The University of Sussex

School of Engineering & Informatics

Masters Dissertation

Programme: MSc in Artificial Intelligence and Adaptive Systems.

Candidate number 2291444

Title of Dissertation: Prison Inmates with Violent Crimes: a new dataset and datasheet for the machine learning fairnss community

Approximate number of words: 13,100 (excluding abstract, captions, references and appendices)

Supervisor(s) Novi Quadrianto

|  |
| --- |
| Date submitted ………………... Time submitted ……………………. |

|  |
| --- |
| **Declaration**  I certify that the information on this cover sheet is correct.  I certify that the content of this dissertation is my own work, and that my work  contains no examples of misconduct such as plagiarism, collusion, or fabrication of results.  Candidate’s signature ………………………………………………………………. |

Inmates with Violent Crimes: a dataset for machine learning fairness

Abstract

Machine bias is an active research topic. Research to date has found a significant amount of machine bias starts with data. As a result, techniques for removing bias from datasets have emerged. Machine learning fairness researchers have also produced data transparency documentation that is designed to highlight dataset provenance and help system designers identify potentially biases or ethical concerns with datasets. A limiting factor in machine learning fairness research is the availability of datasets. This work aims to provide the machine learning fairness community with a new dataset that is supported by data transparency documentation. This paper discusses how to transform the US Survey of Prison Inmates 2016 into a usable machine learning fairness dataset and provides supporting code. In it’s raw form, the Survey documentation and data is difficult to navigate and so the structure of the survey documentation as well as the data is explained. The dataset is designed to predict whether an offender will receive a harsh or not harsh sentence. The process of selecting features for including in the survey involved evaluating classifiers trained on the dataset for accuracy on this task. Some of the most important features (such as sentence length and offenses) are challenging to derive, and the process of identifying how to best process these variables is described in detail. The resulting dataset is evaluated for bias, as are classifiers trained using the dataset. A simple decision tree classifier achieves 80% accuracy on this task; the same score achieved by less explainable models with much higher capability. Bias assessment shows there is bias in the dataset, and that the classifiers amplify this bias for certain protected attributes. Further work is needed to evaluate a broader range of sensitive attributes, derive additional features and apply different feature derivation techniques to existing features.

# Sensitive Topic Advisory & Project Ethics

This paper and the accompanying code makes reference to violent crimes such as murder, rape, sexual abuse, bodily injuries suffered as a result of violent crimes, domestic abuse and crimes with child victims. The references to these topics are limited as they appear in the context of processing data about violent crimes. Many of the variable names are codes (i.e., V1234) rather than semantic descriptions, but the Appendices contains semantic descriptions of the variables, as does the variable configuration file in the codebase (which acts as a metadata dictionary).

This paper describes processing that can be applied to data collected in 2016 Survey of US Prison Inmates (the Survey) [1] in order to produce a version of the dataset that is suitable for use in machine learning fairness research. Participants in the Survey consented to their data being collected and released into the public domain via an informed consent protocol. Participants were provided with information about the research prior to the survey interviews. During the survey interview, the consent process took two minutes on average. [[1]](#footnote-1) Participants had the option to refuse to answer any questions. The version of the dataset used is the public use dataset. Any variables from the full Survey dataset that could potentially lead to participant re-identification (such as date of birth and offense names) were suppressed before publication.[[2]](#footnote-2)

Given the density of the documentation and dataset, the sensitive nature of some of the content that would be in the final report and dataset was not known at the beginning of the project. The need to reference sensitive topics (as stated above) emerged during the investigation and the sensitive topic advisory above was added to the survey.

Another ethical consideration that arose is that the victim information (sex, age, race, injuries suffered, relationship to offender) was self-reported by the offender. In this sense, victims have little agency in how this data was produced and so it would be advisable not to use this data to predict victim information; rather, the victim information should be considered as proxies that may or not correspond with ground truth.

Offense codes mean it is not possible to distinguish between abortion, child abuse and gang crime as they are grouped under the same offense code in US law (see Figure 0.1). This means that any model trained on the classifier would not be able to distinguish between these offenses.

As final ethical consideration; a great amount of detail is collected about how guns are used in crimes and how they were obtained. Comparatively, there is little detail on areas such as the history of the offender’s relationship with the victim.

Much of this ethical discussion is repeated in the accompanying datasheet.

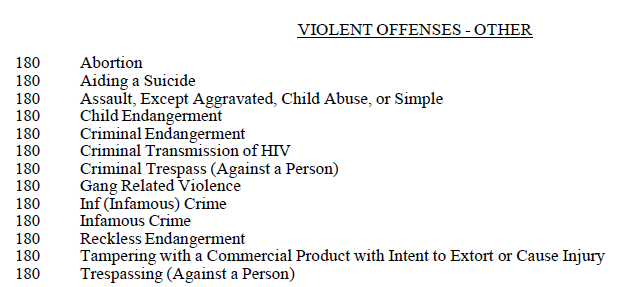


Figure 0.1‑1The offense code 180 refers to abortion, aiding suicide and other offenses, meaning these cannot be distinguished by classifiers.

# Background and Motivation

## Machine bias is an active research topic

There is a growing concern among consumers and governments about the harms of artificial intelligence and automated decision making [2]. Indeed, the use of probability to model human populations, has been contentious both scientifically and politically for over a century [3] [4], with many arguments objecting to the use of statistics in decision making about humans (see Chapter 2 in [5]).

Discrimination is one such concern. In the domain of machine learning fairness, discrimination refers to bias that disadvantages demographic groups that have been subject to ‘unjustified and systematically adverse treatment in the past’ [5]. Machine bias is an active research topic [6] and it’s well evidenced that AI can amplify pre-existing social biases [7]. In recognition of the harms of algorithmic bias, governments in Europe [8], Norway [9],China [10] and the US [11] have introduced algorithmic fairness regulations.

## Machine bias is often rooted in data

The machine learning pipeline involves collecting data, preprocessing data, labelling data and training a model using this curated data. There is a risk of bias being introduced at any of these stages, and therefore bias should be measured at each one. [12]. However, a significant proportion of machine bias starts with data [12].

There are many reasons that data will produce a biased classification [7]. In a typical dataset, many features have a correlation with the sensitive attribute that is protected by law [5], and a classifier may learn a relationship between these correlated (but unprotected) attributes and the target label. The underrepresentation of unprivileged groups in society can lead to under representation of those groups in machine learning datasets [13]. Gebru and Bulowami’s work on facial recognition classifiers is a seminal example [14]. The target label assigned in a dataset is often a proxy for the real truth we are trying to reach in the decision making, and this proxy may be informed by existing bias. There is also the issue of how power dynamics between unprivileged groups and privileged groups lead to sample bias. Consider crime datasets used for predictive policing; such datasets are riddled with sampling bias because data is only sampled from the areas that have historically been surveilled for criminal activity [15] .

## Data debiasing techniques are part of the solution

As a result of the clear link between datasets and bias, the European Union’s Ethical Guidelines for Trustworthy AI recommend that the socially constructed biases need to be addressed prior to using the data [8]. Similarly guidelines produced by the Turing Institute and the UK Office for Artificial Intelligence recommend system designers should ‘use only fair and equitable datasets [2].

Methods available to remove bias from a dataset include oversampling with real data [16] or synthetic data [17], preferential sampling [18], reweighing [18] and. disparate impact removal [19]. Most of these preprocessing techniques aim make the target label independent of the sensitive features.

A challenge many system designers face when selecting and applying data debiasing methods is how to determine when a dataset, and the output of a model it’s trained on, is fair. There are 21 mathematical definitions of fairness [20] that cannot be satisfied simultaneously [21]. A survey of fairness interventions, [22] encourages practitioners to avoid the proliferation of fairness metrics. It also highlights that fairness metrics can be brittle with regards to data preprocessing and training-test comparison. For this reason, data debiasing alone will not resolve machine bias rooted in data.

## Dataset transparency and availability has an important role in countering data bias

Contemporary data dissemination practices tend to abstract away the decisions that lead to the bias; whether that be who a decision about where the collect data from or what labels to apply [5]. As a result, data transparency methods have received increasing attention from academia and industry [23]. Data transparency documentation is used to highlight data provenance and make users are aware of the power dynamics in which a dataset was produced. Examples of dataset documentation include Bender and Friedman’s Data Statements for Natural Language Processing [24] and Gebru et al’s Datasheets for Datasets [25]. Research has shown that these dataset documentation methods help engineers identify ethical considerations in datasets and ethical issues in training data. [26]

Limited dataset availability is a constraining factor for algorithmic development in many domains [27]. There a particular lack of datasets, supported by transparent documentation, in machine fairness; the adult dataset provided the basis of analysis for over 300 fairness research papers [28]. This may be because data production is less incentivized compared to building novel machine learning models [29].

Criminal Justice datasets are of particular interest to the machine learning fairness community given the historical biases and power imbalances in policing. ProPubclia’s analysis of a recidivism algorithm used in the US [30]is often referred to as an example of how historical bias in criminal justice manifests in algorithmic decision making.

The need for more datasets, supported by dataset documentation, and interest in criminal justice datasets in particular is the driving motivation for this project. Zilka et al have conducted a survey on the availability of US Criminal Justice Datasets and provided a datasheet for each dataset [31], including a datasheet for the Survey of US Prisoners 2016. This project aims to provide a deeper look into the Survey and create a dataset from it that can be used in machine learning fairness research.

People in the U.S. are more likely both to be sentenced to prison and to serve a longer sentence than in comparable nations. Between 1994 and 1999, following federal funding incentives, 29 US States removed the ability of parole boards to release people early [32]. This makes sentence length a pertinent variable to predict.

## Contribution

The primary contribution of this work is to make the US Survey of Prison Inmates accessible for use in machine learning fairness research. Although the data and codebook is available online, the documentation is opaque and difficult to navigate, meaning that the effort required to use the dataset makes it impractical for most machine learning researchers.

The final dataset contains a number of features (see Appendix 3 full a full list), a target label of harsh/not harsh sentence, and four binarised sensitive attributes. The Further Work section suggests other features that could be sensitive attributes, and other possible target variables. The Further Feature Derivation Appendix (appendix 16) lists how further features could be derived in

The below artefacts have been created and stored in the US Prisoners Github repository [33]:

1. This paper, which describes how to navigate the survey documentation and how to locate and derive variables of interest to the machine learning community.
2. Datasheet for the Prison Inmates with Violent Crimes Dataset that summarises the dataset contents, provenance, pre-processing applied and highlights ethical considerations
3. Diagrams of the survey’s structure and routing logic that support explanations of derived variables in the Datasheet and to help users understand which Survey variables may be of interest. (Appendices 2a-2c)
4. The Python package US Prisoners, which contains the code required to reproduce the Violent Crimes Subset from the US Prisoner Survey 2016 data files. The package configuration options to control how some variables are derived, the threshold for the ‘harsh’ sentence label and to include new variables from the Survey.
5. In the above, a variable configuration file containing key information for most variables in the SPI (any groups of variables that are excluded are references in the variable\_config README). This variable configuration file makes it easy to include new variables in the dataset.
6. A Python notebook[[3]](#footnote-3) that demonstrates how to:
   1. Load the us\_prisoners package
   2. Find the original SPI data and use the US Prisoners package to transform it into the dataset described in this paper the datasheet.
   3. Assess the Inmate with Violent Crimes dataset for bias using the aif360 toolkit by create custom sklearn scorers using aif360 metrics and then scoring for accuracy and fairness metrics across various model hyperparameters with fivefold cv using sklearn GridSearch and Pipeline
   4. Inspect a given decision tree in order to understand how the model is using features.
   5. Load the automl package and fit a automl estimator to the Violent Crimes Dataset, and then assess this model fair fairness using the aif360 metrics.

Source code for the US Prisoners package is publicly available on GitHub, but the notebook is a convenient way of demonstrating the package's capabilities. The processing applied in the notebook aligns to that described in the Inmates with Violent Crimes Datasheet and this paper. If the processing parameters are adjusted, or if the variable configuration file is updated, the resulting dataset and any models trained on it will vary from the results discussed here and in the Datasheet.

The dataset output by the code is not currently hosted online as the author is unclear about distribution permissions regarding derived versions of the Survey data.

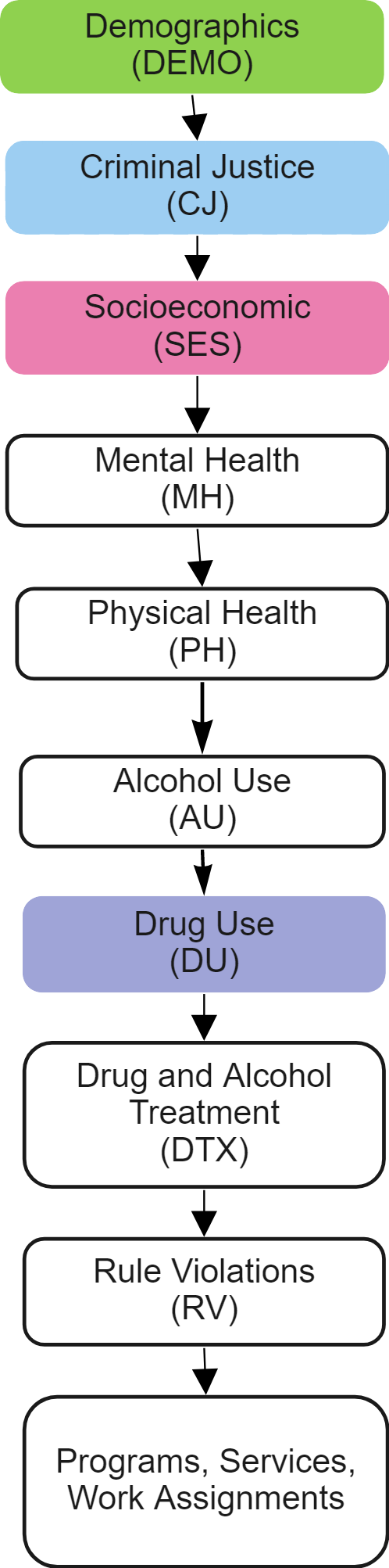
# Exploratory Analysis

## Original Dataset Source

There are two public use versions of the SPI 2016: data one contains data from state and federal institutions combined, and the other data from state-level only. This paper discusses the combined state & federal version of the dataset. If replicating this work, please following the instructions for downloading the dataset as the Survey can be accessed from multiple places on the ICPSR (Inter-university Consortium for Political and Social Research) and BJS (Bureau of Justice Statistics) websites.[[4]](#footnote-4)

## Navigating the Survey

The survey is split into elven sections. Two of these are the interview set up and close out screens, so there are nine sections of interest.



**Figure 3.2.1‑1** High level survey sections diagram

A series of resources produced as part of this project illustrate the survey structure and contents. The Survey Sections diagram (Fig 3-1) shows the sequence of the survey sections. Appendix 1 is a table describing the information gathered in each section.

Some sections contain Blocks. Participants are only asked the questions within a block if they are routed to it based on answers to previous questions. In some cases, there are groups of questions that behave very much like block, but are not officially designated as a block in the survey. These are referred to a pseudo -blocks in this paper; establishing this concept is helpful when explaining how variables have been derived. The routing to blocks and pseudo blocks within sections can be seen in Appendix 2 A, the High Level Survey Routing Diagram. In the case of the Criminal Justice section, there are two additional diagrams with further detail; Appendix 2B, the Inmate Type Determination and Offense Collection and Appendix 2C, the Sentence Collection and Controlling Offense Determination diagram.

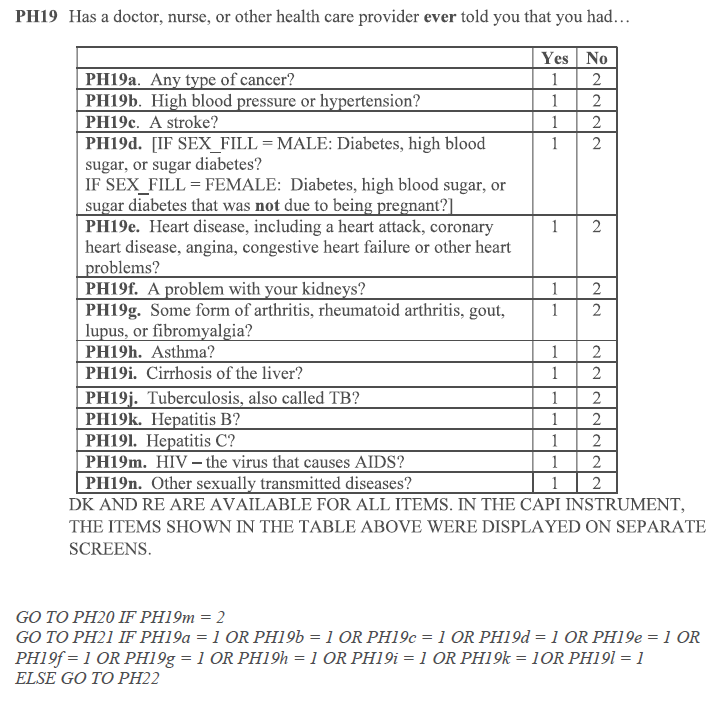
### Mapping variables to survey questions

Two documents need to be cross-referenced in order to understand which variables(s) contain data pertaining to a given question. These are the Variable Description and the Questionnaire, both of which are contained in the Codebook PDF document [34].

The Survey was conducted using a CAPI (computer assisted personal interview) method.

*The Questionnaire* contains the survey questions, the instructions to the surveyor, the information the prisoner was presented with on screen during the survey and the software programming logic for routing and auto-fill throughout the survey sections. The area annotated ‘Routing instructions for programmer’ in Fig 3.1.2-2 is an example. Fig 3.1.2-2 explains the question code syntax; the question code is PH19. ‘PH’ indicates the question is from the ‘Physical Health’ section, and 19 is the question number . The items in the table below the main question are suffixed with a letter, and this indicates that each of these rows is treated as a sub question for the purposes of how prisoners answer and how the data is stored.

The *Variable Description* document contains a list of all variable names and the possible values held within each variable. Most variable names being with a V and are the column name in the dataset. Figure 3-2-2 is an example of a variable listing in the Variable Description document.

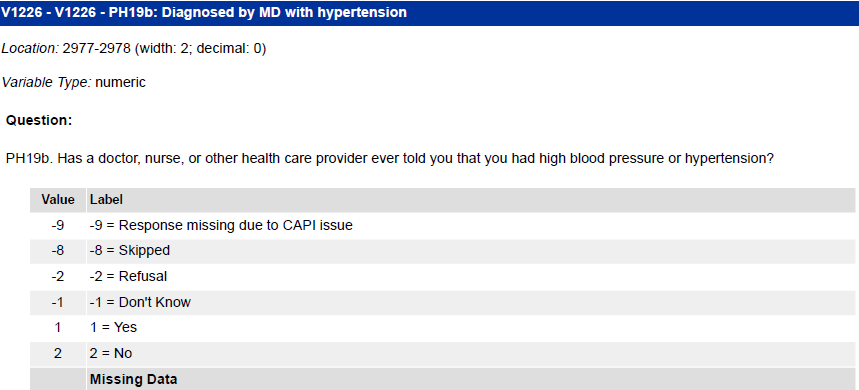


Question code: section code (PH) + question number (19)

Routing instructions for programmer

Sub question code appended with letter

**Figure 3.2.1‑2** PH19 question from the Questionnaire. The text underneath the table (which contains possible answers to PH19) describes the next question the prisoner should be asked (the routing) depending on their answers to this question. Notably, there is no indication of which variable(s) hold the sub-question responses.



Variable name V1226

Corresponding question code

Short description of question

Full question shown to participant

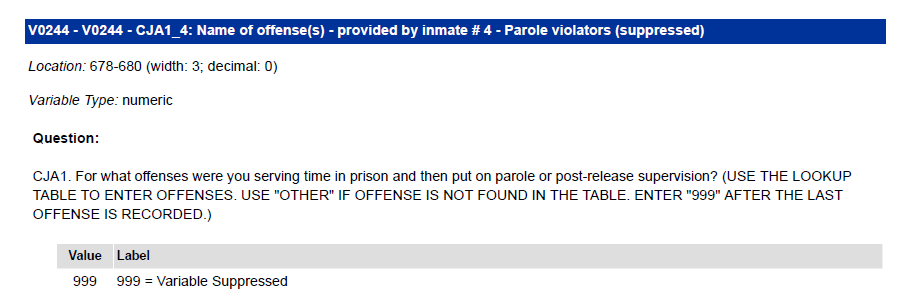
Possible responses and corresponding number encodings

**Figure 3.2.2** V1223, the variable for PH19b. The grey table lists the possible inputs for the question; in this case ‘1’ indicates a positive response and ‘2’ a negative response.

In order to see which variables question responses are stored in, the question number needs to be looked up in the Variable Description document. The question number can be found in the blue band at the top of each Variable Description listing- annotated in Fig 3.2.2 as ‘Corresponding question code’.

When a question is a sub question, it will be suffixed with a small letter- see the caption Fig 3.2.1-2. When a question is part of a block, a letter indicating the block is appended to the section code. Within a block the, question numbers reset at 1. Fig 3-2-3 shows ‘CJA1\_4’; this is the fourth variable holding information relating to question 1 in block A of the CJ (Criminal Justice) section.

A mapping between question codes and variable names is not provided in the survey documentation or the survey data. As part of this project, a dictionary that contains metadata about each variable has been created and is available on the US Prisoners Github as the file variable\_config.csv.



CJA1\_4: this variable holds data relating to question 4 in block A of the CJ (Criminal Justice) section. A in the criminal justice 9Cj) block

This variable has been suppressed, therefore all entries have been replaced with the value 999 and there are no answer codes.

**Figure 3.2.1‑1 3** An example of a variable that corresponds to a question in a block and is also suppressed.

### Answer Encoding

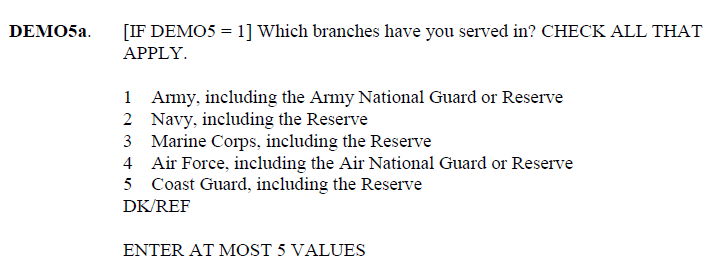
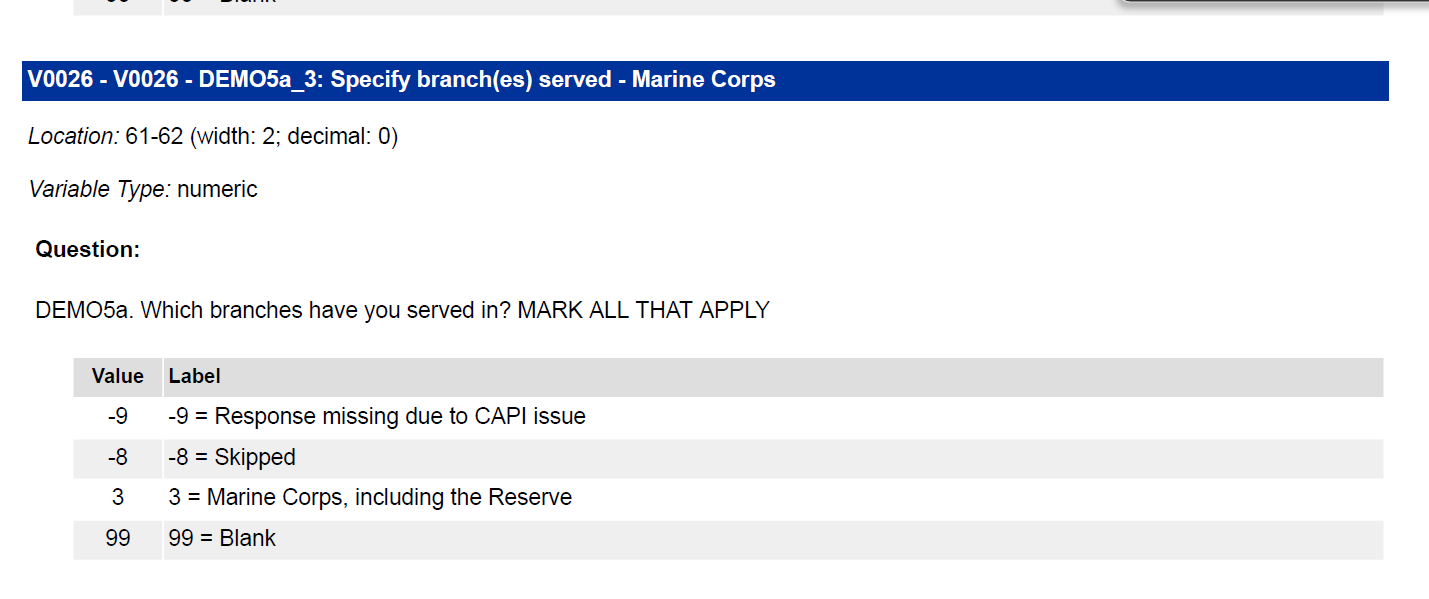
Sub-question Style

Fig 3-2-1-2 is an example of a question (PH19) with multiple sub questions. In such questions, a positive answer to each sub question is coded as 1 and No as 2. However, this mapping of 1 to Yes and 2 to No is not consistent throughout the survey, so must be checked for each variable. For this variable -9,-8,-2,-1 are used to indicate CAPI issue, question skip, refused to answer and don’t know respectively. Once again. these codes are used for most variables, but not all. The variable configuration file in the US Prisoners package allows users to set which values indicate which values represent Don’t Know, Refuse and Question Skip for each variable, as well as how each variable should be encoded or scaled.

Select Multiple Style

Some questions invite participants to select multiple options, but (unlike in the example of PH19), do not treat each option as a sub-question with regards to variable naming and answer encoding. The text ‘Mark all that apply’ indicates such questions. The selected answer is indicated by a number corresponding to the order in which the options were presented on screen. This is illustrated in Fig 3.3, which shows the Demo 5a in the questionnaire and variable V0028, where an answer of 3 indicates that ‘Marine Corps’ was selected as one of the answers. ‘Marine Corps’ was the third option on the screen. There are further variables for each possible answer

For both the sub-question and select multiple style questions, if the participant submits Don’t Know or Refuse to the parent question (or if the question is skipped) the appropriate code (usually -1 or -2, -8) will appear across all variables holding the answers to that question.



**Figure 3.2.2‑1**: The listing for Demo5a in the questionnaire and the corresponding variable for the ‘Marine Corps’ answer.

Upcoded Variables

A number of questions permitted an open-ended response option which was upcoded into existing question response options.

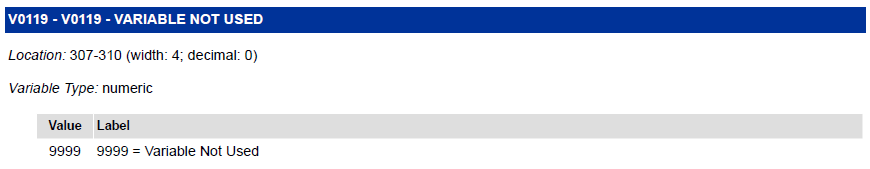
In these cases, the question will have at least twice the amount of variables than possible answers- one set with ‘Original’ in the variable description and another with ‘Upcoded’, and data users should decide which set of variables they wish to use. The upcoded version of the variables are usually listed first in the Variable Description document, followed by the Original versions.

Suppressed Variables

Suppressed variables are still included in the dataset, but all rows are filled with the dummy value 999, as shown in Figure 3-2-3.

Unused variables

A number of variables were not used, but are contained in the dataset. Those variables have a description of ‘Variable Not Used’ in the Variable Description Document and are filled with 9999 as in Fig 3-5.

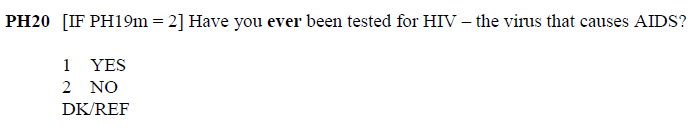


**Figure 3.2.2‑2** An example of how an unused variable appears in the Variable Description document.

### Identifying Survey Routing

There is no contextual information about each variable in the Variable Description document. For example, the routing instructions below PH19 (shown in Fig 3.2.1‑2) say that the inmate should be routed to PH20 if they have ever been told they have HIV. The text above PH20 (shown in Fig 3.2 a) in the Questionnaire does indicate that the route to this question is via PH19- however it refers to the numerical answer codes (‘IF PH19m,=2]), rather than the natural response (i.e., if ‘PH19 (HIV)=‘Yes’), which makes it difficult to follow the routing. When viewing the variable descriptions for both PH19 and PH20 (Fig 3.2‑b), there is no indication of the routing; this can only be inferred from the Questionnaire. Therefore, the user has to go back to PH19 in the questionnaire to find the question and then search for PH19m in the Variable Description document to establish what answer was encoded as ‘2’ in order to understand what PH20 is asking. Figure 3-6 shows an even more complex example of survey routing text.

The survey routing can only be established by cross-referencing both documents and following the survey through different branches. Combined with the need to search through the Variable Description document for question numbers, this means that significant time has to be invested in order to understand what the data held in each variable represented. Documenting this, describing the logic to derive variables of interest and producing diagrams of the survey is one of the main contributions of this work.



**Figure 3.2-a:** The text [IF PH19m=2] indicates that this question should be answered if the answer to PH19m (HIV) was 2 (Yes).

***Figure 3.2‑b*** *The routing that leads to a question is indicated in the questionnaire, but not the variable description document.*

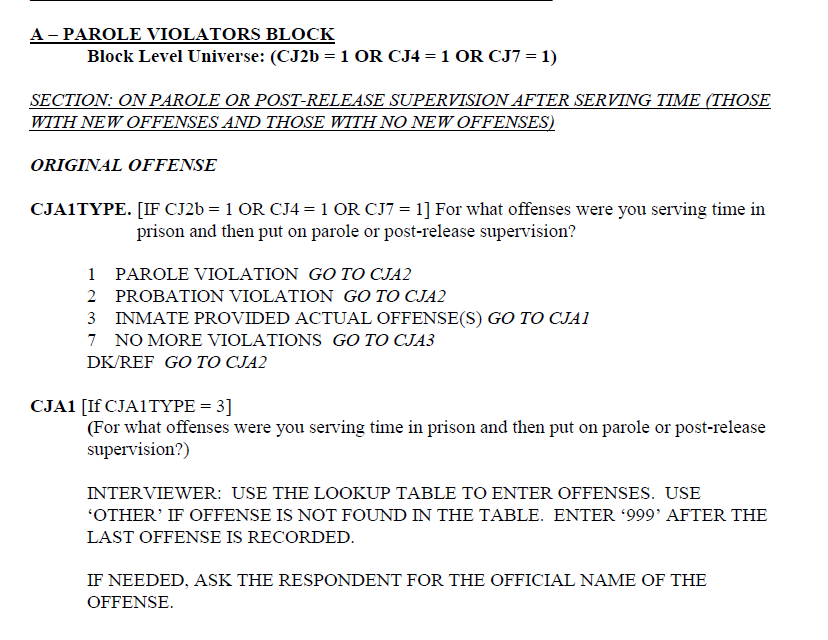
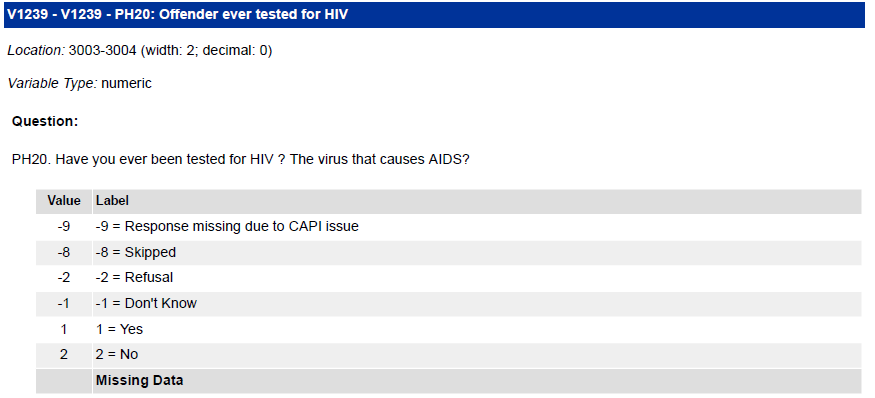
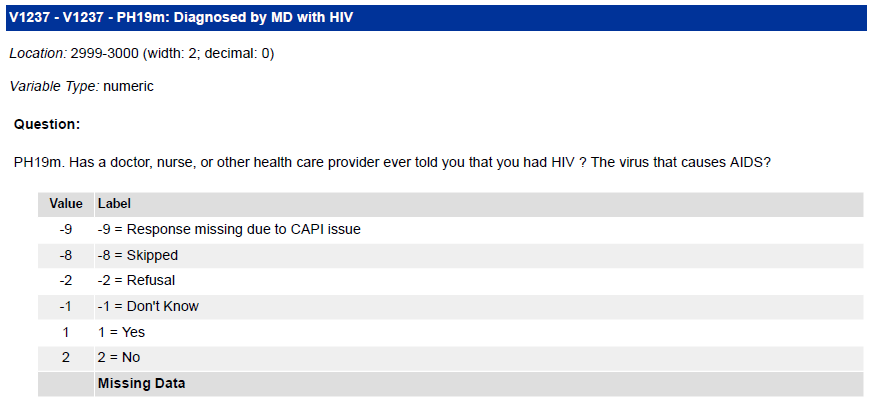


Figure 3.2.3‑1 An example of a larger block of routing information in the Questionnaire.

## Dataset Profiling

As a first step to building an estimator, an attempt was made to profile the data using the pandas profiling library [35].The Pandas profiling toolkit generates a report containing information including missing values, descriptive statistics and correlations between each column. Prior to understanding the shape of the data, the intention was to use this information to guide an initial feature selection.

The process of loading the data and trying to generate the report identified a number of issues relating to the dataset which had to be resolved before any feature selection could begin.

### Mixed data types in a given column

V0772 contains the state the participant was living in at the time of arrest. It contains as a mix of string and number types in the same column. This is because, although state given as a string (i.e, ‘NY’), there are numerical variables representing Missing, Don’t Know and Refuse values.

### *Representation of skipped questions*

The Variable Description document indicates that a question skip is coded as-8 for most variables, and sometimes 99 or 999. Elsewhere in the survey documentation, it’s stated that skipped numerical questions are represented by a period (‘.’) while in string variables it is represented by a blank value that Pandas identifies as nan.

Investigations showed that, where the Variable Description indicates -8 as a skip value, these entries contain missing data which is read as a nan value. Where the Variable Description document indicates 99 or 999 as a skip value, this is the skip value. It would not be wise to convert 99 and 999 to nan values across the entire dataset as these would be valid entries for some variables.

### High dimensionality

Due to the size of the dataset, the Pandas Profiling package was not able to generate a correlation report. Factors contributing to this high dimensionality were the suppressed variables, unused variables, and encoding and the same (or similar) information being collected in multiple branches of the survey.

As will be explained in more detail in the Section 4.3. there are instances where the same question is asked in different blocks or pseudo blocks. This results in multiple variables holding the same information (some of which then have sub questions and original and upcoded versions, multiplying the issue). For example, although only a maximum five current offenses are collected for each prisoner, there are approximately 400 variables which hold this information due to separate pseudo blocks for the 13 different Inmate Types; this is illustrated in the Inmate Type Determination and Sentence Collection diagram (Appendix 2b), where each white oblong with a broken border represents the same questions being asked. In conjunction with answer encoding, this results in the number of variables growing very quickly and a sparse dataset.

From the exploratory data analysis, it was concluded that

* a detailed mapping of the survey routing needed to be undertaken in order to understand which variables might provide useful features to a sentence length predictor
* features would need to be derived by combining data held across a number of variables in order to assess correlation amongst variables and assess the dataset for bias

# Selecting and Deriving Features

Given the number of variables and effort required to understand the survey routing, it was decided to start with fewer variables and build up. Fewer variables is also desirable to produce a less complex model that can be more easily explained [36]. As this dataset was being created to predict sentence length, any variables concerning the period of time after the prisoner had been sentenced needed to be removed. The variable configuration file on Github, and it’s associated Read Me, indicate why each variable was not included. Fig 4.-1 gives a visual summary of which sections were included or part included.

*Demographics Section*

Information about age bracket, sexual orientation, sex, race, marital status and if the offender has served in the US armed forces were included.

Whether or not the offender had served in the US Armed Forces did not show high feature importance using variance threshold, mutual information or chi2 feature selection techniques, so it was determined that further detail in the Military Service block did not need to be included .

*Criminal Justice Section*

The Criminal Justice section has the most complex routing and is the most sparsely populated of all sections*.* As is explained in sections 4.1.2-4.2.4*,* variables in the Criminal Justice section had to undergo extensive pre-processing in order to derive offense and sentence information.

The decision to predict sentence length as the target variable meant that only classes of Inmate Type where the inmate was serving a sentence needed be considered. Consequently, only variables corresponding to some of the offence capture pseudo blocks were needed. The red triangles on the Inmate Type Determination diagram (Appendix 2b) and Offense Capture diagram (Appendix 2c) indicate which pseudo-blocks were included. Table 1 shows the distribution across Inmate types; retaining only inmate types 3,8 and 11 meant a large majority of the dataset was retained whilst reducing the number of pseudo – blocks that had to be processed.

The decision to focus on instances where the controlling offense was a violent offense meant that blocks CJF (Property Crimes) and CJG (Drug Crimes) from the Criminal Justice section were not relevant. Table 2 shows the distribution across controlling offense types; the majority of crimes are violent crimes, so there was not a concern about reducing dataset size.

Variables containing information about whether or not time spent in prison whilst awaiting sentencing or trial was applied to the prison sentence were not included in the dataset as it was not clear from the survey how the answer to this question related to the target of sentence length.

*Socioeconomic Section*

Whether or not the participant was born in the US, number of years lived in the US, second country of citizenship, homelessness, if living in prison upon arrest, if living with children at the inmate of arrest (and how many), if homeless before 18, time in foster care, and if they received welfare or public housing before 18 were retained.

Detail about where children are living now the offender is in prison and the types of communication they have from their children was not included as this refers to a time after sentencing.

Only key information was included from the Education block; if the participant had a GED or High School diploma before entering prison, if they had been diagnosed with, dyscalculia or dyslexia or attended special educational classes.

*Physical Health Section*

Many of the questions in the Physical Health section could not be included as they covered the entirety of the prisoners life, I;e; ‘Have you ever been told you’ve had cancer?’ (PH19 in Fig 3-2-1-1). ‘Ever’ could refer to a period before or after the prisoner was sentenced.

The only information retained from Physical Health was heigh, weight, history of smoking and hospitalization 12 months before arrest.

*Drug Use and Alcohol Use Sections*

A select number of fields referring to if the prisoner was using drugs or alcohol at the time of the offense and 30 days before the offense were retained, but most information from this section was not included.

*Rule Violation and Programmes, Services and Work*

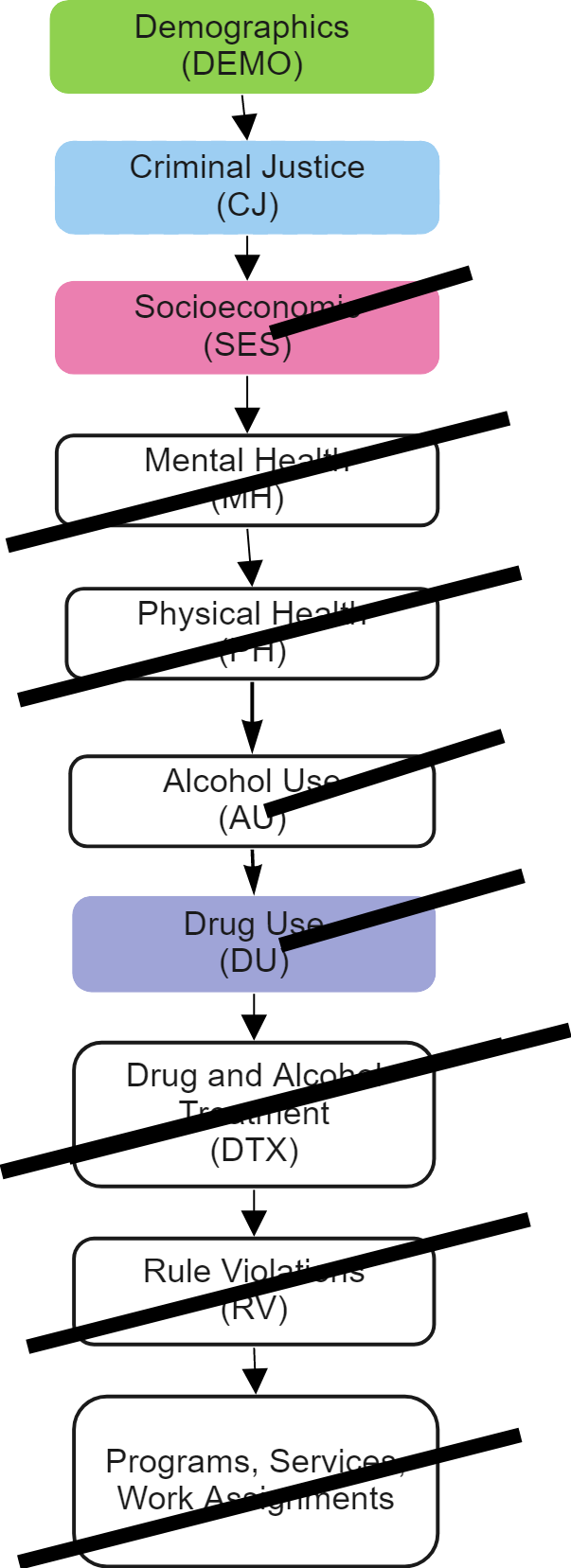
No variables from this section were retained as they refer to a period of time after sentencing.

**Table 1:** Distribution of controlling offense types across SPI 2016 data (federal and state combined).



**Table 2 (Below):** *Distribution of Inmate Types in original dataset. Only instances where the inmate type was 3,11 or 8 was considered. This reduced the number of pseudo sections that had to be mapped and processed, whilst still keeping 22340 of 25353 instances (before further filtering based on controlling offense type).*

|  |  |  |
| --- | --- | --- |
| Code | Description | Freq |
| 3 | Inmate not incarcerated for a parole or probation violation and not on parole or probation at time of arrest | 15597 |
| 11 | Probation violator with new sentenced offenses | 4248 |
| 8 | Parole violator with new sentenced offenses | 2495 |
| 10 | Probation violator with no new offenses | 996 |
| 7 | Parole violator for no new offenses | 430 |
| 12 | Probation violator with new arrest offenses | 291 |
| 9 | Parole violator with new arrested (unsentenced) offenses | 146 |
| 1 | Inmate awaiting trial or awaiting to be sentenced- not on parole, probation or escape at the time of arrest | 105 |
| 2 | Inmate not sentenced, not convicted/awaiting sentenced, not awaiting trial and being held for other authorities | 45 |
| 4 | On escape with no new offences | 0 |
| 5 | On escape with new sentenced offenses | 0 |
| 6 | On escape with new arrested offenses | 0 |



**Figure 3.3.3‑1** Sections immediately excluded from consideration have a strike through them. Selection where select question were included have a half strike.

## Feature Selection Evaluation Methods

Some of the variables in these sections did not require extensive processing to become features. For example, participants were asked their marital status in DEMO4 and the answer was stored in a single variable that could be one hot encoded.

The majority of information that was deemed likely to be useful to a sentence length classifier (such as offenses) needed to be consolidated from many variables or needed to be derived into a binary feature for use in machine learning fairness algorithms.

In order to test if there features (derived and non-derived) constituted a useful dataset for predicting sentence length, tests were conducted throughout the dataset development process. The data was used to train machine learning estimators to predict sentence length, with accuracy and feature importance scores guiding decisions about if the dataset was sufficient.

The intention was to use auto\_sklearn [37] to find the model with the best accuracy score for each version of the dataset. However, computational limitations meant this was not practical. [[5]](#footnote-5)

Therefore, the datasets were fed to logistic regression, multi-layer perceptron and decision tree models. K-best and variance threshold feature selection with various hyperparameters for each model were explored, 5 fold cross validation was used. Five was selected for the k-fold cross validation as 5 is a commonly used value that produces good results [38]. The notebook provides an example of evaluating various parameters for a pipeline with K-best feature selection and a Decision Tree model using scikit learn’s Pipeline and GridSearchCV toolkits.

After feature derivation, the following processing was applied to variables before evaluating how well the data performed on the sentence prediction task:[[6]](#footnote-6)

*Missing Data Don’t Know/Refuse* Although there are skipped columns, there is no truly missing data in the dataset. *Don’t Know and Refuse* were treated as missing values. As the frequency of Don’t Know and Refuse responses was low, there was not a need to create complex imputation, and they were replaced with the most common value.

*Skipped questions*: for categorical variables, skipped was treated as a separate category. This category was dropped in order to avoid collinearity and reduce dimensionality. For continuous variables, the value was set to 0.

*One Hot Encoding:* categorical features were one hot encoded

*Preserving variable descriptions:* by default, scikit learn’s one hot encoder returns a numpy array and therefore removes variable names*.* In the US Prisoners package, the variable names are preserved; in the features that are output from one hot encoding, each column will have the original feature name appended with a number. This allowed feature importances to be examined meaningfully

Original column name

One hot encoded column names

V0122

V0122-x1

V0122-x2

V0122-x3

Figure 3.3.3‑1 Input and output column names from the dataset\_processor class in the US Prisoners Package

*Scaling::*Continuous features were scaled using a min-max scaler.

Selected plots from the preliminary experiments are show in section 4.3. Results that meaningfully informed the variable selection and feature derivation process are discussed within the relevant sub-section in 4.2.

## Derived Features

This section explains the process that was undertaken to develop each derived variable. Appendix 5 provides more detail on other ways features included in the Inmates with Violent Crimes dataset could be derived, and suggests new variables to include. The suggestions in Appendix 5 could not be explored in this project due to time limitations.

### Sentence Length: Harsh or Not Harsh

See the ‘set\_sentence’ method on the dataset\_processor class for corresponding code.

As sentence length was to be the target variable, it was the first feature to be derived. Most bias assessment algorithms and debiasing techniques require a binary target, and so it was converted into a binary target of harsh or not harsh.

There are three pseudo-blocks where sentence is captured depending on if the prisoner had committed multiple offenses or a single offense. The routing between the three pseudo-blocks is illustrated in Appendices 2A and 2B. Within each pseudo-block . yet further distinctions are made about the type of sentence and how the length of the sentence is captured:

Range or Indeterminate Sentences

Some sentences are not for a set length of time, but have a minimum and maximum length of time. If this is the case, separate variables for minimum years, maximum years, minimum months, maximum months, minimum days, maximum days are completed in each pseudo-block. If the maximum is life, this is captured as 997; this should be considered when encoding sentence length.

Single or Flat Sentence

One variable in each pseudo block captures if the sentence is any of; Life, Life Plus Additional Years, Life without Parole, Death, Intermittent (weekends/nights).

For prisoners who had a single sentence with a specific length of time, a binary value indicating if the sentence length is above or below the threshold is set. Prisoners with flat sentences of Life, Life Plus Additional Years, Life without Parole, Death were put in the harsh class. Intermittent sentences were excluded; the amount of instances lost by excluding intermittent sentences was negligible and it simplified the binarization task at hand.

For sentences with a minimum and maximum range, the sentence was considered harsh if either the maximum or minimum amount of time was equal to or greater than the threshold

The sentence\_transform method of the data\_processor class in the US Prisoners package allows the user to specify the threshold for a ‘harsh’ sentence. Setting this to 0 will mean only Life, Life Plus Additional Years, Life without Parole, Death are considered harsh. It is also possible to configure if, for range sentences, the threshold should only apply to the minimum length of sentence (this would be a stricter definition of a harsh sentence).

25 years was found to provide a balance between the ‘harsh’ and ‘not harsh’ classes, and is a length of sentence considered harsh by sentencing campaign reform groups [32]. 25 is the threshold used when constructing the classifiers discussed in the rest of this paper.

#### *Preliminary Results 1 After Sentence Length Derivation*

Once the target variable of sentence length was derived , other variables were one hot encoded and scaled and the data passed to a Multi-Layer Perceptron and a Logistic Regression Classifier, without any feature selection in the pipeline. A grid search was used to measure accuracy across a broad range of hyperparameters. A low accuracy of 62% was achieved on the held-out test set. From this, it was determined that further features needed to be derived in order to improve the accuracy

### Current Offenses

The code to derive current offenses is in the set\_offenses method on the dataset\_processor class. This method calls the offense\_transform class, which contains most of the logic described below. The specific variables required to derive current offenses are contained in code comments.

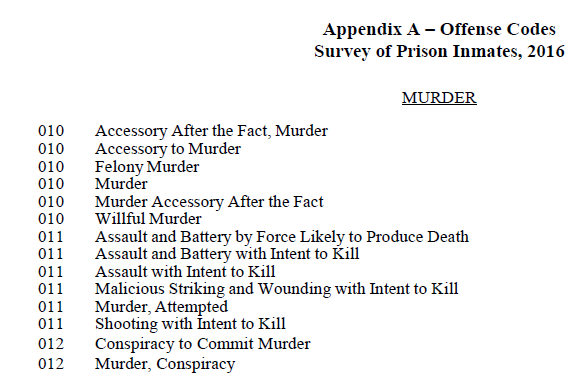
It was assumed that the offenses that the prisoner received their sentence for would be important for a classifier to predict sentence length, and that the sparseness of this data (held across 400 variables) would be weakening a signal to the classifier. In the Survey, there is the concept of current offenses and controlling offense. The current offenses are the group of offenses about which sentence information (from which the target variable is derived) is collected, and the controlling offense is the current offense (if there is more than one current offense) about which data is collected about in Victim, Weapon, Drug Crime and Property Crimes blocks.

Where participants were on parole, probation or escape before their current stay in prison, information is also collected about their previous offenses (that they were on parole or probation or escape from) as well as any offenses they are currently awaiting trial or sentencing or have been sentenced for (which may be parole or probation alone, or may be additional new crimes). This is visualized in the Inmate Type Determination and Offense Collection diagram (Appendix 2c); in each pseudo block (marked with a broken line) there are sets of offense collection questions (marked in a solid black line). As the routing arrows show, some participants will enter information about multiple sets of offenses.

The Questionnaire text, if read carefully, provides a mapping from Inmate Type to the set of offense variables that should be considered the current offenses that the prisoner is asked about (and from which the controlling offense is then established). The key factors that determine inmate type are if the prisoner is currently serving a sentence versus awaiting trial of sentencing and if they were on parole, probation or scape before being arrested. Inmate Type is stored in variable V0063. In broad terms, the current offenses are the offenses which the prisoner has most recently been sentenced for that are not court violations. For example, if a prisoner had escaped after being sentenced for two counts murder, the murder offenses listed in the ‘offenses before escape’ set of variables should be considered their current offense. The escape offense would not be considered the current offense.

Only inmate types 3,11 and 8 are included in the Inmates with Violent Crimes dataset. For inmate type 3, who had ‘No CJ Status’ (meaning they were not on parole, probation or escape at the time of arrest), the current offenses are those sentences which they have been sentenced for. For inmates type 8 and 11 (who were on parole or probation, respectively, when arrested), the current offenses are the ones they have most recently been sentenced for unless the new offenses are only parole or probation violation- in which case it would be their original offenses.

Up to five offenses can be captured in each offense collection pseudo block. Therefore, there are five variables to capture offense names (suppressed), five variables to capture offense codes and five variables to capture the corresponding offense types (Violent, Drug, Property, Public Order, Other), three variables to capture arrest date and three variables to capture admission date. In the Public Use dataset, offense name is suppressed in every set of offense variables. This means that it is not possible to establish the specific offense as one offense code corresponds to multiple specific offenses. Fig4-2-0 shows an extract of the Offense Codes Appendix in the Survey documentation; offense code 10 corresponds to multiple murder type offenses.



**Figure 4.2** An offense code corresponds to multiple specific offenses

Given that there are around 80 offense codes to be one hot encoded, leaving the current offense information in separate variables increased the dimensionality of the dataset. After completing the first round of preliminary experiments, with a top accuracy of 63%, it was decided to condense this information. The pseudo blocks marked with red triangles in the Inmate Type Determination and Offense Collection diagram indicate the pseudo sections that were collapsed as part of offense feature derivation.

Two approaches for densifying current offense variables are described in the sections below. In the final data processing code that has been provided, it is possible to choose which approach to use. The final classifiers described in section 7 use the ‘offense code count’ method.

Ordinal encoding would have been one way of dealing with the dimensionality created by one hot encoding offense codes. However, this was ruled out. Ordinal encoding would require a judgement about what offenses were worse than others. Offense seriousness is a multivariate issue that considers the details of the offense and number of victims, rather than the category of the offense; this is why sentencing guidelines exist [39]. Given that violent crimes such as rape, murder and bodily harm are contained within the dataset, it did not feel appropriate to make a judgement about how offense codes should be ranked.

#### One hot encoding

The first method was to collapse the three sets of offense variables (5 x offense codes, 5 x types) into a single set of five columns which could be one hot encoded. Fig 4-3 illustrates how one of the offense code columns was derived from the variables containing offense 1 (of up to five offenses) for each inmate type included in this experiment.

Location of first offense code for inmate type 11

Location of first offense code for inmate type 8

Location of first offense code for inmate type 3

off\_code\_1

V0124

V0280

V356

Derived variable

**Figure 4.2.2‑1** An illustration of how one of the off\_code\_n columns is derived. The above was repeated for all five offense code columns and offense type columns.

Using the same grid search method described above to exhaust all hyperparameters, this dataset with collapsed offenses (that were then one hot encoded) produced a MLP classifier with accuracy of 75% (see section 4.3.1).

There was a clear relationship between the number of features selected by the feature selector and the classifier accuracy. The highest ranking features according to both chi2 and mutual information were reviewed. Combined with feature importance information gathered from classifiers, this suggested that only some offense codes provided a signal to the classifier. This motivated trying a second approach to representing offenses. .

#### Offense codes count columns

The second method of representing current offense information was to have one column for each offense code populated with the count of how many times the inmate had listed this code in their current offenses. If a prisoner did not have this offense code listed, the count would be 0. This was then treated as a continuous variable and scaled- drastically reducing the size of the dataset. The same approach was applied to offense types.

Offense code frequency distribution was examined in order to determine which offense code columns to retain; the highest frequency offense codes are listed in Fig 4-4. The counts of all the columns with low frequency offenses were retained in another column so that this information was not lost. .

Dropping or retaining low frequency codes can be controlled through the set\_offenss method on the dataset\_processor class in the US Prisoners package.

|  |  |
| --- | --- |
| Code | Offenses under this code |
| 90 | Aggravated robbery, car jacking, heist |
| 10 | Murder, Manslaughter, |
| 120 | Aggravated assault, assault with a motor vehicle, domestic violence, mayhem |
| 70 | Aggravated sexual abuse, fondling, indecent assault, sex by deception, sexual misconduct |
| 480 | Weapons offense: carrying a firearm, selling a weapon |
| 11 | Attempted Murder, Assault with Intent to Kill |
| 40 | Abduction, kidnapping, false imprisonment |
| 190 | Burglary, House breaking, illegal entry |
| 50 | Aggravated Rape, Simple Rape, Attempted Rape, Rape of a Child with Force |
| 180 | Abortion, aiding suicide, child endangerment, illegal HIV transmission, tampering with a commercial product with intent to injure |

Figure 4.2.2‑2High frequency offense codes

### Controlling offense

The controlling offense determines which further sections of the survey the prisoner will answer, and is the offense about which specific offense data (such as victim demographics and injuries and weapons used) is collected. It was assumed that the controlling offense would be important to the classifier.

The survey instructions indicate that a controlling offense would have been entered into the survey software in order to route the surveyor and prisoner to the appropriate questions. However, controlling offense was not stored as a variable in some branches of the survey. The logic for determining Controlling Offense from the Current Offenses can be inferred by following the survey through different routes and reading the programmer instructions.[[7]](#footnote-7)

1. If the prisoner has one current offense listed, then this is always the controlling offense.[[8]](#footnote-8)
2. If the prisoner has more than one current offense, the survey asks prisoners to identify the offense(s) with the longest sentence. If one offense had a longer sentence than other offenses, this is the controlling offense.

If b identifies that multiple offenses received an equally long sentence (that was longer than the sentences for other offenses), or in any of the below scenarios, a multi-offense hierarchy was used to determine the controlling offense. In this hierarchy, the controlling offense is 1) A Violent Offense 2) A Property Offense 3) A Drug Offense 4) A Public Order Offense. There is no specification about how to determine the controlling offense if there are multiple offenses at one level in the hierarchy.[[9]](#footnote-9)

1. Following on from b; if two or more offenses, sentenced separately, received the same sentence length and the length is longer than any other sentences
2. Each offense was sentenced separately, but each sentence was the same length.
3. One sentenced was received for all sentences.

Once the current offense variables have been collapsed (as described in 5.2.1), the logic above was used to determine the controlling offense in many cases.

*For scenario a*, only one of the derived current offense variables will be populated with an offense code. This can be set as the controlling offense code.

*For scenario b*, the offense that the participant indicated was the longest offense will be captured in the controlling offense section. [[10]](#footnote-10)

*For scenarios c,d and e*, the hierarchy is used. The Violent Crimes subset only includes rows where the controlling offense is a violent offense. As violent is the first offense in the hierarchy, there will always be a violent offense as the controlling offense. Therefore, the below logic was used to determine the controlling offense in instances where the previous steps have not located it:

* If the count of violent offenses count is 1, we can infer that the one violent offense in the ‘current offenses’ columns derived in section 5.2.2 is the controlling offense.
* If there are multiple violent offenses, but all have the same offense code, then this offense code can be used as the violent offense

A challenge with this is that the ‘Lookup Table’ that maps an offense code to offense type referred to in the interviewer instructions in the Questionnaire is not provided in the Codebook. Therefore, assumptions had to be made using the lookup between offense names and offense codes provided in the documentation.

The controlling offense code and controlling offense type were treated as a categorical variables and one hot encoded. Where the controlling offense was not determined, this was treated as its own category. There were 786 instances (around 7% of the dataset) where a controlling offense could not be determined.

Preliminary investigations into feature importance showed that particular offense codes had a higher feature importance than others. As a result, low frequency offense codes in the controlling offense columns were replaced with a dummy value and the sum of low-frequency controlling offense columns was stored in a separate variable.

### Parole and Probation

Parole and probation information is held throughout many variables. In each sentence collection pseudo-block, the routing indicates if one of the most recent offenses was parole or probation- but the offense codes are not captured. This is indicated by the ‘P’ symbols in circles on the Inmate Type Determination and Offense Collection Diagram (Appendix 2b). As significant time had been spent deriving other variables, it was decided to take a simpler approach to incorporating parole and probation information into the dataset.

There are separate variables indicating the year an inmate was arrested depending on if the inmate was on parole or probation when they were arrested. These were kept in the dataset alongside the master arrest date variable.

The variables ‘parole’ and ‘probation’ were also created. These are set to ‘1’ if the prisoner falls into any of the prisoner types that were on parole or probation before their current arrest.

### Violent Type

Violent Type indicates if the offense was Murder, Rape or Other. Violent Type was included after preliminary investigations showed that an offense code of murder was important in separating between harsh and not harsh sentences. Similar to controlling offense, whilst Violent Type was entered into the programming software in order to route the interviewer to the correct questions, there are cases where it was not stored in the data. It can be inferred from survey routing; in blocks CJEA (Single Victim) and CJEB (Multiple Victims) in the Criminal Justice section, there is a question that routes the participants to pseudo-blocks that capture victim injuries. This is illustrated in the lower right section of the criminal justice section of the ‘Offense Collection and Inmate Determination’ diagram, indicated by marker triangles 3 and 4. The victim injury questions are phrased differently depending on if there was one or more victims. Which of these pseudo-blocks is populated can be used to determine the violent type.

### Offender race

The corresponding code is in the set\_offender\_race method on the dataset\_processor class

Offender race is a variable that we want to assess for bias in the final dataset and classifiers.

The offender is asked about their race, and is able to select multiple categories. This follows the pattern of a question being split into sub questions, with each sub question (in this case, each race category) stored as its own variable (as described in 3.2.2 Answer Encoding under the ‘Select Multiple’ heading).

For the purpose of defining a privileged and unprivileged class for fairness assessment, offenders who class themselves as white were determined to be the privileged class. This was binarised and renamed as ‘offender\_white’ whilst other offender race columns were dropped.

### Offender Sex

Offender sex is a variable that we want to assess for bias in the final dataset and classifiers.

Offenders are asked about their sex assigned at birth and their self-described sex identity in the physical health section. Offender\_male was set to true if the offender's self-described sex identify was male.

### Victim Injuries

The corresponding code is in the set\_victim\_injuries method on the dataset\_processor class

The Survey routes inmates whose controlling offense type is a violent offense to the ‘Violent Crimes and Victims’ section (CJE) . The same victim injuries information is collected the sections for a single victim and multiple victims (CJEA and CJEB)- this is illustrated at triangle markers 3 and 4 on the High Level Survey Routing diagram (Appendix 2A). As with offense information, this meant that the same information was stored in multiple places.

In initial experiments, victim injuries data was not collapsed. However, inspecting the feature importances indicated that some of the victim injury data was being used by the classifier, and so further work was done to densify this data.

Prisoners could select multiple injuries suffered by their victims. For each type of injury, a new binary feature was created and set to 1 if either of the corresponding variables in the multiple victim or single victim section had a positive value.

Table 4-2-8: Derived victim injury column name and source variables

|  |  |
| --- | --- |
| New feature name | Original Survey of Prison Inmates 2016 variables |
| broken\_bones | V0593','V0615','V0501','V0523' |
| bruises\_swelling | 'V0597','V0618','V0505','V0526' |
| gunshot\_bullet | 'V0592','V0614','V0500','V0522' |
| internal\_injuries | 'V0595','V0616','V0503','V0524' |
| knife\_stab | V0591','V0613','V0499','V0521' |
| other\_injuries | 'V0598','V0619','V0506','V0527' |
| victim\_died | 'V0589','V0497' |
| not\_physically\_injured | 'V0620','V0528' |
| rape\_sexual\_assault | 'V0590','V0498','V0510','V0602' |
| yeeth\_chipped | V0594','V0502' |

If a violent type of ‘Murder’ or ‘Rape’ has been indicated, then the victim injury pseudo section did not ask about injuries Rape and Victim Died (see the ‘Violent Type’ routing questions to the left of triangle markers 3 and 4 on the High Level Survey Routing diagram). Therefore, in addition to the victim injuries variables being collapsed, the presence of answers in the murder/manslaughter and rape survey branches was used to infer if the sexual assault and victim\_died features should be set to 1.

### Victim relationship

The Violent Crimes and Victims block in the Criminal Justice section collects information about the offender’s relationship to the victim. Similar to victim injuries, there are separate branches for single and multiple victims, but the same type of information is collected in both branches, and therefore it was deemed valuable to collapse this and see if it had any predictive importance.

In each single/multiple victim branch, the first group of questions establishes how well the inmate knew the victim. If the victim (or any of the victims) were known, it’s then established how well known. If the victims were well known, the specific relationship is established.

The below binary features were derived from the options available in the survey:

|  |  |
| --- | --- |
| New feature | Source variables |
| ‘victim\_spouse’ | V0493 Single Victim Relationship at time of crime is 1 (spouse), 2 (ex-spouse, 8 boyfriend/girlfeidn, 9 ex boyfriend/girlfriend  OR multiple victim spouse V0561=1, ex spouse V0562=7, ex boyfriend or girlfriend, V0569=9,V0568=8 boyfriend or girlfriend |
| ‘victim\_child’ | V0493 (single victim relationship) is 4 (own child) or 5 (step child).  OR V0564=4 (own child), V0565=2 (step child) |
| ‘other\_well\_known’ | V0494- 6 Brother/Sister/SteoBorhter Step Sister, 7 Other Relative, 10 Friend/Ex Friend, 11- Other |
| ‘victim\_stranger’ | V0491 (single victim known or not known)=2, V0557 Multiple victims)-2,3 some knonw/all strangers |
| ‘victim\_sight\_only’ | V0491 single victim, kn  V0492 (muitple victims, how well they knew) 1-Knew Well, 2-Casual Aq, Well Known 3=1 |
| 'victim\_casualacq' | V0492, multiple ictims ==2, V0559=1 |

‘Victim\_relationship’ did not seem to be an important feature when inspecting logistic regression and decision tree classifiers, and so no further work was done on these categories.

### Victim below 12

The corresponding code is in the set-victim\_age method on the dataset\_processor class

The Violent Crimes and Victims block asks inmates about the age range of their victim(s). Inmates with multiple victims are asked about the age range of the youngest victim (stored in variable V0555) and the age range of the oldest victim, and inmates with a single victim are asked for the age range of that victim (variable V0490).

Victim age information was collapsed into the column ‘victim\_below\_12’. If either the youngest of multiple victims was below 12 or if the victim (in the case of a single victim) was below 12.

Note that ‘victim\_below\_12’ is a different concept to ‘victim\_relationship\_child’. The latter indicates if the victim was a child, step-child or adopted child of the offender, but does not reflect their age. For example, if a stole jewelry from a child aged 40, this would be captured as 'victim\_relationship\_child', but ‘victim\_below\_12’ would be 0.

### Victim Race

The corresponding code is in the set\_victim\_race method on the dataset\_processor class

Victim race is a variable that we want to assess for bias in the final dataset and classifiers trained on it.

For single victims, the prisoner was asked what race they thought victim was, and they were able to select multiple categories; this is mapped to variables using the same ‘Select Multiple’ approach described in the Answer Encoding section (3.2.2) [[11]](#footnote-11) For multiple victims, the prisoner was asked both what race most of the victims were and, for each race category, if any of the victims were of that race.

***Victim Race Questions asked to offenders with multiple victims****:*

*CJEB3. Were any of the persons of Spanish, Latino, or Hispanic origin?*

*CJEB4. Were all the persons Hispanic, were most of them Hispanic, were equal numbers of them Hispanic and non-Hispanic, or were*

*most of them non-Hispanic?*

*CJEB5. What race or races were the persons? You may answer yes to one or more of these categories. (CHECK ALL THAT APPLY)*

*CJEB6. What race were most of the persons?*

For the purpose of producing a dataset that can be tested for bias, the victim race columns were collapsed into a single feature indicating if the victim was white. This was set to 1 if the single victim was white or if the victims were mainly white for multiple victims. This means that, if some of the victims were white but not all of them, the ‘victim white’ feature would be set to 0

### Victim Sex

The corresponding code is in the set\_victim\_sex method on the data\_processor class

Victim Sex is a variable that we would like to assess for bias. As with other victim variables, victim sex would not have been part of the sampling strategy.

The single victim block asked *‘Was the person male or female?[[12]](#footnote-12)* The multiple victim block asked *‘Were most of the persons male or female?,.[[13]](#footnote-13)* If the inmate answers ‘Both Male and Female’, they are then asked which sex the majority of victims are, or if the split was evenly divided.

A new column ‘victim\_male’ was created as a binary feature. This is set to ‘1’ , to indicate the privileged male class if:

* the single victim was male (‘V0489’=1)
* the multiple victims were all male (‘V0553’=1)
* most victims were male or the victims’ sex was evenly divided (V0554=1 or 3).

## Preliminary Results Used to Inform Feature Selection

### Preliminary Results 1- after collapsing and one hot encoding offenses

In this round of experiments, a pipeline was created that used k-best feature selection before passing the data to a logistic regression classifier and then a multi-layer perceptron. A grid search was used to review various combinations of logistic regression penalties and solvers for values of k up to 400. For both MLP and logistic regression, the best accuracy score was 75% with 300 features.

It appeared that variation in the results was being caused by the k-means scorer function. Plotting results from the scorer functions separately showed that chi2 produced better results than mutual information regardless of other hyperparameters.

Different values of alpha did not have a significant impact on accuracy on the MLP classifier; the best results were achieved with the adam solver regardless of other hyperparameters. best results came with the best 300 features selected accorded to chi2

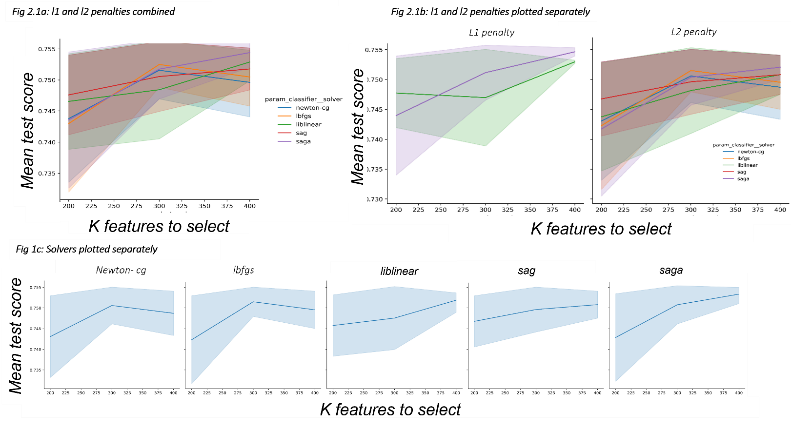


Figure 4.3.1‑1 A: Results for an MLP model across all solvers and activation functions. The top score achieved was 75%. Issues with saving the document have led to low image fidelity. Both scorer functions

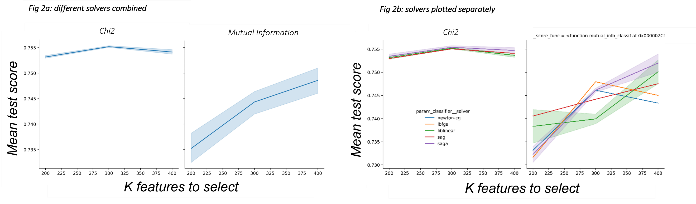


Figure 4.3‑1 B: When separating the grid search results by the scorer function used to select the k best features, it shows that chi2 produces significantly better results

### Preliminary Results 2-with controlling offense information

The following results were produced once controlling offense information had been derived. Varying the decision tree criterion did not have a significant impact on results. The most important hyperparameter was the score function used to select the top k features, with mutual information outperforming chi2.

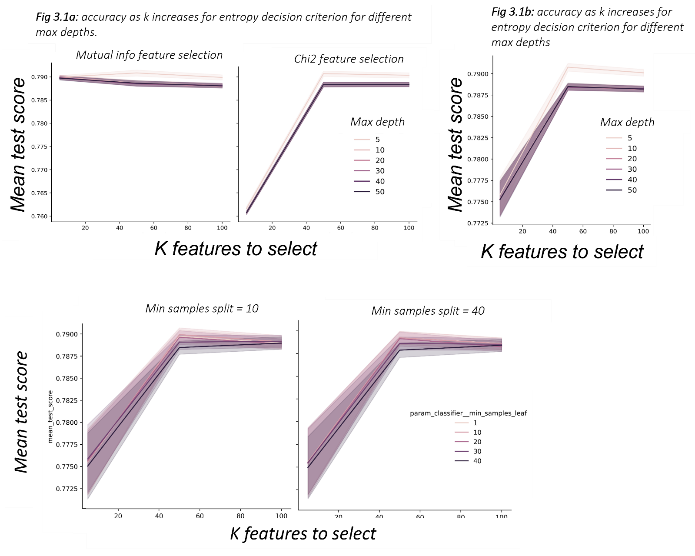


Figure 4.3.2‑1 Decision Tree

#### Preliminary results Round 3

After offenses were changed to variable counts and other features had been derived, both logistic regression and decision tree classifiers were able to achieve an accuracy of 80%. At this point, work was stopped on feature derivation and a more detailed inspection of bias in the dataset and models commenced.



Figure 4.3.2‑2 Results from Logistic Regression for increasing values of k. The penalty used did not impact accuracy; the value of k was the controlling factor.

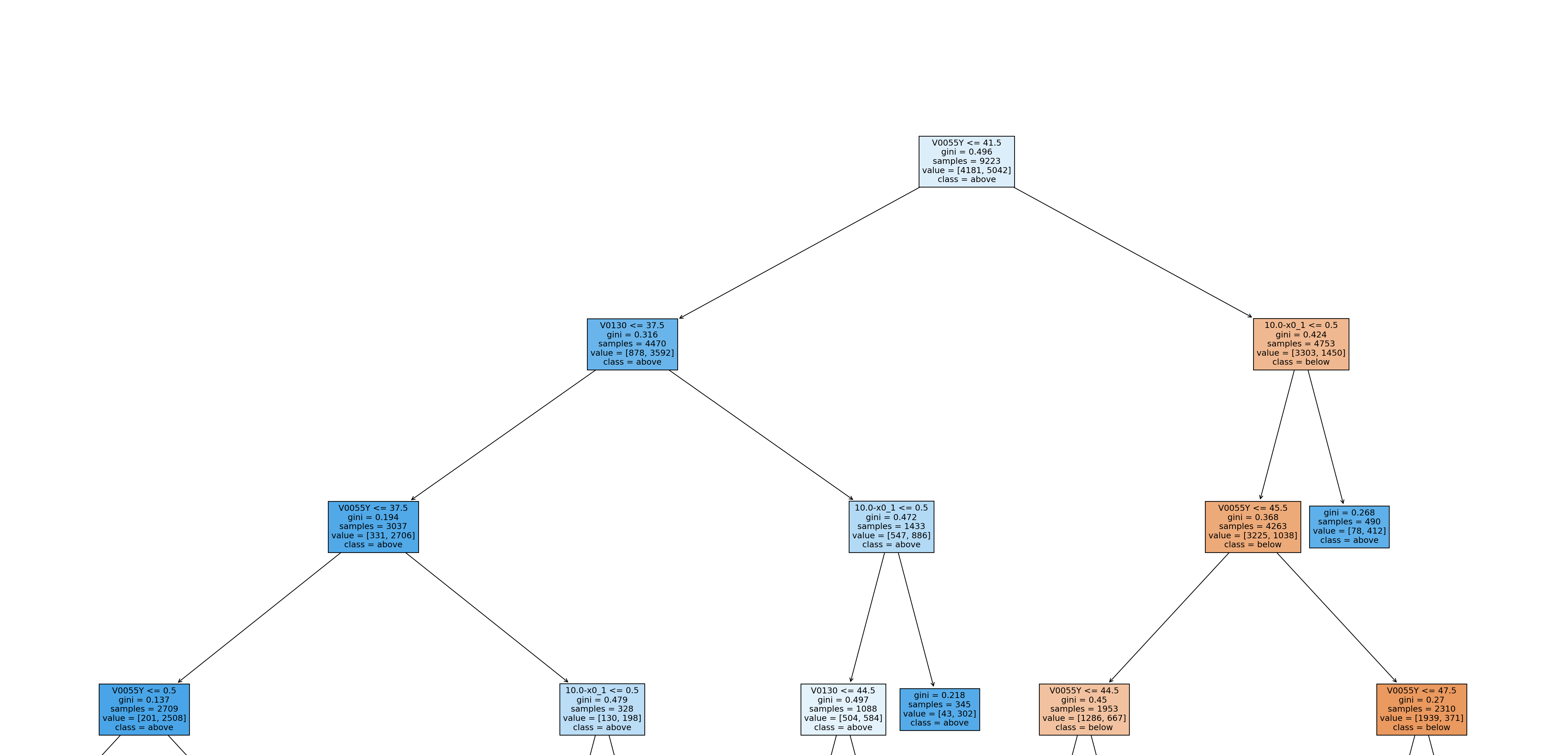


Figure 4.3.2‑3 A decision tree classifier produced after controlling offense had been derived and current offenses were stored as offense counts (for each offense code) rather than one hot encoding..

|  |  |  |
| --- | --- | --- |
| Variable Name | Description | Score (top 5 features) |
| V0055Y | Arrest year | 0.60 |
| 10.0x\_1 | Offense code 10 (various murder offenses) count | 0.26 |
| V0130 | Year arrested (for inmates not on parole, escape or probation at time of arrest) | 0.05 |
| V0362 | Year arrested (for probation violators) | 0.03 |
| V0069\_x\_3.0 | Violent type other (not murder or rape) | 0.015 |

# Description of final dataset

A dense version of the Inmates with Violent Crimes dataset (without encoding and scaling applied to the non-derived values) has 133 columns (132 features and the target label). The notebook demonstrates which parameters to pass to the dataset processor in order to produce a dense version of the data, and how to update the variable\_config dictionary to add and remove variables. After one hot encoding has been applied, there are 282 features and one target column. 190 columns have a variance below 5%. Appendix 4 shows the top features ranked by mutual information and chi2 metrics. Note that there are 188 variables with a mutual information score of 0.

The binarised sensitive attribute columns in the encoded version are**:**

|  |
| --- |
| offender\_white-x0\_1 |
| offender\_male-x0\_1.0 |
| victim\_white-x0\_1 |
| victim\_male-x0\_1.0 |

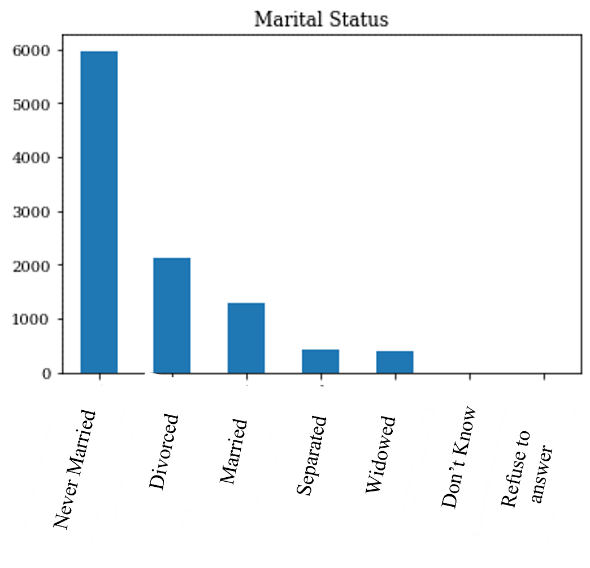
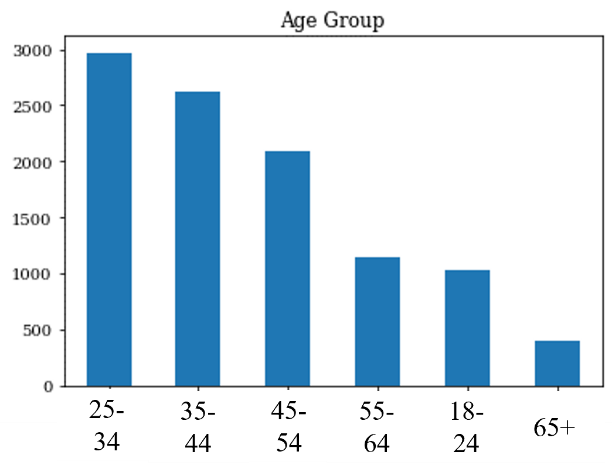
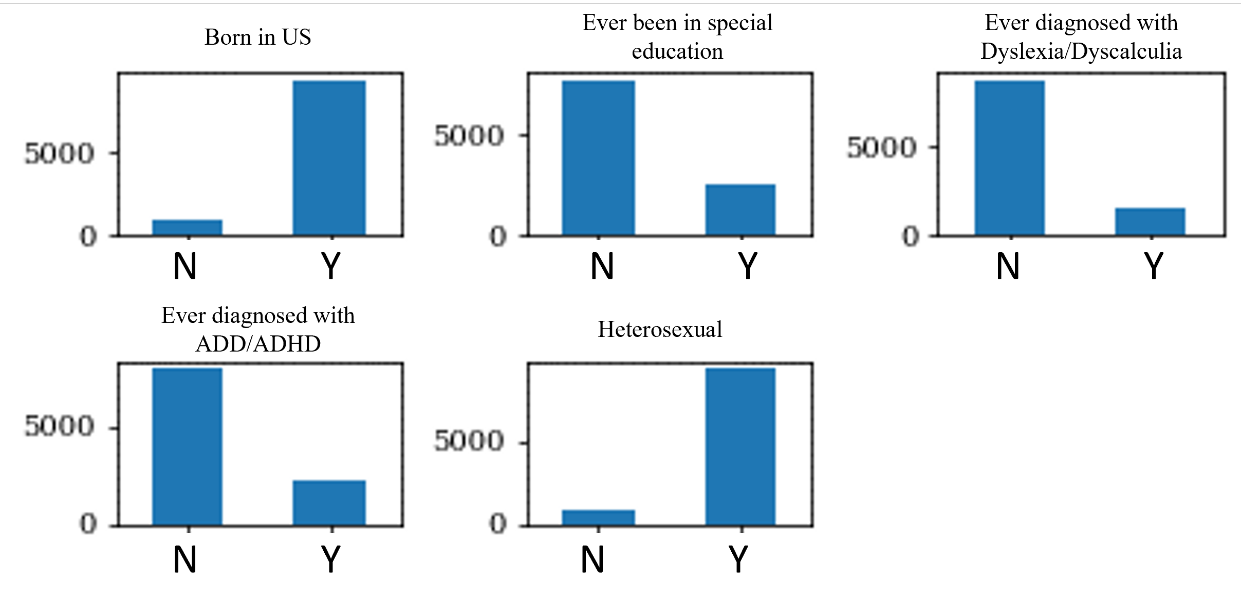
Other potential sensitive attributes are:

|  |  |
| --- | --- |
| RV0002 | offender age category |
| V1213-x0\_1.0 | Lesbian or Gay |
| V1213-x0\_2.0 | Heterosexual |
| V1213-x0\_3.0 | Bisexual |
| V1213-x0\_4.0 | Sexual orientation something other than lesbian, gay, heterosexual or bisexual |
| V1213-x0\_5.0 | Doesn’t know sexual orientation |
| V1208, V1209 | Height (Feet, Inches) |
| V1210 | Weight |
| V0022-x0\_2 | Widowed |
| V0022-x0\_3 | Divorced |
| V0022-x0\_4 | Separated |
| V0022-x0\_5 | Never Married |
| V0942-x0\_1.0 | Diagnosed with ADD |
| V0943-x0\_1.0 | Diagnosed with dyslexia or dyscalculia |
| V0945-x0\_1.0 | Born in the US |
| V0951 | How many years lived in US |

The figures in section 5.1 show distribution of the binarised sensitive features variables and the other potential sensitive attributes.

The target label, offender race, victim race and victim sex are well balanced. Born in the US, Special Education classes, Dyslexia/Dyscalculia, ADD/ADHD and Heterosexual are all imbalanced, with the privileged class being more represented. The exception is offender sex, where there is a much higher number of men (the historically privileged class).

## Distribution of Target and Sensitive Labels:



# Dataset Bias

As discussed, there are a variety of ways to measure bias in data and machine leaning models. Three intuitive and commonly used metrics are:

*Statistical parity difference (SP)* requires that the same proportion of each group receive a positive outcome. A negative *SP* means bias to the unprivileged group. A bigger difference shows more bias

*Disparate impact (DI) DI* divides the proportion of favourable outcomes for the unprivileged group by that of the privileged. If there is no disparate impact, *DI* will be 1. A number below 1 shows bias to the unprivileged group. A number higher than 1 shows bias to the privileged group.

*Equal opportunity difference (EO)* is useful in evaluating machine learning predictions. Equality of opportunity is satisfied if both groups are equally as likely to receive a false negative. *EO* is the difference in the proportion of false positives between the groups. A negative *EO* shows bias is to the unprivileged group. A bigger absolute value shows more bias.

The author uses similar descriptions of these equality metrics in [40]

The aif360 toolkit was used to calculate SI and DI for the Inmates with Violent Crimes dataset (this can be reproduced in the notebook).

**In the case of prisoners, the favourable outcome is a less harsh sentence. In the case of victims, a harsh sentence was considered favourable.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | SP | DI |  |
| Offender race | -0.04 | 0.93 | Bias in favour of the privileged class |
| Offender sex | 0.12 | 1.24 | Bias in favor of the historically unprivileged class |
| Victim race | 0.01 | 1.02 | No bias |
| Victim sex | -0.05 | 0.89 | Bias in favour of the privileged class |

Figure 4.3.2‑1Fairness metrics for the processed dataset

With regards to offenders, there is a bias against non-white offenders (which is expected given the research in [15] and [30]), but there is a bias in favor of the unprivileged (female) sex class.

With regards to victim race, the scores are very close to ‘perfect’ fairness. However, there is a bias in favour of the historically privileged class (male). This suggests that male victims are more likely to have their perpetrator receive a harsher sentence.

# Model Search and Assessment

Deriving the features reduced the size of the dataset, and it was then possible to use auto\_sklearn to search for a model. The ‘vanilla’ version of auto\_sklearn was used, meaning that the search was for a single model rather than an ensemble. This also reduced the computational complexity of the task. The notebook demonstrates how to install auto\_sklearn into Google Colab (which is useful as auto\_sklearn does not run on Windows) and set the appropriate hyperparameters. The auto\_sklearn toolkit found a Linear Discriminant Analysis (LDA) model and a Random Forest Classifier that both achieved 81% accuracy. Sklearn GridSearch also found models that achieved 81% accuracy using Logistic Regression and Multi Layer Perceptron. 80% accuracy was achieved with a simple Decision Tree.

Decisions trees can be easily inspected and explained to humans. [8]Given the importance of explainability as an ethical AI principles in [41] and the fact that explainable models are easier to check for fairness [42], it seems prudent to proceed with analysing the decision tree models, which are easier to both inspect and explain than LDA, RF and MLP models.

## Decision Tree GridSearch

In order to find the best decision tree parameters, the same grid search method with fivefold cross validation as used in the preliminary experiments was applied. Between 0 and 250 features were selected using both mutual information and chi2. These were evaluated with max depths of 5,10 and 20; max leaf nodes of 10, 20 and 30, gini and entropy decision tree criterion.

All parameter combinations classifiers achieved at least 80.005% accuracy on the held out test set (not used in the Grid Search). The best score was 80.040%. The most accurate decision tree used only 5of the 282 encoded and scaled features, with a loss of just 0.01 between training and test sets.

When reviewing the results of the grid search (section 8.1), it is clear that the optimal value of k is different depending on whether features are selected using chi2 or mutual information scores. Figure 8.1.1 shows accuracy as more features are selected by the k-best feature selector. The plots in the left hand column show accuracy k when mutual information is used to select features, and the right hand column shows the results when chi2 is used to select features. Whilst max leaf nodes and decision criterion both impact accuracy, the value of k is the most significant factor – which is consistent with preliminary experiments. When mutual information is used to select features, accuracy decreases as soon as more than 5 features are selected. When chi2 is used, accuracy peaks at both 5 and 50 features, and the decline in accuracy after 50 features is slower compared to mutual information. Instead of inspecting only the most accurate mutual information classifier for bias, the most accurate chi2 classifier was also inspected.

Tables scoring each variable by mutual information and chi2 were produced in gain more insight into what features were driving high accuracy with low values of k, and what caused the step change at k=50 in the chi2 classifiers. These are in Appendix 4.

The variables: year of arrest; inmate type 8 year of arrest; count of murder offenses and if the controlling offense was murder were the top four ranking variables for both scores. The fifth variables were if the victim died (for mutual information) and the age the offender was the first time they were admitted to prison (chi2). The victim age, whether a weapon was carried with the intent to kill and if the controlling offense was aggravated assault also appeared in the top 10 variables for both scores. Most of these seem like reasonable features to inform sentence length.

Plots for both the most accurate classifier and the most accurate chi2 classifier were produced so that the logic of both trees could be inspected- the parameters are given in Table 7-1.

|  |  |  |
| --- | --- | --- |
|  | Mutual info | Chi2 |
| K features | 5 | 150 |
| Max depth | 5 | 5 |
| Max leaf nodes | 10 | 30 |
| Min samples per leaf | 10 | 10 |
| Min samples split | 10 | 10 |
| Criterion | Entropy | Entropy |

Table 7‑1 Parameters for the most accurate decision tree classifiers using mutual info and chi2 feature selection.

## *Decision Tree Grid Search Results*

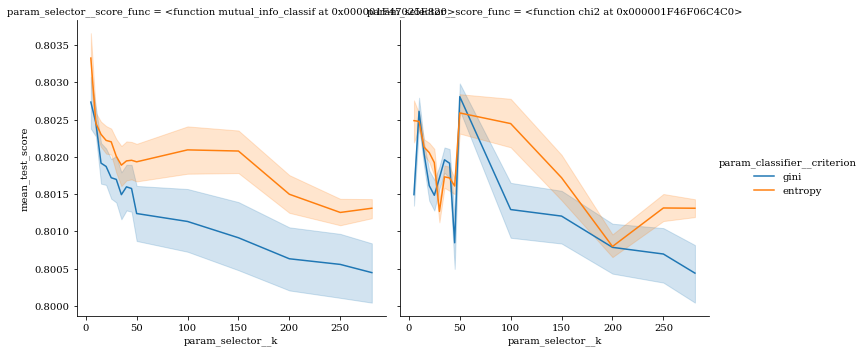
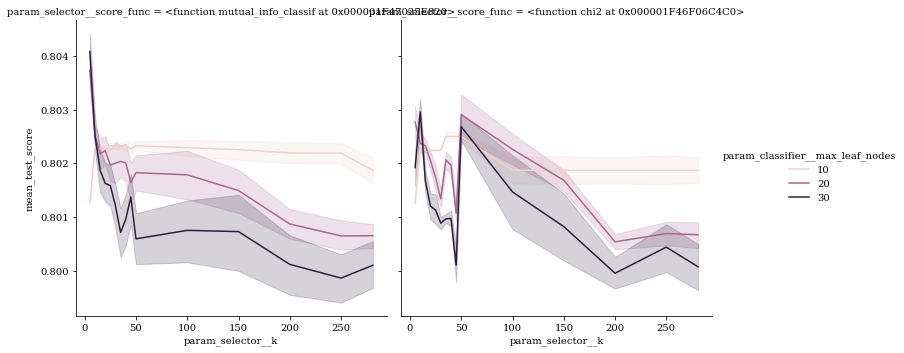


Figure 8.1‑1 Results from the decision tree grid search. The left column shows the results when features were selected using mutual information and the right chi2.

Figure 8.1‑2 Results from a decision tree parame

## Decision Tree Inspection

As the US Prisoners package retains feature names after encoding, it was possible to inspect the decision tree structures.

Although the chi2 k was set to 150, the classifier only used 7 features: count of murder offenses, year arrested, year arrest (inmate type 8 only), current age 18-24, if the offender lived in the US State NC at the time of arrest, count of offense code 90 (mugging) and if the inmate was divorced. Divorced is one of the variables that ranks between 45 and 55 in the chi2 table, so this variable is potentially the reason for the jump in accuracy when 50 features are selected. Given that divorce is marital status, it could be considered a sensitive attribute

Consistent with the preliminary experiments, the variables relating to year of arrest and if the controlling offense was murder were of high importance in both trees. In both trees, the year of arrest is referenced at

multiple splits- they are marked with a red triangle (year of arrest) and black triangle (inmate type 8 year of arrest) on the decision tree diagrams in Fig 7-2-1-1 and 7-2-1-2..

When decision tree plots were generated from sklearn, the years values at each leaf were the ordinally encoded values. For ease of inspection, these were cross referenced and the diagrams updated actual year. This reveals that ‘Arrested after 2007’ appears as a decision point multiple times in the chi2 decision tree, and once (in the second layer) of the mutual information decision tree. 2004 and 2003 is used twice in both decision trees. 2013 appears once in both decision trees. The significance of these years is not apparent from inspecting the Survey or the data alone, but would require an understanding of Criminal Justice policy in the United States. Similarly, the significance of the US State NC (assumed to be North Caroline) is not apparent from the data. This supports the point made by Zilka et al in [43], Criminal Justice data is complex and its importance to consider contextual factors when using it in machine learning.

### Final Decision Tree Plots

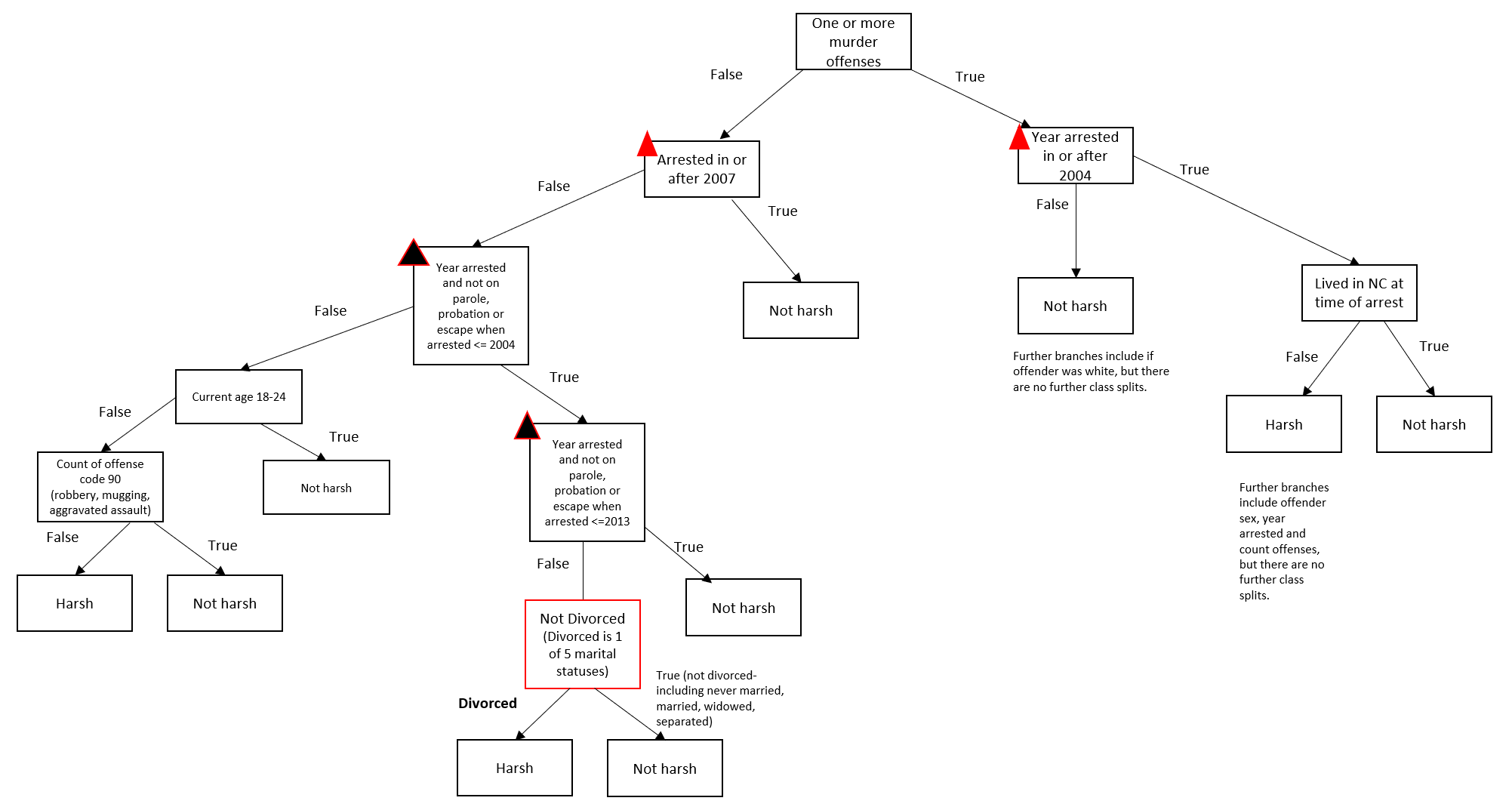


Figure 7.2.1‑1 Decision Tree plot for best chi2 Decision Tree. Due to issues inserting the plot generated by sklearn, it was reproduced in Microsoft Office.

Diagram

Description automatically generated

Figure 7.2.1‑2 Decision Tree Plot for best mutual info decision tree

## Classifier Accuracy and Bias

The confusion matrices for both classes showed that the F1 was poorer the positive class, which was caused by a c70% recall of the positive class. In high stakes decision where the favourable label is the positive class, it is preferable for accuracy to suffer as a result of a false positive- i.e., it is better for the classifier to predict a less harsh sentence than a harsh sentence. However, if we consider this with regards to victim justice, then poorer recall for the positive class is less fair.

The classifiers did amplify bias for some groups. The figures in 7.4 show the EO, DI and SP fairness metrics for each protected attribute.

For offender race, there is a small increase in disparate impact and statistical parity scores (disadvantaging the less privileged class).

In the original dataset, there was bias towards male offenders. The model predictions show less bias is for both DI and SP.

The most bias amplification is shown for Victim Sex. Disparate Impact and Statistical Parity scores both indicate more bias towards female victims (meaning their perpetrators are less likely to receive a harsh sentence) than the original dataset. In addition, the EO score for the predictions indicates inequality of opportunity.

## Accuracy and Confusion

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Best mutual info** | | | | |
|  | Prec | Rec | F1 | Acc | Loss |
| 0 | 0.78 | 0.90 | 0.84 | 80.85 | 0.15 |
| 1 | 0.86 | 0.70 | 0.77 |

Figure 7.2.1‑1 Results for the decision tree classifier trained using the top 5 features by mutual information scores. A high accuracy of 80.5% was achieved with a small loss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Best chi2** | | | | |
|  | Prec | Rec | F1 | Acc | Loss |
| 0 | 0.79 | 0.90 | 0.84 | 81.23% | +0.01 |
| 1 | 0.85 | 0.71 | 0.78 |

Figure 7.2.1‑2 Results for the decision tree classifier trained using the top 150 features by chi2 score. Accuracy increased by 1% between train and test sets.

## Fairness Metrics

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Offender Race** | | |
|  | EO | DI | SP |
| Original | / | 0.93 | -0.05 |
| Mutual info |  | -0.91 | -0.05 |
| Chi2 |  | -0.93 | -0.04 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Offender Sex** | | |
|  | EO | DI | SP |
| Original | / | 1.24 | 0.12 |
| Mutual info | -0.02 | 1.06 | 0.04 |
| Chi2 | -0.03 | 1.06 | 0.04 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Victim Race** | | |
|  | EO | DI | SP |
| Original | / | 1.02 | 0.01 |
| Mutual info | -0.04 | 1.06 | -0.03 |
| Chi2 | -0.050 | 1.03 | -0.03 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Victim Sex** | | |
|  | EO | DI | SP |
| Original | / | 0.89 | -0.05 |
| Mutual info | **-0.15** | 0.75 | -0.12 |
| Chi2 | **-0.13** | 0.76 | -0.10 |

# Discussion

The Inmates with Violent Crimes dataset provides machine learning researchers with new criminal justice data to work with. This paper outlines a brief investigation only one target variable and four sensitive variables in the dataset, but many more targets and sensitive attributes could be created. The explanations, diagrams and code provided as part of this project will be available to researchers wishing to work with this data.

Much research has shown that simple models are often as good as black box models [43]. Rudin urged practitioners to ‘Stop Explaining Black Box Models for High Stakes Decisions and Use Interpretable Models Instead’ in a paper of the same name. [44]. Whilst this project did not set out to demonstrate this point, the findings have undoubtedly done so.

This dataset also demonstrates the importance of datasheets, in particular providing information about protected groups in the dataset and how they are represented. The models trained on this dataset did amplify bias in some areas, and used a potentially sensitive attribute (marital status) to make a high-stakes decision. It’s important that anyone using the data should be aware that this is possible. The presence of data about secondary data subjects (the offender’s victim(s) and child(ren)) is not obvious upon initial reading of the Survey documentation, but the datasheet allows it to be highlighted.

# Limitations

All of the information in the Survey was self-reported by offenders; as such, any conclusions drawn from the data or models trained on it may not reflect ground truth.

The bias assessment of the dataset and models is contingent on not only how the target label of ‘harsh’ or ‘not harsh’ is derived, but also how the race and sex variables were determined. There was particular ambiguity introduced by the fact offenders could indicate multiple races they identified with, and the fact that many offenses had multiple victims. There are many other ways that the target label, race feature and sex feature could be derived.

Victim age was not considered as a sensitive attribute in this project. The rationale was that there are different offense codes for crimes with child victims, and it follows that there may be a difference in sentence length. However, bias that wouldn’t be acceptable is if the age of a victim above 18 influenced the sentence the perpetrator received. It became apparent late in the project that it may be possible to determine an age range by calculating the inmate’s current age using V0012 (calculated age based on year of birth) and V0013 (the age given by the prisoner if they believed the calculated age to be incorrect) and cross referencing this with year of arrest.

The survey contains information about how long the offender was in prison (awaiting sentencing or trial) before they received their sentence, and how much of this time was applied to their sentence. These variables were not used on the assumption that the sentence length provided into the survey is the sentence prior to any time deduced. The survey instructions to do not state this explicitly however, so this should be validated.

The Appendix provided in the US Prisoners Codebook does not explicitly state which offenses are classed as violent offenses, but the author made assumptions based on the title of each offense group. The assumed mapping between offense codes and offense types can be seen in set\_violent\_type method on the offense\_transform class.

The classifiers rely heavily on the year the prisoner was arrested in. This doesn't reflect the human reasoning that would have been applied at the time of sentencing (i.e., sentencing guidelines). It would be preferable to retain a fair accuracy without the classifier having this information. Investigation as to. The derived variables section provides numerous ways in which the way variables are derived could be adjusted. Implementing these different approaches and re-assessing classifier accuracy, dataset bias and classifier bias would be the next logical step in this work. As discussed, it is not possible to determine the specific offenses a prisoner has been sentenced for. This may prevent the classifier from achieving an accuracy higher than 80%.

# Further Work

## Sensitive Features

Section 6 lists further variables in the dataset that could be considered sensitive features. The dataset and decision tree model predictions should be assessed for bias with regards to these features, and the datasheet updated. The possibility of deriving victim age was identified after significant progress had been made in the project. This variable should also be derived and assessed for bias. Intersectional groups could also be identified from the dataset and the data and models assessed for bias with regards to these groups. Examples include non-white female offenders or female victims who were spouses or ex-spouses of the offender.

The high capability models that were found (MLP, LDA, RandomForest) should also be assessed for biased predictions and the datasheet updated.

## Target Variables

The Inmates with Violent Crimes data contains lots of other information about the sentence in addition to the custodial sentence length; whether or not sentence includes court fees, fines, community service, requirements for drug or alcohol treatment, conditions such as listing on the sexual offenders register. These sentence attributes could also be predicted. It may be possible to create a multi-output classification predicting the sentence length and other sentence attributes.

The survey also asks prisoners about things they do in jail, including if they have received visits. Some consideration would need to be given to how long the prisoner has been in prison for, but this could also be a potential target variable. Potential signalling information may be found in sections about family, children, who they were living with and where they were living before arrest.

V1010-V1039 contain information about where children that were living with participant at the time of arrest are living now, which could be consolidated into fewer features and used as a target label.

V1040-V1052 contain information about communication, visits and frequency of them from children since prison and reasons for a lack of in person visits. It’s possible to envisage this kind of data being used to predict if a prisoner is likely to get visits and communication with children.

# References (885 words)

|  |  |
| --- | --- |
| [1] | United States. Bureau of Justice Statistics, *Survey of Prison Inmates, United States, 2016,* Inter-university Consortium for Political and Social Research, 2021. |
| [2] | D. Leslie, *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector.,* 2019: The Alan Turing Institute. |
| [3] | A. Desrosières, *The Politics of Large Numbers: A History of Statistical Reasoning,* Harvard University PresHarvard University Pres, 1998. |
| [4] | D. Bouk, *How Our Days Became Numbered: Risk and the Rise of the Statistical Individual,* University of Chicago Press, 2105. |
| [5] | S. Barocas, M. Hardt and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, fairmlbook.org, 2019. |
| [6] | T. Baer, *Understand, Manage, and Prevent Algorithmic Bias,* Apress, 2019. |
| [7] | E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, AlaniHarith, B. Berendt, T. Kruegel, C. Heinze and S. Staab, "Bias in data‐driven artificial intelligence systems—An introductory survey," *WIREs Data Mining and Knowledge Discovery,* vol. 05, 2020. |
| [8] | European Commission and Directorate-General for Communications Networks, Content and Technology, *Ethics guidelines for trustworthy AI,* Publications Office, 2019. |
| [9] | Ministry of Local Government and Modernisation (Norway), *The National Strategy for Artificial Intelligence,* 2020. |
| [10] | B. Sterling, "The Beijing Artificial Intelligence Principles," Wired (Conde Nast Interational), 01 June 2019. [Online]. Available: https://www.wired.com/beyond-the-beyond/2019/06/beijing-artificial-intelligence-principles/. |
| [11] | R. T. Vought, "MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES: Guidance for Regulation of Artificial Intelligence Applications," [Online]. Available: https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf. |
| [12] | T. Mahoney, K. Varshney and M. Hindi, AI Fairness, O'Reilly Media, Incorporated, 2020. |
| [13] | A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," *arXiv preprint arXiv:1810.08810,* 2018. |
| [14] | J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018. |
| [15] | W. D. Heaven, "Predictive policing algorithms are racist. They need to be dismantled.," MIT Technology Review, [Online]. Available: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/. |
| [16] | L. Dixon, J. Li, J. Sorenson, T. Nithum and L. Vasserman, "Measuring and Mitigating Unintended Bias in Text Classification," in *AAAI/ACM Conference on AI, Ethics, and Society*, 2018. |
| [17] | S. Sharma, Z. Yunfeng, J. M. Rios Aliaga, D. Bouneffouf, V. Muthusamy and K. R. Varshney, "Data Augmentation for Discrimination Prevention and Bias Disambiguation," in *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society New York NY USA February 7 - 9, 2020*, 2020. |
| [18] | F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discriminatio," *Knowledge and Information Systems,* vol. 33, no. 1, p. 1–33, 2012. |
| [19] | M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger and S. Venkatasubramanian, "Certifying and Removing Disparate Impact," New York, NY, USA, 2015. |
| [20] | A. Narayanan, "21 fairness definition and their politics," in *ACM FAT\* (Fairness, Accountability and Transparency) Conference in 2018*, 2019. |
| [21] | J. Kleinberg, S. Mullainathan and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *CoRR,* vol. abs/1609.05807, 09 2016. |
| [22] | S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton and D. Roth, "A comparative study of fairness-enhancing interventions in machine," in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT 2019, Atlanta, GA, USA, January 29-31, 2019*, 2019. |
| [23] | "2021. ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT). https://facctconference.org/," [Online]. |
| [24] | E. M. Bender and B. Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," *Transactions of the Association for Computational Linguistics,* vol. 6, pp. 587-604, 2018. |
| [25] | T. Gebru, J. Moregenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daume III and K. Crawford, Datasheets for Datasets, arXiv, 2018. |
| [26] | K. L. Boyd, "Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data," in *Proceedings of the ACM on Human-Computer Interaction Issue CSCW2*, 2021. |
| [27] | A. Paullada, D. R. Inioluwa, E. Bender, E. Denton and A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," *Patterns,* vol. 2, no. 11, 2021. |
| [28] | https://proceedings.neurips.cc/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf. |
| [29] | K. Wagstaf, *Machine Learning that Matters,* 2012. |
| [30] | J. Larson, S. Mattu, L. Kirchner and J. Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," ProPublica, 23 May 2016. [Online]. Available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. |
| [31] | M. Zilka, B. Butcher and A. Weller, "A Survey and Datasheet Repository of Publicly Available US Criminal Justice Datasets," *https://openreview.net/.* |
| [32] | M. Mauer, "Long-Term Sentences: Time to Reconsider the Scale of Punishment," The Sentencing Project, 6 November 2018. [Online]. Available: https://www.sentencingproject.org/publications/long-term-sentences-time-reconsider-scale-punishment/#:~:text=A%20key%20issue%20in%20assessing,(2011).. |
| [33] | S. B. Durcan, "https://github.com/sdurcan/us\_prisoners," [Online]. |
| [34] | *Survey of US Prison Inmates 2016, 37692-0001-Codebook'.* |
| [35] | https://pandas-profiling.ydata.ai/docs/master/index.html. |
| [36] | I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research,* vol. 3, pp. 1157-1182, 3 1 2003. |
| [37] | M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum and F. Hutter, "Efficient and Robust Automated Machine Learning," *Advances in Neural Information Processing Systems 28 (2015),* pp. 2962--2970, 2015. |
| [38] | M. Kuhn and K. Johnson, Applied Predictive Modeling, New York, NY: Springer New York. |
| [39] | United States Sentencing Commission, 2018. |
| [40] | S. B. Durcan, *ML\_A3\_202\_Report\_Student\_229144,* None: part of coursework submission, 2022. |
| [41] | European Parliament, *EU guidelines on ethics in artificial intelligence,* 2019. |
| [42] | F. D.-V. a. B. Kim, *Towards A Rigorous Science of Interpretable Machine Learning,* arXiv, 2017. |
| [43] | C. Wang, B. Han, B. Patel and C. Rudin, "In Pursuit of Interpretable, Fair and Accurate Machine Learning for," *CoRR,* vol. abs/2005.04176, 2020. |
| [44] | C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *arxiv.* |
| [45] | *Survey of Prison Inmates, United States, 2016 Codebook for Public-Use U.S. Data (state and federal combined).* |

# Appendix 1: Survey Contents (990 words)

Table 3 Content of each survey section and sub-blocks.

|  |  |  |
| --- | --- | --- |
| **Section Name** | | **Section Code** |
| **Interview Setup** | | **IS** |
| Inmate consent  If an incentive is being offered for survey participation  OMB Approval information  Facility information (state/federal, gender) | | |
| **Demographics** | | **DEMO** |
| Age  Race  Marital status  Military service status | *A-Military Service Block*  Branches of military served in  Type of discharge | |
| **Criminal Justice** | | **CJ** |
| *Q1-10, CJA-Parole Violators Block, CJB-Probation Violators Block*  Reason in prison  Offenses currently being held for  Month and year arrested or charged with current offenses  Month and year admitted to prison for current offenses  Original offenses (if one of their current offenses is parole violation, probation violation or escape)  Month and year arrested or charged with original offenses  Month and year admitted to prison for original offenses  *C-Sentencing Block*  How much of time spent in jail before sentencing was applied to jail sentence  Length of custodial sentence  Drug and Alcohol treatment condition  Sex offenders register condition  Fines and victim payments  Controlling offense (offense referred to in Blocks C-J)  *Q26-31*  State living in at the time of arrest  If they ever expect to be released from prison, and when and any release conditions  *CJEA-Violent Crime and Victims Block*  Violent crime type- murder, rape or other  Victim age, sex and race  Victim injuries  If the victim was known, and relationship  *CJEB-Multiple Victims Block*  Violent crime type- murder, rape or other  Youngest and oldest victims’ age  Victims sex: all female or all male, mostly male or female, split between male and female  If the victims were known, and relationship to the victims if known  *CJF-Property Crime Block (not reviewed in detail)*  *CJG-Drug Crime Block (not reviewed in detail)*  *H-Weapons Block*  Weapons in possession during the crime  Weapons used during the crime  How weapons were used during the crime  Injuries as a result of weapons used during the crime  How weapons used during the crime were obtained  *CJ40-CJ53* Criminal history  Age first arrested  How many times ever arrested for different types of crime (violent, property, drug, other, any)  How many times placed on probation  How many times sentenced to serve time in a juvenile facility  How many times sentenced to serve time in a local or county jail and age first admitted  How many names sentenced to serve in a state or federal prison  Age first admitted | | |
| **Socioeconomic Characteristics** | | **SES** |
| How others would describe race  *SESA-Education Block*  Highest level of school attended before admission to prison and if that year was completed  High school diploma or GED  If received GED whilst in prison  If ever diagnosed with ADD or ADHD, dyslexia or other learning disability  If ever enrolled in special education classes  *Q4-10*  Country of birth  US Citizenship Status and Countries of citizenship  How long lived in the US  Placed lived 30 days before arrest  If homeless 12 months before arrest  If in a correctional facility for 30 days before arrest  How many people living with at time of arrest, and relationship to them  If they have any children  *SESB-Children Block*  How many children and ages  If child was living with them when arrested and where those children are living now, and if they are living together  Visits and communication from children since being in prison, including frequency  If they’ve not had in-person visits, why they think this is  If paying child support at time of arrest  *SESC Employed Block*  Number of jobs and business and full time or part time  *Q12-27*  If not looking for work before arrest, why  When last working and if this was full or part-time work  If any income was received from these sources before arrest: wages, self-employment, social security, public assistance or other government grants, illegal activities  Total income in 30 days before arrest  If income was half the household income  If they had any health insurance 30 days before arrest and where it came from and what type if so  If they expect to be eligible for health care benefits on release from prison  If they lived in foster home, agency or institute before 18  Where they lived most of the time before 18  If household when under 18 received welfare  If they lived in public housing before 18  Relatives and step relatives ever sentenced to serve time in jail or prison (mother, father, siblings, spouse, children) | | |
| **Mental Health (not reviewed in detail)** | | **MH** |
| Mental health within the last 30 days prior to interview  History of mental health diagnoses and treatments | | |
| **Physical Health, Treatment and Disabilities** | | **PH** |
| Height and weight  Sex, Gender, Sexual Orientation  If pregnant when admitted to prison  Medical care received since being admitted to prison  Current Difficulties due to mental, physical and emotional problems  History of major diseases (cancer, diabetes, HIV and others)  Smoker and use of cigarettes  Exercise | | |
| **Alcohol Use** | | **AU** |
| If they’ve ever had alcohol, and the earliest age they did  History of alcohol dependency and abuse | | |
| **Drug Use** | | **DU** |
| Use of recreational drugs ever  Use of recreational drugs 30 days prior to arrest  Use of recreational drugs at time of offense  Use of recreational drugs 12 months prior to admission  *D- Substance Dependency and Abuse Items Block (not reviewed in detail)* | | |
| **Drug and Alcohol Treatment** | | **DTC** |
| If ever received counselling for drug or alcohol use and type of counselling received  If received counselling for drug or alcohol use since admission to prison | | |
| **Rule Violations** | | **RV** |
| If the inmate has been found guilty of breaking prison rules any time since admission and/or any time in the last 12 months  If broken rules relate to drugs, alcohol, weapons, stealing, banned substances, assaulting prison officers, assaulting other inmates, escape, food strikes, rioting, work slowdowns, arson, orderliness  Disciplinary action that took place, such as solitary confinement, confinement to cell, higher security level, new sentence, extra work, change to work assignments, loss or privileged, physical restraints, suspended punishment | | |
| **Programs. Services & Work Assignments** | | **P** |
| If participated Job training. education programmes and work assignments- ever or the last 12 month and/or currentlys, if this was required.  Reasons for not participating and not participating | | |
| **Interview Closeout Screens** | | **IC** |

# Appendix 2 :Survey Routing Diagrams

# Appendix 3: Dense Prisoners with Violent Crimes Variables (1613 words)

The below parameters should be specified in order to produce a dense subset, where the features have not been encoded and scaled.

*dense subset=prep.prep(violent,enc=0, scale=0, impute=1, years='ordinal', th=20,low\_freq\_code\_counts=0)*

In the encoded subset, feature names will be appended with the categorical values. For example variable name ‘V0022-x0\_3’ is the binary column for divorced.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Description** | | **Values (if categorical)** |
| **Offender Demographics and Socioeconomic Info** | | | |
| RV0002 | age bracket | | 1: 18-24, 2:25-34, 3: 35-44, 4: 45-54, 5: 65+ |
| V1213 | sexual orientation | | 1: lesbian or gay, 2: heterosexual, 3: bisexual, 4: something else, 5: don't know, -8 sip |
| offender\_male | if the offender is male (1) or not (0) | | male (1) not (0) |
| offender\_race | if the offender is white (1) or not (0) | | white (1) not (0) |
| V0022 | Marital status | | 1: married, 2:widowed, 3: divorced, 4: separated (not due to incarceration), 5: never married |
| V0023 | served in US armed forces | | 1: yes, 2: no, 8: skipped |
| V0940 | GED or high school diploma | | 1: GED, 2: high school diploma,- 8: skipped |
| V0942 | ADD | | 1: yes, 2: no,- 8: skipped |
| V0943 | dyslexic | | 1: yes, 2: no,- 8: skipped |
| V0944 | Ever enrolled in special ed | | 1: yes, 2: no,- 8: skipped |
| V0945 | Born in US or other country | | 1: US, 2: other,- 8: skipped |
| V0950 | also a US citizen | | 1: yes, 2: no,- 8: skipped |
| V0951 | #years lived in US | |  |
| V0961 | Homeless 12 months before arrest | | 1: yes, 2: no,- 8: skipped |
| V0962 | In correctional facility for entire 30 days before arrest | | 1: yes, 2: no,- 8: skipped |
| V0983 | # children | |  |
| V1005 | living with any children at time of arrest | | 1: yes, 2: no,- 8: skipped |
| V1008 | #children living with at the time of arrest | |  |
| V1080 | # different jobs 30 days before arrest | |  |
| V1138 | income from veteran's administration 30 days before arrest | | 1: yes, 2: no,- 8: skipped |
| V1145 | range of income for 30 days before arrest | | in USD. 1: less than 200, 2: 200-599, 3: 600-999, 4:1000-1999, 5:2000-4999, 6:5000+ |
| V1146 | do they provide half or more of the financial support for the household? | | 1: yes, 2: no,- 8: skipped |
| V1147 | total income from illegal activity 30 days before arrest | | 1: all of it, 2: most of it, 3: some of it, 4: just a little, -8: skipped |
| V1165 | homeless before 18 | | 1: yes, 2: no,- 8: skipped |
| V1166 | in foster home before 18 | | 1: yes, 2: no,- 8: skipped |
| V1170 | welfare/public assistance before 18 | | 1: yes, 2: no,- 8: skipped |
| V1171 | lived in public housing before 18 | | 1: yes, 2: no,- 8: skipped |
| **Personal & Family Criminal History** | | | |
| V1172 | father ever been sentenced or serve time | | 1: yes, 2: no,- 8: skipped |
| V1173 | offenders month ever served time | | 1: yes, 2: no,- 8: skipped |
| V0909 | # times arrested for any offense | |  |
| V0910 | # times placed on probation | |  |
| V0911 | # times sentenced/served in juvenile facility | |  |
| V0912 | # times sentenced/served in local/county jail | |  |
| V0913 | age first admitted to local/county jail | |  |
| V0915 | # times sentenced/served in state/federal prison] | |  |
| V0917 | date first admitted to prison (year) | |  |
| V0918 | age first admitted to state/federal prison | |  |
| V0919 | date first admitted to prison (season) | |  |
| V0899 | age at first arrest for any offense | |  |
| **Health, Alcohol and Drugs** | | | |
| V1200 | Hospitalised 12 months before arrest | | 1: yes, 2: no,- 8: skipped |
| V1201 | Taking a prescribed medication at time of arrest | | 1: yes, 2: no,- 8: skipped |
| V1208 | height (feet) | |  |
| V1209 | height (inches) | |  |
| V1210 | weight (pounds) | |  |
| V1252 | Age first started smoking regularly | |  |
| V1264 | age first drank an alcoholic beverage | |  |
| V1268 | spent time binge drinking in 12 months before admission | | 1: yes, 2: no,- 8: skipped |
| V1369 | repeated problems with law due to drug use in 12 months before arrest | | 1: yes, 2: no,- 8: skipped |
| **Victim Information** | | | |
| V0480 | one or more victim | | 1: one victim, 2: multiple victims, 3: no victim/victim not a person, -8: skipped |
| victim\_white | if the victim is white (1) or not (0) | | white (1) not (0) |
| victim\_male | if the victim is male (1) or not (0) | | male (1) not (0) |
| victim\_below\_18 |  | |  |
| victim\_spouse | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| victim\_child | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| victim\_stranger | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| victim\_sight\_only | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| victim\_casualacq | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
|  |  | |  |
| **Victim Injuries** |  | |  |
| broken\_bones | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| bruises\_swelling | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| internal\_injuries | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| knife\_stab | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| other\_injuries | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| knocked\_unconscious | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| victim\_died | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| not\_physically\_injured | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| rape\_sexual\_assault | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| teeth\_chipped | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| gunshot\_bullet | if the victim(s) sufferred this injury | | 1: yes, -8: skipped/don't know, refuse |
| **Controlling Offense I** |  | |  |
| ctrl\_count | the number of controlling offenses | |  |
| ctrl\_off\_code | controlling offense code | | see offense codes PDF |
| ctrl\_apply | how the controlling offense was determined | |  |
| ctrl\_off\_type | type of controlling offense (violent, property, drug, public, other) | | (1: violent, 2: property,;3 drug, 4:public order, 5:other) |
| V0478 | Act alone/received help during offense | | 1: alone, 2: someone helped, -8: skipped |
| V0479 | how many people helped | |  |
| V0772 | state where living at time of arrest | |  |
| V0774 | where controlling offense took place (with victim) (upcoded) | | 1:at home shared with victim, 2: victim's home, 3: offender's home, 4: commerical place, 5: public place, 6: other place, 7: multiple incidents |
| V0776 | where controlling offense happened (no victim) | | 1: in offenders home, 2: in commerical place, 3: in public place, 4: some other place, 5: multiple incidents |
| **Current Offenses** |  | |  |
| V1267 | drinking at time of offense | | 1: yes, 2: no,- 8: skipped |
| V1326 | under drugs at time of offens | | 1: yes, 2: no,- 8: skipped |
| count\_offenses | the number of current offenses (max 5) | |  |
| 10 | counts of offense code 10 (variations of murder offense- see offense codes PDF) | |  |
| 11 | counts of offense code 11 (attempted muder and variations of intent to kill) | |  |
| 40 | counts of offense code 40 (various kidnapping and imprisonment) | |  |
| 50 | counts of offense code 50 (variations of rape) | |  |
| 70 | counts of offense code 70 (variations of sexual assault) | |  |
| 90 | counts of offense code 90 (variations of robbery and armed robbery with assault) | |  |
| 120 | counts of offense code 120 (variations aggravated assault) | |  |
| 180 | counts of offense code 180 (includes abortion, aiding suicide, child endangerment and others) | |  |
| 190 | counts of offense code 190 (various burglary- see offense codes PDF) | |  |
| 480 | counts of offense code 280 (various weapons offenses- see offense codes PDF) | |  |
| probation | count of offenses listed in probation sections | |  |
| parole | counts of offenses listed in parole sections | |  |
| V0055Y | arrest year | |  |
| V0063 | inmate type | | 3,8,11 |
| V0130 | year arrested- 'standard' inmate | |  |
| low\_freq\_codes\_sum | count of low frequency offenses (that do not have own column) | |  |
| **Sentence** |  | |  |
| V0314 | probation violators, if previous probation was a stand alone sentence | | 1: yes, 2: no,- 8: skipped |
| V0315 | probation violators- if original was part of split sentence | | 1: yes, 2: no,- 8: skipped |
| V0450 | Sentence include court costs/fees (upcoded) | | 1: yes, 2: no,- 8: skipped |
| V0451 | Sentence include fines (upcoded) | | 1: yes, 2: no,- 8: skipped |
| V0452 | Sentence include restitution to victim (upcoded) | | 1: yes, 2: no,- 8: skipped |
| V0453 | Sentence include any other type of fee/payment (upcoded) | | 1: yes, 2: no,- 8: skipped |
| V0458 | Sentence include community service | | 1: yes, 2: no,- 8: skipped |
| V0459 | Does sentence include mandatory drug testing | | 1: yes, 2: no,- 8: skipped |
| V0460 | Does sentence include drug/alcohol treatment options | | 1: yes, 2: no, -8: skipped |
| V0461 | Does sentence include sex offenders treatment program | | 1: yes, 2: no, -8: skipped |
| V0462 | Does sentence include psychiatric/psychological counseling/treatment | | 1: yes, 2: no,- 8: skipped |
| V0463 | Does sentence include other special conditions/restrictions | | 1: yes, 2: no, -8: skipped |
| **Controlling offense- weapon** | | | |
| V0778 | | Did you carry, possess or use a weapon during the offense | 1: yes, 2: no,- 8: skipped |
| V0779 | | # weapons carried during offense | 1: yes, -8: no/skipped |
| V0780 | | specify weapon- firearm | 1: yes, -8: no/skipped |
| V0781 | | specify weapon- toy or bb gun | 1: yes, -8: no/skipped |
| V0782 | | specify weapon- knife | 1: yes, -8: no/skipped |
| V0783 | | specify weapon- other sharp object | 1: yes, -8: no/skipped |
| V0784 | | specify weapon- blunt object | 1: yes, -8: no/skipped |
| V0785 | | specify weapon- another weapon | 1: yes, -8: no/skipped |
| V0786 | | specify weapon- DK/REF | 1: yes, -8: no/skipped |
| V0796 | | # firearms carried in offense |  |
| V0797 | | one or more firearm carried during offense | 1: yes, 2: no,- 8: skipped |
| V0810 | | stole firearm | 1: yes, -8: no/skipped |
| V0884 | | planned to use firearm during offense | 1: yes, 2: no,- 8: skipped |
| V0885 | | did you show or point the firearm during the offense | 1: yes, 2: no,- 8: skipped |
| V0886 | | was the firarm fired during offense | 1: yes, 2: no,- 8: skipped |
| V0887 | | was anyone shot during offense | 1: yes, 2: no,- 8: skipped |
| V0888 | | did any person that was shot die | 1: yes, 2: no,- 8: skipped |
| V0889 | | weapon was used to scare someone | 1: yes, 2: no,- 8: skipped |
| V0890 | | weapon was used to kill someone | 1: yes, 2: no,- 8: skipped |
| V0891 | | weapon used to get away | 1: yes, 2: no,- 8: skipped |
| V0892 | | weapon used to protect yourself | 1: yes, 2: no,- 8: skipped |
| V0893 | | weapon used in any other way | 1: yes, 2: no,- 8: skipped |

# Appendix 4: Mutual Info and Chi2 Tables (369 words)

*20 Most informative variables after one hot encoding according to Mutual Information*

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Name** | **Description** | **Score** |
| 1 | **V0055Y** | Year of arrest | 0.18 |
| 2 | **V0130** | Year of arrest (only populated for inmates who had no previously been on probation or parole) | 0.15 |
| 3 | **10.0-x0\_1** | Count of murder offenses | 0.13 |
| 4 | **ctrl\_off\_code-x0\_10.0** | Controlling offense of murder | 0.12 |
| 5 | **victim\_died-x0\_1.0** | Victim died | 0.09 |
| 6 | **RV0002** | Offender current age category | 0.05 |
| 7 | V0888-x0\_1.0 | Someone that was shot died | 0.03 |
| 8 | **ctrl\_off\_code-x0\_120.0** | Controlling offense of aggrated assault | 0.03 |
| 9 | V0917 | First time offender was ever admitted to state or federal prison | 0.03 |
| 11 | 120.0-x0\_1 | Count of agrravated assault offenses | 0.02 |
| 12 | ctrl\_off\_code-x0\_90.0 | Controlling offense of armed robbery | 0.02 |
| 13 | V1005-x0\_1.0 | Living with children at the time of arrest | 0.02 |
| 14 | ctrl\_apply-x0\_one\_of\_n | One controlling offense of multiple offenses | 0.02 |
| 15 | V0772-x0\_MD | Living in MD (assumed to be Maryland) at time of arrest | 0.02 |
| 16 | V0778-x0\_1.0 | Carried a weapon during offense | 0.02 |
| 17 | V0887-x0\_1.0 | Someone was shot during offense | 0.02 |
| 18 | offender\_male-x0\_1.0 | Offender is male | 0.02 |
| 19 | V1147-x0\_1.0 | All income from illegal activity in 30 days before arrest | 0.02 |
| 20 | V0452-x0\_2.0 | Did not pay restitution to victim | 0.02 |
| 21 | 50.0-x0\_1 | Count of Rape offense | 0.02 |

*Rank 1-10 and 47- 55 Most Informative Variables after one hot encoding according to chi2*

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Feature** | **Description** | **Chi2 score** |
| 1 | **V0055Y** | Year of arrest | 4021 |
| 2 | **V0130** | Year of arrest (only populated for inmates who had not previously been on parole or probation) | 3034 |
| 3 | **10.0-x0\_1** | Count of murder offenses | 1328 |
| 4 | **ctrl\_off\_code-x0\_10.0** | Controlling offense of murder | 1215 |
| 5 | V0917 | First time prisoner was ever admitted to prison | 906 |
| 6 | **victim\_died-x0\_1.0** | Victim died | 850 |
| 7 | **RV0002** | Age range of prisoner | 510 |
| 8 | V0888-x0\_1.0 | Something that was shot died | 351 |
| 9 | **ctrl\_off\_code-x0\_120.0** | Controlling offense of aggravated assault | 326 |
| 10 | **V0890-x0\_1.0** | Weapon was carried with intent of killing someone | 305 |
|  | 11-47 not shown |  |  |
| 48 | V0022-x0\_3 | Divorced | 34 |
| 49 | 40.0-x0\_1 | Count of kidnapping offense | 33 |
| 50 | knife\_stab-x0\_1.0 | Victim was stabbed | 32 |
| 51 | 90.0-x0\_1 | Count of armed robbery offenses | 32 |
| 52 | ctrl\_off\_code-x0\_30.0 | Controlling offense manslaughter | 31 |
| 53 | V0458-x0\_1.0 | Sentence includes community service | 28 |
| 54 | V0778-x0\_2.0 | Did not carry a weapon during offenes | 27 |
| 55 | V0774-x0\_2.0 | Offense took place in the victim's home or apartment | 27 |

# Appendix 5: Further Feature Derivation

### Further work on features already in the dataset

Controlling offense

Other methods to populate the controlling offense could be using a nearest neighbor approaches, or analysing the data about victim injuries and violent type in the Violent Crimes block.

#### *Parole Probation*

Offenses listed in the Previous Offenses pseudo sections could be derived in the same format as current offenses; with a column for each offense code and a count of how many times that offense is listed as a previous offense

*Violent Type*

It should be possible to determine the violent type for all instances where a controlling offense code has been established, if assumptions are used about which offense codes correspond to which violent types offense code

#### *Victim age*

The effect of using other victim age categories in the dataset could be explored. If the multiple victim age variables and single victim age variable could be combined, this could be ordinally encoded.

The effect of distinguishing between a child below the age of 12 and a child below the age of 18 could be explored.

*Offender Race*

Consideration might be given as to how to treat instances where the inmate has selected other races, as well as white.

In the Socioeconomic section, participants are asked what races other people would describe them as. These responses could be considered when constructing a privileged and unprivileged class.

*Offender Sex*

Consider sex assigned at birth when determining if the offender is part of a privileged group.

#### Victim race

The effect of the below could be explored:

* Setting ‘victim\_white’ to 1 if any of multiple victims were white (indicated in variable V0545). The current implementation only considers if most victims were white.
* Keeping the single victim white and multiple victims any white and multiple victims most victims white as three separate columns, but dropping other victim race variables.

*Where the offense took place*

V0774 and V0776 hold the responses to where the offense took place in the single and multiple victims sections respectively. This could be condensed into a single variable.

#### Victim sex

The effect of including ‘even divided’ victim groups in the privileged or unprivileged class could be explored.

*Mother and Father served time in jail*

Condense the separate variables for mother and father served time into a parent served time variable.

*Educational Needs*

V0944, V0943, V0942 could be collapsed (ADD, dyslexia, special education l needs) and provide another sensitive attribute.

*Height and Weight*

Height and weight could be used to work out BMI and set a threshold to turn this into a protected attribute

### Additional Features

*Drug and Alcohol Use*

In the Drug and Alcohol use sections, it may be possible to infer if any of the answers refer to instances before the prisoner entered jail by cross referencing the ‘ever’ and ‘since admission’ questions.

V1269-V1279 ask questions that are similar to questionnaires medical professionals use to identify possible alcohol addiction. A score based on these questions could be derived into a single variable.

V1315-V1321 ask about different drugs used in 30 days before arrest, which might be useful to include.

*Offender Criminal History*

The criminal history questions in the socioeconomic section go through each offense type (Violent, Drugs, Property, Public Order, Other) and ask about arrests, first arrest, time in local jail and type in federal jail for each type. This information could be condense.

*Age of Offender’s Children*

The ages of the prisoner’s children are collected. V0984 contains the age of a single child, and V0894- V1004 allows the capture of ages up to 20 children. This could be condensed into the ages of the oldest and youngest chid, or how many children the prisoner has in age categories.

*Family Criminal History*

V1171-V1174 ask if different family members have ever been sentenced or served time. This could be condensed perhaps into any family members or just parents.

*Health Insurance*

V1148-V1164 ask about different types of health insurance before arrest. These could be densified and included in the dataset.

*Living with at Time of Arrest*

This is captured in variables V0952-V0960. This information could be condensed into categories such as ‘homeless’, ‘with partner’ etc.

*Where Living Before Arrest*

Condense the variables which capture information about where the offender was living in the 30 days before arrest.

1. As Stated in the Codebook p.13- 14 [↑](#footnote-ref-1)
2. A full list of suppressed variables is in the Codebook p.25 [↑](#footnote-ref-2)
3. Intended for use in Google Colab [↑](#footnote-ref-3)
4. The full download ( zip file) for the SPI can be found herehttps://www.icpsr.umich.edu/web/NACJD/studies/37692#Assuming the above link is used, a zip file named ICSPR\_37962-V4 will be downloaded. This contains a file ICSPR\_37962.Documentation including sampling strategy, survey weight files, user guides and licensing can be found at this location. There are three sub-folders containing the different versions of the datasets. File DS001, which contains the public use combined state and federal data. The primary artefacts referenced in this paper are: Codebook 37692-0001-Codebook [34] and Data file 37692-0001-Data [35] [↑](#footnote-ref-4)
5. Auto scikit learn requires a Linux operating system, which was not available. A subscription to a virtual Linux environment (Colab Pro+) was used to install auto scikit learn. Auto scikit learn recommends 24 hours to run a search, but only 12 hours is possible with Google colab. The data was passed to auto scikit learn, but no algorithms were successfully evaluated, even with a restricted search space. It is assumed that this was due to the size of the feature set. [↑](#footnote-ref-5)
6. Some derived features do not require pre processing after the derivation process. [↑](#footnote-ref-6)
7. Many of the instructions are contained on Questionnaire p.50, p.51,p.54,p.59, p.60. [↑](#footnote-ref-7)
8. There are a small number of cases where a prisoner has a single offense that is not stored in the ‘offense 1’ variables (in the set of variables that store the offense codes and types for that prisoner type). If the prisoner only has one offense listed in the current offense columns corresponding to their prisoner type, the current implementation pulls the code from the first offense code and offense type variables. Therefore, if the surveyor has stored the offense type and offense code in one of the other variables (offense 2, offense 3 etc), it would not be captured. Further code could be added to address this. [↑](#footnote-ref-8)
9. The logic to apply when cases where the inmate has no current sentences, because they are awaiting trial, sentencing or are in prison for some reason other than sentenced can be found in the Criminal Justice section, but is not relevant to the prisoner types included in this excercise. Survey of Prisoner Inmates ICPSR, p. 43 of Questionnaire [↑](#footnote-ref-9)
10. There are five controlling offense variables. If the third offense the prisoner listed is the controlling offense, the third of the five offense variables will contain the offense code. [↑](#footnote-ref-10)
11. Question CJEA2 [↑](#footnote-ref-11)
12. *CJEA3.* [↑](#footnote-ref-12)
13. CJEB8. [↑](#footnote-ref-13)