning* usados em ambientes reais foram criados e testados com dados que por si já podiam conter um certo viés, por essa razão, qualquer ferramenta criada vai também conter esse viés, ainda que umas sejam classificadas como mais ”justas” que outras. Concluindo, o problema da justiça e viés da inteligência artificial em algoritmos de apoio a decisão tem como por base o dados usados já serem de alguma forma enviesados. Sendo este problema difícil de combater por estes dados serem baseados na realidade em que estão inseridos, portando a utilização de ferramentas como as descritas neste documento tem que ser devidamente analisadas e

adaptadas à realidade do ambiente onde estão a ser implementadas.

6.4 Trabalho Adicional

Durante a escrita da dissertação estive inserido num projeto participado com uma bolsa de investigação que estava de certo modo relacionado com o tema da mesma. Durante este projeto participei na construção de um algoritmo de apoio à decisão baseado em regras. Desenvolvi uma API, com o uso de *pyhton* e *swagger*, para a disponibilização dos serviços oferecidos pelo algoritmo de apoio à decisão. Fui também desenvolvedor de uma aplicação prototipo para a aprovação de resultados obtidos pelos sistemas referidos anteriormente. Mais recentemente participei na implementação de métodos de *machine learning* desenvolvidos por outro membro do projeto, na API já referida. Para além destes trabalhos práticos, participei também na escrita de alguns artigos, mais intensivamente na escrita de uma revisão sistemática sobre o tema da minha dissertação em conjunto com outro membro inserido no projeto.

Bibliografia

[1] J. Palter. (2021) Law enforcement technology and the future of prison systems. [Online]. Available: https://www.realtimenetworks.com/blog/law-enforcement-technology-and-the-future-of-prison-systems v, 1

[2] S. Dharmaraj. (2019) Tech to make singapore prisons smart. [Online]. Available: https://opengovasia.com/tech-in-singapores-prisons/ v, 1

[3] P. Puolakka. (2022) Smart prison: From prison digitalisation to prison using, learning and training artificial intelligence. [Online]. Available: https://justice-trends.press/smart-prison-from-prison-digitalisation-to-prison-using-learning-and-training-artificial-intelligenc v, 1

[4] L. K. Jeff Larson, Surya Mattu and J. Angwin, "How we analyzed the compas recidivism algorithm," 2016. [Online]. Available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm xi, 11, 12, 26

[5] Master the adaboost algorithm: Guide to implementing understanding adaboost. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/#:~:text=AdaBoost%2C%20also%20called%20Adaptive%20Boosting,are%20also%20called%20Decision%20Stumps. xi, 39, 40

[6] Logistic regression in machine learning. [Online]. Available: https://www.javatpoint.com/logistic-regression-in-machine-learning xi, 42

[7] Decision trees. [Online]. Available: https://www.ibm.com/topics/decision-trees xi, 43, 44

[8] Understanding random forest. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 xi, 43, 44

[9] K. HOUSER. (2019) Hong kong has a plan to make all of its prisons "smart". [Online]. Available: https://futurism.com/smart-prisons-hong-kong 1

[10] C. Metz and A. Satariano. An algorithm that grants freedom, or takes it away. [Online]. Available: https://www.nytimes.com/2020/02/06/technology/predictive-algorithms-crime.html 2

[11] A. Chohlas-Wood. (2020) Understanding risk assessment instruments in criminal justice. [Online]. Available: https://www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice/ 2

[12] B. of Justice Assistance. What is risk assessment. [Online]. Available: https://bja.ojp.gov/program/psrac/basics/what-is-risk-assessment#ehgvzl 2

[13] Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin. (2016) How we analyzed the compas recidivism algorithm. [Online]. Available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm 3

[14] E. M. Lowder, M. M. Morrison, D. G. Kroner, and S. L. Desmarais, "Racial bias and lsi-r assessments in probation sentencing and outcomes," *Criminal Justice and Behavior*, vol. 46, no. 2, pp. 210–233, 2019. [Online]. Available: https://doi.org/10.1177/0093854818789977 7

[15] P. Smith, F. T. Cullen, and E. J. Latessa, "Can 14,737 women be wrong? a meta-analysis of the lsi-r and recidivism for female offenders*," *Criminology & Public Policy*, vol. 8, no. 1, pp. 183–208, 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-9133.2009.00551.x 8

[16] E. J. Palmer and C. R. Hollin, "The level of service inventory— revised with english women prisoners: A needs and reconviction analysis," *Criminal Justice and Behavior*, vol. 34, no. 8, pp. 971–984, 2007. [Online]. Available: https://doi.org/10.1177/0093854807300819 8

[17] K. W. Whiteacre, "Testing the level of service inventory–revised (lsi-r) for racial/ethnic bias," *Criminal Justice Policy Review*, vol. 17, no. 3, pp. 330–342, 2006. [Online]. Available: https://doi.org/10.1177/0887403405284766 9

[18] Northpointe. (2012) Compas risk need assessment system selected questions posed by inquiring agencies. [Online]. Available: https://www.labecedaire.fr/wp-content/uploads/2017/08/FAQ_Document.pdf 11

[19] T. L. Fass, K. Heilbrun, D. DeMatteo, and R. Fretz, "The lsi-r and the compas: Validation data on two risk-needs tools," *Criminal Justice and Behavior*, vol. 35, no. 9, pp. 1095–1108, 2008. [Online]. Available: https://doi.org/10.1177/0093854808320497 13

[20] S. Tolan, M. Miron, E. Gómez, and C. Castillo, "Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia," 2019. [Online]. Available: https://doi.org/10.1145/3322640.3326705 14

[21] M. Karimi-Haghighi and C. Castillo, "Enhancing a recidivism prediction tool with machine learning: Effectiveness and algorithmic fairness," 2021. [Online]. Available: https://doi.org/10.1145/3462757.3466150 17

[22] Python. [Online]. Available: https://www.python.org/ 21

[23] Guido van rossum. [Online]. Available: https://computerhistory.org/profile/guido-van-rossum/ 21

[24] scikit-learn - machine learning in python. [Online]. Available: https://scikit-learn.org/stable/ 21

[25] Scipy. [Online]. Available: https://scipy.org/ 21

[26] David cournapeau. [Online]. Available: https://scholar.google.com/citations?user=ua46uL0AAAAJ&hl=en 21

[27] Google summer of code. [Online]. Available: https://summerofcode.withgoogle.com/ 21

[28] Jetbrains. [Online]. Available: https://www.jetbrains.com/ 21

[29] Api development for everyone. [Online]. Available: https://swagger.io/ 22

[30] Tony tam. [Online]. Available: https://wellfound.com/p/tony-tam 22

[31] Apis together. [Online]. Available: https://www.postman.com/ 22

[32] Abhinav asthana. [Online]. Available: https://www.crunchbase.com/person/abhinav-asthana-2 22

[33] Nij's recidivism challenge full dataset. [Online]. Available: https://data.ojp.usdoj.gov/Courts/NIJ-s-Recidivism-Challenge-Full-Dataset/ynf5-u8nk 23

[34] Compas recidivism risk score data and analysis. [Online]. Available: https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis 25

[35] compas-recidivism. [Online]. Available: https://huggingface.co/datasets/imodels/compas-recidivism 25

[36] The ultimate guide to adaboost algorithm | what is adaboost algorithm? [Online]. Available: https://www.mygreatlearning.com/blog/adaboost-algorithm/#:~:text=AdaBoost%20algorithm%2C%20short%20for%20Adaptive,assigned%20to%20incorrectly%20classified%20instances. 39

[37] What is logistic regression? [Online]. Available: https://www.ibm.com/topics/logistic-regression#:~:text=Related%20solutions-,What%20is%20logistic%20regression%3F,given%20dataset%20of%20independent%20variables. 41