01TXHSM DATA ETHICS AND PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam 18 Jan. 2021

A. CASE (7,5points). Prioritizing patients for intensive care programs

The health system of USA relies on a risk-prediction tool to target patients for "high-risk care management" programs, where identified patients are treated with additional resources. These programs are widely considered effective at improving successful treatment outcomes and satisfaction while reducing costs. Because the programs are very expensive (dedicated nurses, extra appointment slots, etc.), the USA health ministry uses the risk-prediction tool to identify patients who will benefit the most. To achieve such a goal, two key assumptions are made: i) those with the greatest care needs will benefit the most from the program; ii) those with the highest medical expenditures are those with the greatest care needs. Under these assumptions, the targeting problem becomes a pure prediction problem: predict future medical expenditures of patients.

Developers built the algorithm relying on past data to build a predictor $E_{i,t}$: total medical expenditures for patient *i* in year *t*. The algorithm is given a dataset with the following data on care utilized and billed to patient *i* over the year *t-1*

a. Demographics: age, sex.
b. Health insurance type: basic insurance provided by USA or commercial insurance.
c. Code of the main illness diagnosed.
d. Prescribed medications.
e. Appointments made, with type of medical service (e.g., surgical, radiology, etc.).
f. Billed amounts, categorized by type (e.g., outpatient specialists, dialysis, etc.).

It is not known how these features were combined to make the prediction. An independent audit looked at the algorithm's results (patients identified for the intensive care program) from the point of view of race. The published analysis results are reported in the following table:

	White	Black
Nr of patients	45000	5000
Average age (years)	55,5	45,5
Female (%)	60%	70%
Race composition of the intensive care program (%)	80%	20%

1. Provide possible explanations for the results of the audit: clearly state your own hypotheses, and any other information that you suppose in addition to the provided data, in order to coherently support your reasoning. (5p)

2. Which data and/or which part of the process would you change to better achieve the goal of the system? (2,5p)

B. THEORY (TOT 7,5 points). Answer briefly to the questions:

The Algorithmic Accountability Act:

i. Summarize the main points (no need to go into the details of numbers) (5 points)
ii. Briefly comment on the possible use of algorithmic fairness criteria (independence, separation, sufficiency) and/or bias measures (Gini and Shannon index) for the prescriptions of the law proposal (2,5 points).

Exam 18 January 2021 – Indications for a possible solution

A. CASE (7,5points). Prioritizing patients for intensive care programs.

The following comments focus on the interpretation of the case: they should be used as a guide and not as the only possible solution, which -indeed- is not unique, and it is dependent on the reasoning presented and the hypotheses made.

1. Provide possible explanations for the results of the audit: clearly state your own hypotheses, and any other information that you suppose in addition to the provided data, in order to coherently support your reasoning.

The results of the audit show that the set of patients selected for intensive care program is highly imbalanced with respect to race, being 80% white and 20% black.

Two protected attributes were used as features: age and gender. We cannot assume any relationship between gender and medical expenditures, unless childbirth is considered and under the assumption that this category of expenditure is a relevant proportion of the total amount of medical expenditures. Regarding age, it is reasonable to expect that -on average- younger people are healthier than older ones.

Then, looking at the other features, it is possible to make further observations:

b) *Health insurance type*: it is well known that the basic health insurance in US covers less costs and types of medical services than commercial insurances. Considering also the structural race inequalities of the US society, it is reasonable to expect that this variable is a proxy for race, because many more white people own a private insurance than black people do. In addition, private insurances will correlate with higher medical expenditures.

c) *Code of the main illness diagnosed*: some illnesses certainly require higher medical expenditures, but the type of knowledge and data required to verify this correlation is so specific that the effect of this feature cannot be considered here.

d) *Prescribed medications*: this is related to the diagnosed illness; thus, the same limitations apply also here. The type of knowledge and data required to verify the correlations with medical expenditures is so specific that the effect of this feature cannot be considered here.

e) Appointments made, with type of medical service (e.g., surgical, radiology, etc.): the type of knowledge and data required to verify the correlations with medical expenditures is so specific that the effect of this feature cannot be considered here.

f) *Billed amounts, categorized by type (e.g., outpatient specialists, dialysis, etc.):* this is probably the main driver for predicting the medical costs in the next year. It is a direct consequence of feature c, d and e, but it is also dependent on the type of insurance owned by the patient, i.e., feature b).

Given these considerations on the features, we can conclude that the bias arises because of the key assumptions made in design phase: the algorithm predicts health care costs, but costs are highly affected by unequal access to care. In fact, although on average sicker patients need more care, needing health care does not imply receiving health care, which in the context at hand is correlated with race: black patients -on average- generate fewer medical expenses because they have lower access to healthcare, especially for economic reasons (it has been widely debated in the lectures that race and socioeconomic status are correlated, especially in US).

More in general, poor patients face substantial barriers to accessing health care: for example, because they cannot afford expensive insurance plans. Other mechanisms by which poverty can lead to disparities in use of health care are geography and differential access to transportation, competing demands from jobs or child-care, or simply lack of the knowledge for the reasons to seek care.

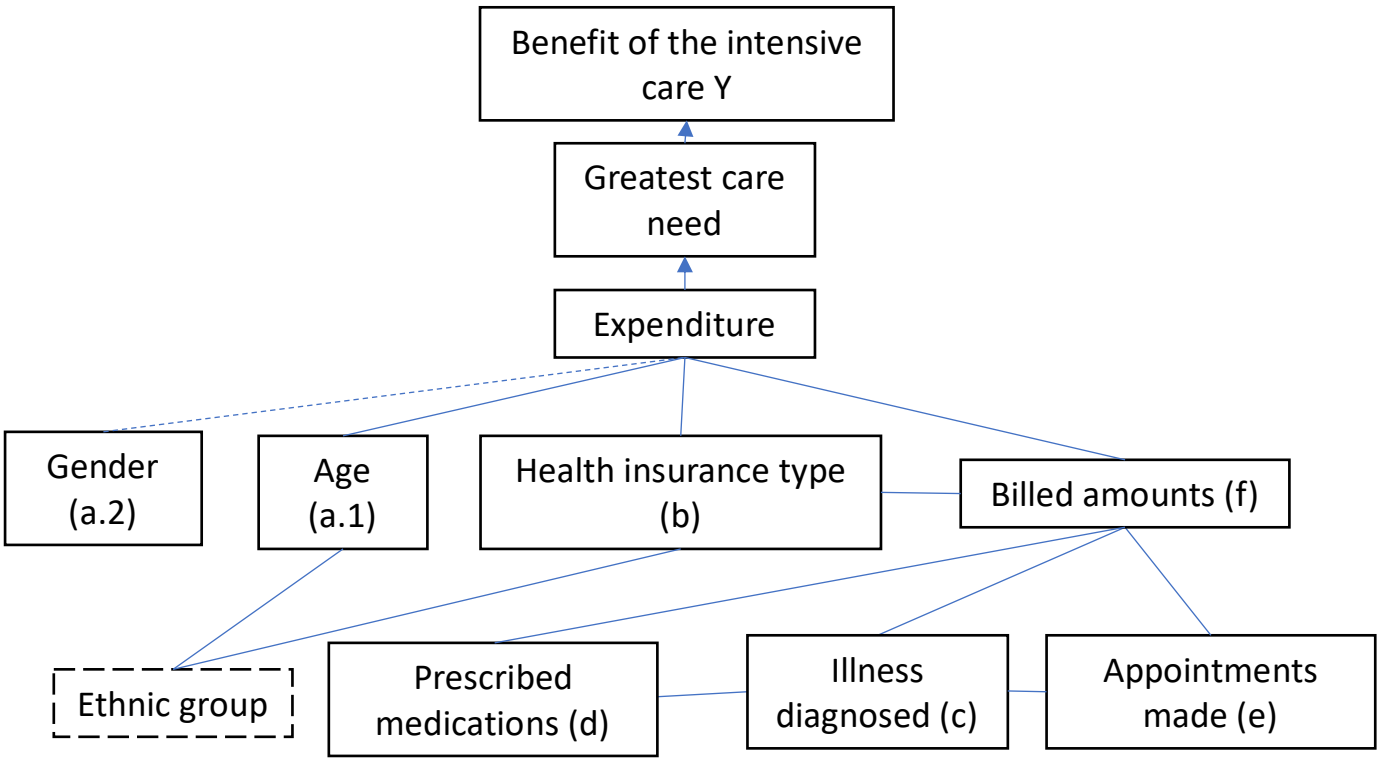
Notice also that it is wrong to interpret the numbers in the table as characteristics of the training set: the table show only audit results (it could be interpreted as test set). Therefore, it is wrong to use the imbalance of people in the set (45000 white vs 5000 black) as cause of discrimination.

2. Which data and/or which part of the process would you change to better achieve the goal of the system?

Although the target variable is medical expenditures, the goal of the program is to include in the intensive care program those patients who would benefit more from it because they need greater care. The previous part demonstrated that the designers' assumption that patients with the greatest future costs could benefit more from the program is the main cause of the biased behavior of the tool.

A more suitable target variable would be a measure of health. For example, the number of active chronic/serious health conditions: those patients are a promising group on which to deploy preventative interventions. This variable could serve the program goal without being so heavily affected by the structural inequalities of US society, as medical expenditures are. Another option would be to create an index variable that combines health prediction with cost prediction, to mitigate the problems of the latter.

Finally, regarding the process related to the usage of the tool, doctors could be presented the result of the tool together with contextual information from patients' electronic health records: in this way they can make a wider consideration of their health status and care need, without solely relying on the prediction of the algorithm.



01 URZSM DATA ETHICS AND PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam 25 June 2021

A. CASE (7,5points). Student Success Predictive Model.

The company ABCsoftware.com (in short: ABC) sells a service based on the Student Success Predictive Model (SSPM) to predict the graduation likelihood of university students in the USA. Once the software produces predictions, no further guidance is given to professors who serve as student advisers, and it is up to the school to decide what to do with predictions: typically, the service is used by universities to recommend students changing major or to withdraw their career.

Predictions are made with a university-specific model: a customized set of predictors is constructed from student records of the specific university, and then combined and weighted using an automated training process designed to maximize predictive accuracy. Historical student records that are used in the training set satisfy the following criteria: i) matriculated between 1st Sept. 2005 to 31st July 2010; ii) had at least three registered successful exams passed; iii) yearly subscription taxes regularly paid. Predictions are made on any type of student, i.e. from incoming freshmen to nearly-graduating seniors, and a final rank is provided with success scores from low to high.

Among the predictors used for training the model in University X, the ones which had higher impact on the score of the software (i.e., responsible for more than 5% of the variance in scores) are the following ones:

A. *Admit Code*: a student's admission type, which can be: first time freshman, transfer from another university, conditional admit (documentation missing)
B. *First Generation Indicator*: "Yes" or "No" indicator of whether any of an individual's parents have ever earned a bachelor's degree.
C. *Median Income by Admission Zip Code*: The median household income in the zip code of a student's home at the time of her admission
D. *High School GPA*: a student's high school grade point average
E. *Legacy Indicator*: "Yes" or "No" indicator of whether a student is someone whose parent or other family member attended the same college.
F. *International Indicator*: Yes" or "No" indicator of whether an individual is an international student
G. *High School Percentile*: a student's high school rank in terms of percentile (in that specific high school)
H. *High School Size*: the size of an individual's high school student body
I. *Veteran Indicator*: "Yes" or "No" indicator of whether a student is a veteran of the United States Armed Forces
J. *In State Resident Indicator*: a "Yes" or "No" indicator of whether a student is a resident of the university home state.

An audit found large disparities in the software outcomes for students of different races, according to the following table:

Table 1

	Black	White	Latino	Asiatic	*TOT*
Success	2500	6500	1200	2400	*12600*
Moderate success	1050	3500	650	2100	*7300*
Failure	1300	220	1100	260	*2880*
Unknown	10	50	30	30	*120*
TOT	*4860*	*10270*	*2980*	*4790*	*22900*

Table 2

	Black	White	Latino	Asiatic	*TOT*
Success	2200	6600	1100	2300	*12200*
Moderate success	1350	3400	650	2000	*7400*
Failure	1600	210	1100	260	*3170*
Unknown	10	45	30	25	*110*
TOT	*5160*	*10255*	*2880*	*4585*	*22880*

Please briefly answer to the following questions:

1. Provide possible explanations for the results of the audit: clearly state your own hypotheses, and any other information that you suppose in addition to the provided data, in order to coherently support your reasoning. (3p)

2. Which measurement issues do you observe? (3p)

3. Which data and/or which part of the process would you change to make the impact of the system less harmful? (1,5p)

B (7,5 points). What sequence of steps should be instituted to ensure a democratic oversight over the development and deployment of new technologies (including digital technologies)?

Indications for a possible solution

A. CASE (7,5points). Comment on the SSPM case.

The following comments should be used as a guide to the exam and not as the solution, which is not unique, and it might be dependent on the reasoning presented and the hypotheses made (if valid).

N.B. exam was provided in 4 versions, according to the following combinations:

- Version 1: Table 1 and variables A, B, F, G, H
- Version 2: Table 2 and variables A, B, F, G, H
- Version 3: Table 1 and variables C, D, E, I, J
- Version 4: Table 2 and variables C, D, E, I, J

Tables were designed to be different in computations but equivalent in their meaning.
Variables were designed to be equivalent when assigned to the exam version. Other variables were very similar in the reasoning required in order to have the same level of difficulty in the analysis.

1) Explanations for discriminations

The text says that the predictions made by the software are *typically […] used by universities to recommend students changing major or to withdraw their career*. Because of such usage, it is possible to focus only on the number of failures. The tables show that the composition of the group of students predicted with failure is highly imbalanced: 45% (table1, 1300/2880) and 50% (table2, 1600/3170) for black people, 38% (table 1, 1100/2880) and 35% (table 2, 1100/3170) for latino, and below 10% for white and asiatic (both tables). We do not have information on error rates. Therefore we consider black and latino as groups discriminated. We reason on possible correlations between the disadvantaged ethnic groups in the output of the algorithm and the predictors listed as "high impact" (thus, we already know they have a relevant correlation with the output):

A. *Admit Code*: we don't have assumptions to believe in a correlation with ethnic groups.
B. *First Generation Indicator*: the generation of parents of students between 2005 and 2010 should have obtained a degree in the '60s or '70s, a period where ethnic (black and latino, mainly) and gender disparities in US colleges were (even) bigger than today, as discussed during the course. This fact sustains the assumption that this variable might correlate with ethnic groups.
C. *Median Income by Admission Zip Code*: correlation of zip code with socio-economic status and with ethnic groups in US has been largely shown during the course.
D. *High School GPA*: while it is reasonable that high school location would correlate with socio-economic status of families, there is no reliable evidence that show that schools in neighborhoods characterized by skewed ethnic compositions give systematically higher/lower marks to students.
E. *Legacy Indicator*: reasoning equivalent to variable B.
F. *International Indicator*: we cannot assume that international students' distribution correlates with the specific ethnic groups involved in the case.
G. *High School Percentile* reasoning equivalent to variable D
H. *High School Size*: we might expect a moderate correlation with ethnic group because usually private schools are smaller and attended by students coming from white families.
I. *Veteran Indicator*: we expect that composition of US army, in general, reflects the composition of the population, i.e., large majority of white people.
J. *In State Resident Indicator*: reasoning like variable F.

2) Measurement issues

Training data from students matriculated between September 2005 and July 2010 and then using that data to base decisions more than 10 years later is intrinsically a very limited choice because the characteristics of the student population may have considerably changed meanwhile. Take, for example, the case of the global economic crisis started in 2008, which might have drastically changed the economic situation of the students and of their families.

Another relevant issue regards an aggregation bias. In fact, the model is unique for incoming freshmen and nearly-graduating seniors, which constitute very different subgroups: the predictors might be not equally important for the two subgroups for instance, we might expect high school GPA to be highly relevant for freshmen, but minimally important for seniors.

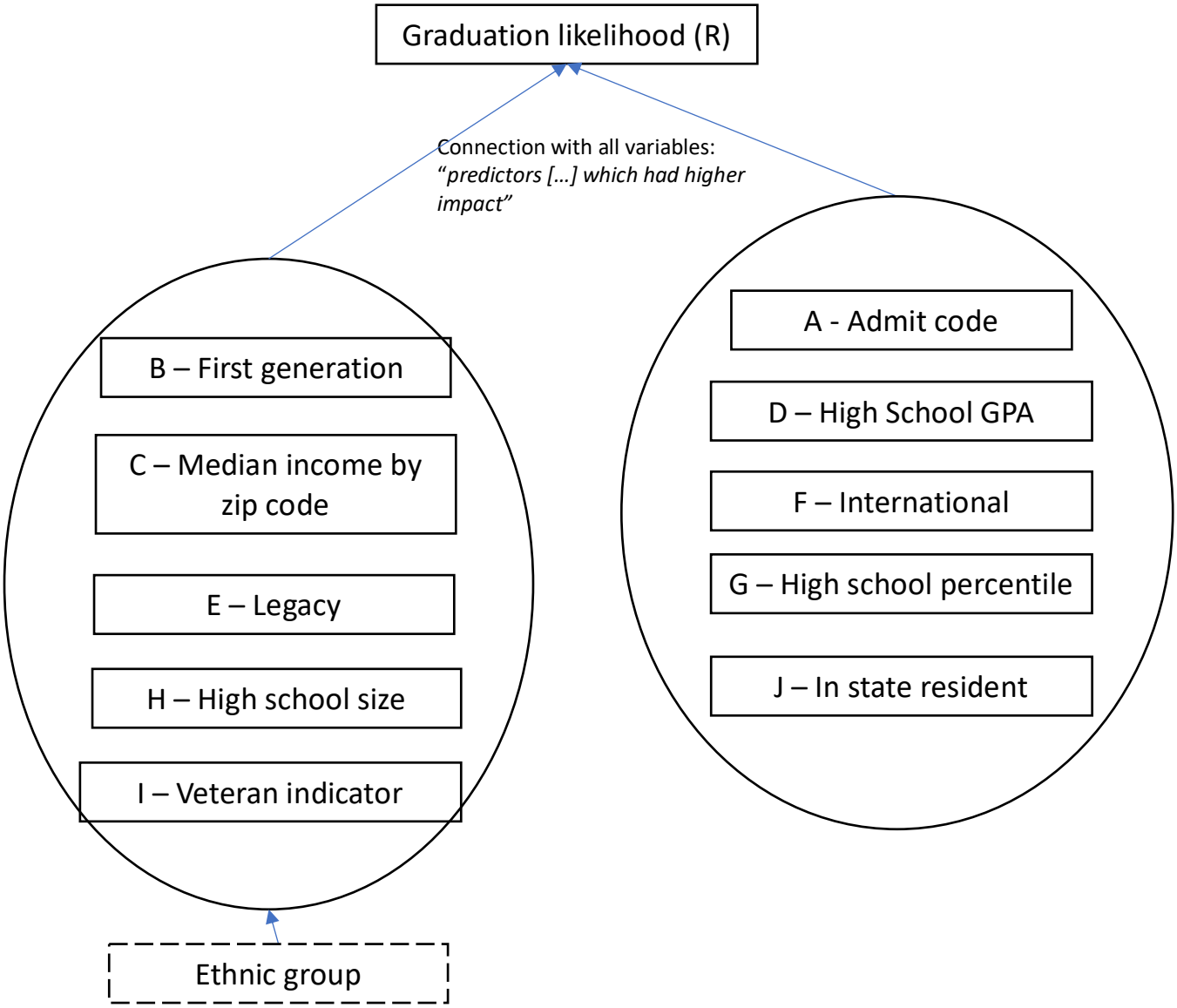
3) System improvement

To overcome the measurement issues specified previously, it would be necessary to train different models for specific sub-groups of the students' population, and of course updating training data whenever possible to better reflect the new characteristics of the distribution. Data should be anonymized whenever possible, and predictions could be given to students' sub-groups and not to individuals to make a different usage of the predictions: instead of *"recommend students changing major or to withdraw their career"* , universities could use the information on the most recurrent patterns of failures to make ad-hoc policies towards the student population.

Most frequent errors observed in the case correction:

- **Question 1 - Explanations for discriminations**
  - The three data collection criteria are not predictors (although they can produce selection bias in the training set, but it should have been clarified)
  - The text clearly states that features listed have a relevant correlation with the classification output, so reasoning on that aspect is unnecessary or even wrong
  - Missing analytical description of feature correlation with ethnicity, or missing/very poor explanation of it
  - The output of the software is used to identify students at "risk of failure" and not success
  - Given the previous point, Asiatic people are not disadvantaged in this case
  - Data provided is not the training data, but the result of the audit
  - Wrong probabilities computed and not reported how they were computed
- **Question 2 - Measurement issues**
  - Repeating the discrimination issues of Q1
- **Question 3 - Improvements**
  - No explanations provided for improvements
  - Improvements are mainly about removal of current set predictors

Finally: many provided a unique answer to the three questions, or they did not clearly identify which part of their text was related to which question. **It should be avoided**!



B. Question on technology and Society (7,5points)

Check slides and video-lectures.

01URZSM DATA ETHICS AND PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam 9 July 2021

A. CASE (7,5points). Car insurance classification model.

An Italian car insurance provider experimented the usage of machine learning for predicting the risk of repaying the costs of caused accidents in a year for a person. In case of high risk, a higher premium is offered to the applicant (who can accept the quote or not). The company used part of its own historical data (from 2010 to 2020) to train a classifier. During the experimentation, applicants could get their quote after filling a form with data about the car and themselves (fields filled are the same used by the classifier). In addition to that data, applicants were required to input information on ethnic group: although this sensitive data was not used by the classification algorithm, it was collected to test the classifications against discrimination. In fact, national and international regulations require that prices should not vary depending on ethnic group and they also forbid training of classification/prediction algorithms with it. The ethnic group had the following possible values: Caucasian, Black, Asiatic.

The features used by the classifier are the following ones.

A. Birthplace: driver's nation of birth (list of all possible countries in the world)
B. Age: driver's age (integer between 18 and 100)
C. City: driver's residence (list of all Italian cities)
D. Car: insured vehicle type, deduced from the car model and year (two possible values: small cars and large powerful cars)
E. Claim history: number of previous claims (values: less than 3, between 3 and 6, more than 6)
F. Yearly distance: estimation of kilometers driven in a year.

The analysis of the data gathered during the 6-months experimentation showed that offered prices varied significantly when only ethnic group differed while all other characteristics were equal.
Therefore, the company dismissed the development of the system, and kept applying insurance costs with respect to traditional, static, risk models.

Please briefly answer to the following questions (GIVE A SEPARATE ANSWER FOR EACH QUESTION):

1. Provide possible explanations for the results of the experimentation: clearly state your own hypotheses, and any other information that you suppose in addition to the provided data, to coherently support your reasoning. (3p)

2. Which measurement issues do you observe? (3p)

3. Which changes in the experimentation data collection process would you introduce to check fairness in terms of separation? (1,5p)

B (7,5 points). Comment the following statement: "technology is essentially the direct application of the results of fundamental science".

Indications for a possible solution

A. CASE (7,5points). Car insurance classification model.

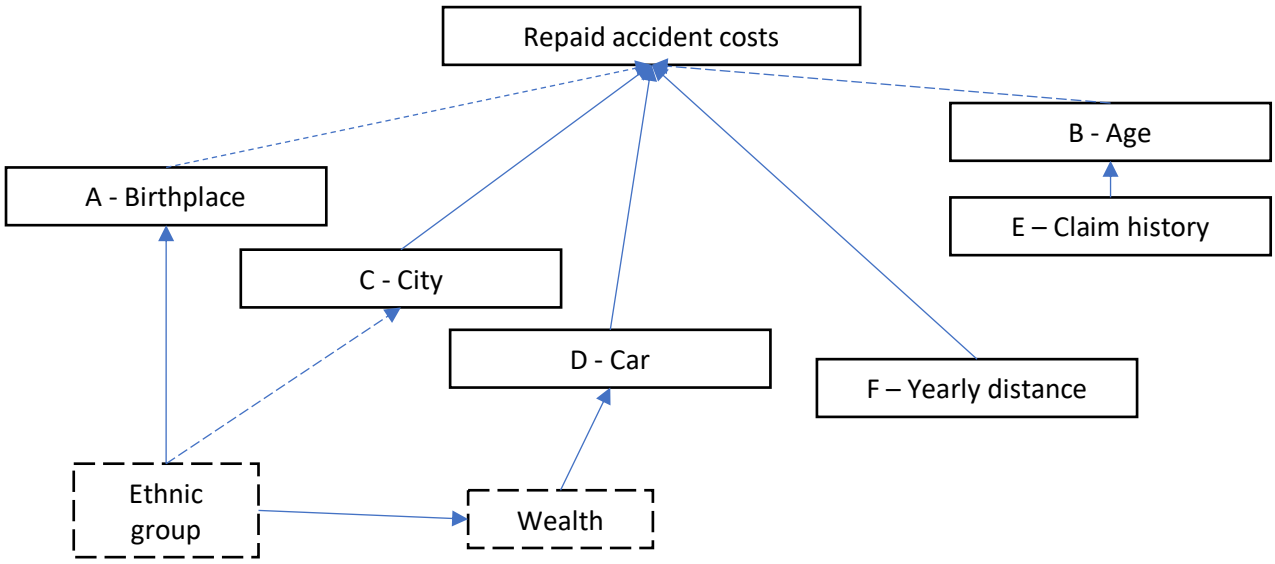
The following comments focus on the interpretation of the case. It should be used as a guide and not as the solution, which is not unique, and it might be partially dependent on the reasoning presented and the hypotheses made (if valid). The exam, given the available time, requires synthetic (but precise and logically coherent) answers, therefore the length of the analysis reported herein should not be taken as a reference.

1) Possible explanations for the experimentation result

The system offers a higher premium to applicants predicted with higher risk of accidents costs. Therefore, repaid accident costs are assumed to be the target variable for which the company has historical data. The following analysis shows the possible correlations of the used features with ethnic group.

A. Birthplace: it can be a proxy for the country where drivers obtained their license, and it can be assumed that learning to drive under different traffic rules and road signs represents an important risk factor which could slightly correlate with higher probability of accidents. The driver nation of birth correlates with ethnic group.
B. Age: young people pay higher premium because the risks of accidents is higher. However, besides that, information provided is not sufficient to make assumptions on correlations with ethnic group.
C. City: the driver's residence can correlate with costs (as in many countries, also in Italy variability of insurance premiums among regions and cities is high). Certain cities could have a much higher percentage of non-Italian residents than other, but detailed statistics should be consulted: hence, out of caution, we assume a weak correlation with ethnic group.
D. Car: larger and more powerful cars get higher insurance costs because repairing them after an accident is more expensive; owning these cars is also more expensive, and wealth – on average – correlates with ethnic group,.
E. Claim history: it is expected to correlate with age. Hoverer, besides that, information provided is not sufficient to make assumptions on correlations ethnic group.
F. Kilometers driven: it correlates with costs, due to increased time on the road and consequent higher chance of being involved in an incident. Hoverer, besides that, information provided is not sufficient to make assumptions on correlations with ethnic group.

Based on the reasoning and hypotheses above, it is possible to draw the following causal model:



Most common errors/imprecisions

- Most answers were accepted as correct, although in some cases it was not clearly shown which features could have caused the discrimination. Full correlational analysis -with hypothesis argumentation, as shown in the guide- is appreciated but remember that the focus should be kept on answering to the question: in this case, it means providing explanation for racial discrimination.

2) Measurement issues

Examples of measurement issues are:

- The feature space is very limited in terms of representativeness of the type and number of risk factors in car incidents; examples are the age of the car, presence of continuous maintenance, driving style, status of the roads and of the safety signals.
- Luxury cars should be treated separately, since they are likely associated with the most expensive quotes but are also far from a representative choice for the average Italian driver.
- Number of claims should be normalized by age, or by number of years driving, to avoid that younger people have a systematic advantage. Also, a mobile/weighted average could be used to filter/weigh less the incidents made many years before.
- Data about ethnic groups was collected without considering a category “other”, whose absence could have had an impact on the quality of the gathered data: for example, people of mixed or missing ethnic groups might have difficulties in checking that field.

Most common errors/imprecisions

- Correlational analysis for explaining discrimination should not be repeated/done here
- No or very poor explanations provided when indicating an issue

3) Changes in the data collection process

To check discrimination in terms of separation, the company should have also tracked the real presence of incidents (and, additionally, related costs reimbursed), thus including in the computation only those quotes that were accepted and followed by a contract. Considering that car insurances are usually valid for one year, the resulting experimentation would have lasted much more than 6 months.

Most common errors/imprecision

- Most people did not read the question until the end: so, their answers were out of scope.

If interested in the real study that inspired the case, check it at https://arxiv.org/pdf/2105.10174.pdf

B. Question on technology and Society (7,5points)

Check slides and video-lectures.

01URZSM DATA ETHICS AND DATA PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam 22 October 2021

A. CASE (7,5points). Facebook advertisement platform

In 2020, Facebook removed the feature that allowed advertisers to choose the race of the people who could see their ads: for example, they eliminated the ad targeting categories "African-American", "Asian", "Hispanic" and so on. This action was taken after years of critics to the company: in fact, researchers and journalists showed that users of certain races were systematically denied opportunities to see advertisements about employment and housing, despite the fact that US federal law prohibits racial discrimination in these domains.

However, despite the removal of the racial categories from the targeting options, discrimination issues on employment and housing advertisements were still reported. An external audit on a sample of users' Facebook feeds reported the following keywords used for targeting: Bollywood movies; Hip pop music; Gospel music; Latin Music; Maldives holidays; Thai cuisine; Black lives matter; Telemundo; Harvard; Black girls rock; Noticiero Univision; Anime movies; Country music; OS Mac; Boca Juniors soccer team.

Please briefly answer to the following questions (give a separate answer for each question):

1) Explain why the discrimination issues remained also after the removal of racial categories, and for which ethnic groups: clearly state your own hypotheses, and any other information that you suppose in addition to the provided data, in order to coherently support your reasoning (5p.)

2) Which intervention, either technical or socio-technical, can you imagine to mitigate the problem? Please also explain why the intervention might help in decreasing or removing the systematic discrimination issue (2,5 p.)

B (7,5 points). What distinguishes professions from other jobs?

Indications for a possible solution

A. CASE (7,5points). Facebook advertisement platform

The following comments focus on the interpretation of the case. It should be used as a guide and not as the solution, which is not unique, and it might be partially dependent on the reasoning presented and the hypotheses made (if valid). The exam, given the available time, requires synthetic (but precise and logically coherent) answers, therefore the length of the analysis reported herein should not be taken as a reference.

The proposed case was based on the real discriminatory issue of the Facebook/Meta advertising platform:

- https://themarkup.org/citizen-browser/2022/05/12/facebook-promised-to-remove-sensitive-ads-heres-what-it-left-behind
- https://www.brookings.edu/research/solving-the-problem-of-racially-discriminatory-advertising-on-facebook/

Question 1

The keywords used for targeting clearly correlate with categories for race/ethnic group:

- African-American: Hip pop music; Gospel music; Black lives matter; Black girls rock;
- Asian: Bollywood movies; Thai cuisine; Anime movies;
- Hispanic: Latin Music; Telemundo; Noticiero Univision; Boca Juniors soccer team
- White Caucasian: Maldives holidays; Harvard; Country music; OS Mac;

Question 2

The most obvious intervention is to remove all the race dependent categories, and constantly monitor the race dependent categories in the delivery of ads: whenever a correlation between race and category arises, the category should be immediately dropped out.

Another possible intervention would be to constantly analyze the number of ad views aggregated by race, and balancing recommendations whenever a given threshold is not respected. For example: if, after a certain amount of time, an advertisement is delivered to 50% African American and significantly smaller percentages to all the other ethnic groups, the delivery to African Americans should be stopped until it will be more balanced with the other races. Notice that this intervention would be possible only in advertising domains where a different treatment with respect to race/ethnic group is admitted by the law (e.g., not in job advertisement, house advertising, etc.).

B. Question on technology and Society (7,5points)

Check slides and video-lectures.

01URZSM DATA ETHICS AND DATA PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam special session 2nd November 2022

A. CASE (7,5points). Predicting severe flu symptoms for elderly patients.

Data-driven algorithms can be employed to screen and predict the risk of various forms of diseases, including common flu. They can find patterns and links in medical records that previously required visiting a human doctor. Algorithmic predictions can be utilized by patients to decide whether to see a doctor for their flu. Suppose we have two different algorithms predicting the severity of flu symptoms in patients and would like to decide which one should be deployed in the real world.

Each algorithm was tested on 400 elderly persons (>= 75 years old), and results are shown in Table 1 and Table 2, divided by gender (with a binary classification M/F). Prediction to develop severe flu symptoms is marked with R=1 in the tables, and otherwise R=0. The actual development of severe flu symptoms is marked with Y=1 in the tables, and Y=0 otherwise.

Algorithm 1:

	Y=1			Y=0			
Gender	R=1	R=0	tot	R=1	R=0	tot	TOT
Female	20	80	100	10	40	50	150
Male	60	90	150	20	80	100	250
tot	80	170	250	30	120	150	400

Algorithm 2:

	Y=1			Y=0			
Gender	R=1	R=0	tot	R=1	R=0	tot	TOT
Female	20	80	100	10	40	50	150
Male	60	120	180	50	20	70	250
tot	80	200	280	60	60	120	400

Answer concisely to the following questions (give a separate answer for each point).
In doing that, clearly state your own hypotheses, and any other information that you suppose in addition to the provided data, in order to coherently support your line of reasoning.

1) Which of the two algorithms is preferable, in terms of lower systematic discrimination by gender, and why? Support your reasoning by showing computations of a fairness criteria at your choice (only one). (4 p.)

2) Now consider the possible errors of the predictions and their social implications in the context of application: which common limitation of the two algorithms do you observe? Explain why. (3,5 p.)

***

B (7,5 points). Briefly describe what is data-driven predictive policing and why it is controversial (7,5p.).

Indications for a possible solution

A. CASE (7,5points). Predicting severe flu symptoms for elderly patients.

The following comments focus on the interpretation of the case. It should be used as a guide and not as the solution, which is not unique, and it might be partially dependent on the reasoning presented and the hypotheses made (if valid). The exam, given the available time, requires synthetic (but precise and logically coherent) answers, therefore the length of the analysis reported herein should not be taken as a reference.

Question 1

Algorithm 1 is preferable in all fairness criteria, because it has lower differences (delta) between genders. The only case when a difference is found to be higher in Algorithm 1 is the difference of true positives. However, the distance between the two delta is 0.07, which compared to the 0.51 of the other condition of separation, i.e. false positive rates, is negligible.

Independence	**Alg. 1**	**Alg.2**
R = 1, Female	0,20	0,20
R = 1, Male	0,32	0,44
delta	-0,12	-0,24

Separation	**Alg. 1**	**Alg.2**
TP Female	0,20	0,20
TP Male	0,40	0,33
delta	-0,20	-0,13
FP Female	0,20	0,20
FP Male	0,20	0,71
delta	0,00	-0,51

Sufficiency	**Alg. 1**	**Alg.2**
PPV female	0,67	0,67
PPV male	0,75	0,55
delta	-0,08	0,12

	Alg. 1	**Alg.2**
FOR female	0,67	0,67
FOR male	0,53	0,86
delta	-0,14	-0,19

Question 2:

There are two types of errors -i.e., false predictions- with two different implications:

- patients who are falsely predicted to develop severe flu symptoms (false positives): they may unnecessarily seek medical intervention
- patient falsely labeled as developing only mild symptoms (false negatives): they will have to cope with severe symptoms for a longer period of time, or even incur in death due to the age of the people involved in the experimentation.

Both algorithms are not accurate, and especially for the false negative rates – FN/(FN+TP) reported below –, which pose the most dangerous threat.

	Alg. 1	**Alg. 2**
FN female	0,80	0,80
FN male	0,60	0,67
FN tot	0,68	0,71

B. Question on theory(7,5points)

Check slides and video-lectures about the PredPol case. In addition, although the SaferRoute case is not about predictive policing, it has similarities with PredPol in terms of type of crimes reported and used for predictions.

01URZSM DATA ETHICS AND DATA PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam 01 February 2023

A. CASE (7,5points). Skin Cancer risk prediction.

Data-driven algorithms are increasingly employed to screen and predict the risk of various forms of diseases, such as skin cancer. They can find patterns in medical records that previously required great levels of expertise and time from human doctors. Algorithmic predictions are then utilized by health-care professionals to create the appropriate treatment plans for patients.

Suppose we have two skin cancer risk prediction algorithms and would like to decide which one should be deployed for cancer screening of patients in a hospital. Each algorithm was tested on 600 persons, and results are shown in Table 1 and Table 2, divided by skin color. Prediction to have high-risk of skin cancer is marked with R=1 in the tables, and otherwise R=0. The actual presence of skin cancer is marked with Y=1 in the tables, and Y=0 otherwise.

Algorithm 1:

Skin	Y=1			Y=0			TOT
	R=1	R=0	tot	R=1	R=0	tot	
White	120	5	125	5	300	305	430
Black	5	45	50	110	10	120	170
tot	125	50	175	115	310	425	600

Algorithm 2:

Skin	Y=1			Y=0			TOT
	R=1	R=0	tot	R=1	R=0	tot	
White	70	70	140	90	90	180	320
Black	40	40	80	150	50	200	280
tot	110	110	220	240	140	380	600

Answer concisely to the following questions (give a separate answer for each point).
In doing that, clearly state your own hypotheses, and any other information that you suppose in addition to the provided data, in order to coherently support your line of reasoning.

1) From an ethical standpoint, which one of the two algorithms do you think is more desirable for deployment in real-world hospitals? Why? Support your reasoning by showing computations of a fairness criteria at your choice (only one). (4 p.)

2) Now focus only on the tests results of the algorithm that you indicated as the most preferable: considering the context of application, identify and explain a limitation that you still observe (only one). (3,5 p.)

***

B (7,5 points). Explain why Facebook has been sued by the US Department of Housing and Urban Development.

Indications for a possible solution

A. CASE (7,5points). Skin Cancer risk prediction.

The following comments focus on the interpretation of the case. It should be used as a guide and not as the solution, which is not unique, and it might be partially dependent on the reasoning presented and the hypotheses made (if valid). The exam, given the available time, requires synthetic (but precise and logically coherent) answers, therefore the length of the analysis reported herein should not be taken as a reference.

Question 1

Algorithm 2 is preferable in all fairness criteria, because it has lower differences (delta) between skin colors. The following table shows all computations (refer to the slides for the computations steps). Remember that only one criterion had to computed.

Independence	**Alg. 1**	**Alg. 2**	**Separation**	**Alg. 1**	**Alg. 2**	**Sufficiency**	**Alg. 1**	**Alg. 2**
R = 1, White	0,29	0,50	TP White	0,96	0,50	PPV White	0,96	0,44
R = 1, Black	0,68	0,68	TP Black	0,10	0,50	PPV Black	0,04	0,21
delta	-0,39	-0,18	delta	0,86	0	delta	0,92	0,23
			FP White	0,02	0,50	FOR White	0,02	0,44
			FP Black	0,92	0,75	FOR Black	0,82	0,44
			delta	-0,90	-0,25	delta	-0,80	0

Question 2

Algorithm 2, although is more acceptable in terms of equitable performances with respect to skin color, it has very high error rates:

- false positives rate (FP/FP+TN) is 0,63 (respectively 0,50 and 0,75 for white and black): it implies that 6 out of 10 patients may unnecessarily go through high-risk and costly medical interventions;
- false negatives rate (FN/FN+TP) is 0,50 (same for each subgroup): half patients are falsely labeled as cancer-free and may face a lower chance of survival.

These error rates suggest that even Algorithm 2 should not be deployed.

B. Question on theory(7,5points)

Check slides and video-lectures.

01URZSM DATA ETHICS AND DATA PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam 9 September 2021

A. CASE (7,5points). Software CheckTax.

CheckTax software is being tested by France to classify citizens as potential tax evaders or not. It works with data on citizens collected both from social networks and from various e-commerce sites that signed an agreement with the French State (e.g. railways, airlines, amazon, etc.). From this data, an indicator of economic well-being is developed, which is then compared to the average of the last three tax returns. Depending on the result of this comparison, the software returns:

R=1 if the person is considered a potential tax evader
R=0 in the opposite case

The experimentation is done in the department of Alpes-Maritimes using data between 2014 and 2017 and combining it with data on tax evasion controls, available from the Ministry of Economy and Finance.

An algorithm audit is made to verify that no systematic discrimination occurs between citizens of French nationality and citizens of other nationalities (but still residing in France). The results are shown in the table, where Y=1 indicates a person on whom there is at least one verified tax evasion, and Y=0 otherwise.

Table 1

	Y=1		**Y=0**	
	R=1	**R=0**	**R=1**	**R=0**
French	600	200	8000	18000
Foreigners	100	900	1400	3600

Table 2

	Y=1		**Y=0**	
	R=1	**R=0**	**R=1**	**R=0**
French	600	1200	8000	18000
Foreigners	100	900	1400	3600

Table 3

	Y=1		**Y=0**	
	R=1	**R=0**	**R=1**	**R=0**
French	600	1200	8000	18000
Foreigners	100	900	14000	3600

Table 4

	Y=1		**Y=0**	
	R=1	**R=0**	**R=1**	**R=0**
French	6000	2000	8000	18000
Foreigners	1000	900	1400	3600

1. Please briefly comment the results of the audit . (5p)

2. Which other ethical issues do you observe, including measurement issues? (2,5p)

B (7,5 points). Describe the main components of the digital gap between Europe and the US and provide possible explanations for such gap.

Indications for a possible solution

A. CASE (7,5points). Software CheckTax.

The following comments focus on the interpretation of the case. It should be used as a guide and not as the solution, which is not unique, and it might be partially dependent on the reasoning presented and the hypotheses made (if valid). The exam, given the available time, requires synthetic (but precise and logically coherent) answers, therefore the length of the analysis reported herein should not be taken as a reference.

1) Comments on the results of the audit

The question was deliberately open to allow a large variety of comments and computations on the tables (four different versions were provided). Herein, for the sake of simplicity and brevity, we report results of computations for two fairness criteria (independence and separation) followed by examples of comments.

Table 1:

Independence			**Separation**	
R = 1, French	0,32		TP French	0,75
R = 1, Foreigners	0,25		TP Foreigners	0,10
delta	*0,07*		*delta*	*0,65*
			FP French	0,31
			FP Foreigners	0,28
			delta	*0,03*

Table 2:

Independence			**Separation**	
R = 1, French	0,31		TP French	0,33
R = 1, Foreigners	0,25		TP Foreigners	0,10
delta	*0,06*		*delta*	*0,23*
			FP French	0,31
			FP Foreigners	0,28
			delta	*0,03*

Table 3:

Independence			**Separation**	
R = 1, French	0,31		TP French	0,33
R = 1, Foreigners	0,76		TP Foreigners	0,10
delta	*-0,45*		*delta*	*0,23*
			FP French	0,31
			FP Foreigners	0,80
			delta	*-0,49*

Table 4:

Independence			**Separation**	
R = 1, French	0,41		TP French	0,75
R = 1, Foreigners	0,35		TP Foreigners	0,53
delta	*0,06*		*delta*	*0,22*
			FP French	0,31
			FP Foreigners	0,28
			delta	*0,03*

Some examples of comments follow.

- Independence criterion:
  - in cases 1-2-4 it is almost respected (deltas of 6-7%), which means that the two groups have a very similar probability of being classified as a potential tax evader;
  - in case 3 the probability of positive classification is significantly skewed toward non-French persons, and this is later explained by the very high number of false positives: as a consequence, the algorithm systematically disadvantaged foreigners in case 3.
- Separation criterion:
  - in all cases the true positive rate of French people is much higher than non-French people (deltas from 0,23 to 0,56). This fact could be reasonably explained by a different data quality and data availability between the two groups, given the type of information collected: for example, one possible explanation is that foreigners might be still used to online commercial services from their nations of origin, including a different national version of the same platform (e.g., amazon.de); as a consequence, a positive classification is less reliable for foreigners and this should be taken into account by the authorities;
  - in cases 1-2-4 the difference in false positive rates is very limited (3% more for french people), but not for case 3 (49% more for foreigners), suggesting an even stronger imbalance in the quality of data for not French people in that case.

2) Other ethical and measurement issues

Some examples:

- only purchases on high-volume e-commerce sites (given the examples) are analyzed; purchases on other types of e-commerce sites, or physical stores, could be very different and still equally relevant for the goal of the classification task;
- social network access is not uniform across the population;
- patterns of usage of social network are also not uniform across the population;
- the wealth indicator is compared with the average of the last three tax returns, however for foreigners who moved recently to France is more likely to have no data in that period;
- the use of this type of data is questionable from an ethical point of view because people probably use of social networks without considering such possible usage of the data; however this is potentially legal because to register to the main social networks and e-commerce platforms the user gives consent to the processing of data and the transfer to third parties. This is actually happening and here are some references for those who might be interested in knowing more on the topic:
  - https://www.politico.eu/article/france-starts-scrapping-social-media-to-catch-tax-fraudsters/
  - https://algorithmwatch.org/en/france-tax-automated-dgfip/
  - https://news.bloombergtax.com/daily-tax-report-international/how-french-tax-authorities-search-online-social-networks-to-detect-fraud

01URZSM DATA ETHICS AND DATA PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam 23 June 2022

A. CASE (7,5points). Mortgage Application Eligibility Score

A mortgage is an amount of money loaned to buy a house. Mortgage applications in USA are automatically analyzed at first by a credit scoring algorithm to determine whether an applicant meets the minimum threshold necessary to be later considered by a clerk. The Mortgage Application Eligibility Score (MAES) has a maximum of 1000 points, and the minimum threshold to enter manual review by a clerk is set to 600. The criteria and respective weights used for computing the score are described below. Consider that negative points for each criterion are not possible, the minimum is always set to 0.

m1 Missed Payments: up to 300 points.

This component of the MAES score counts the number of times people got behind on rent, utilities and medical bills. Every 3 missed payments and until 9 missed payments, the maximum score is lowered by 30 points; after the 9^{th} missed payment, each missed payment counts for -30 points.

m2 Loan size: up to 200 points.

The size of the loan relative to the value of the property the applicant wants to buy.

100% gives 0 points;

90-99% gives 25 points,

70-89% gives 75 points;

50-69% gives 150 points;

less than 50% determines the maximum.

m3 Current employment status: up to 200 points.

Points are given according to the following categories.

Unemployed: 0 points;

employed < 1 year: 50 points;

1 <= employed < 4 years: 100 points;

4 <= employed < 7 years: 150 points;

employed >= 7 years: 200 points.

m4 Amounts Owned: up to 200 points

Total amount owned in bank accounts, relative to the loan size. The following criteria are used.

No bank account or less than 20% of loan size: 0 points;

20% <= loan size < 40% : 50 points;

40% <= loan size < 60% : 100 points;

60% <= loan size < 80% : 150 points;

>= 80% loan size : maximum points;

m5 Length of Credit History: 100 points.

The longer a borrower holds an account at any bank/institute for repaying a debt, the more points he/she gets, according to a linear increase between 1 and 50 years (i.e., 2 points for each year).

Please briefly and clearly answer to the following questions (give a separate answer for each point):

1) Please identify at least a measurement issue for each of the five indicators used to build the MAES (one issue for each indicator is enough, if correct). (5 p.)

2) The US 1974 Equal Credit Opportunity Act makes it illegal for lenders to discriminate based on race, color, religion, national origin, sex, marital status, age. Identify one potential systematic discrimination towards one of these groups, derived from the application of the MAES score, and explain why (2,5p.).

***

B (7,5 points). Briefly describe the general organization of the "ACM Code of Ethics and Professional Conduct" (2.5 points). Then, highlight one aspect of the code that can be related to data-driven automated decision making systems, explaining why (5 points). It is not necessary to remember the exact id or title of the code articles.

Indications for a possible solution

A. CASE (7,5points). Mortgage Application Eligibility Score

The following comments focus on the interpretation of the case. It should be used as a guide and not as the only possible solution, which is not unique, and it might be partially dependent on the reasoning presented and the hypotheses made (if valid). The exam, given the available time, requires synthetic (but precise and logically coherent) answers, therefore the length of the analysis reported herein should not be taken as a reference.

The case proposed is a simplification and adaption to the exam of the problems of the FICO score used in mortgage approvals in USA: https://cpb-us-e1.wpmucdn.com/sites.suffolk.edu/dist/3/1172/files/2014/01/Rice-Swesnik_Lead.pdf

With the advent of mortgage approval algorithms on top of FICO scores, the problem has worsened: https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms

1) Examples of measurement issues for each indicator.

m1 Missed payments:

- it is not normalized by total debts: therefore it does not take into account the on-time payments;
- it does not consider the amount of the missed payments, which is a relevant factor;
- as a consequence of the previous point, it gives the same importance to very different types of debts (e.g. , phone bills are reasonably different from mortgage installment)
- it does not take into account whether past unpaid debt has been paid back and with how much delay.

m2 Loan size:

- it does not consider loan amount (e.g., 50% of 200.000 € is very different from 50% of 20.000 €);
- the 1% difference at border values on 89-90% and on 69-70% determine much larger points differences (respectively, 50 p. and 75 p., equivalent to 25% and almost 40% of the total score).

m3 Current employment status:

- it is very severe sensitive to temporary unemployment at the time of the application;
- it refers only to the duration of the last employment, and more in general, it does not take into account the whole employment history;
- all border values between intervals (e.g., on 4 years) can determine a large difference of points (50 points, which is ¼ of the total score for this criterion).

m4 Amounts owned:

- same percentages but with different absolute values of the loan could make a difference in terms of affordability of the payments;
- all border values (e.g., on 40%) can determine large differences in the points (50 points, which is ¼ of the total score for this criterion).

m5 Length of credit history:

- measurement is on number of years, but it is not normalized by the age of the applicant, which makes a cap for the maximum points achievable;
- any other existing bank account not linked to a debt repayment is not considered (0 points);

updated text (improved answer)

2) Possible discrimination with any of the following attributes protected by the US 1974 Equal Credit Opportunity Act: race, color, religion, national origin, sex, marital status, age.

Indicator on employment status (m3) could disadvantage female because of more frequent interruptions of work (on average). Length of credit history (m5) makes a systematic disadvantage towards younger people. Indicators related to wealth and socio-economic conditions highly correlate with ethnic group, and potentially also with national origin: in this case variable "amounts owned" (m4), however, is expressed only in relation to loan size, therefore the discrimination risk is proportional to it.

Most frequent errors:

Measurement issues (each indicator was 1 point worth):

- comments not in line with the measurement construct, which is identified by the explicit name of the measure/variable (e.g., if the desired construct is employment status, commenting solely on income is not acceptable);
- no explanations (e.g., why you consider proportionality of points among intervals desirable, why you consider categorical scale not appropriate in the specific context);
- too vague and general statements, not specific to the criterion: e.g. writing everywhere "arbitrary thresholds" is not acceptable;
- the rational of the indicator should not be contested per se and be the only content of the answer;
- focus requested by the question was not on discrimination issues, only on measurement issues.

Discrimination issue:

- identifying a discrimination with a non-protected attributed of the given list.

B. Question on theory (7,5points)

Check slides/material and video-lectures.

01URZSM DATA ETHICS AND DATA PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam 15 July 2022

A. CASE (7,5points). Job Recruitment Recommender

The IT company Lockedin, based in US, developed and runs an online job platform that is used by US tech companies and American citizens to offer/find jobs, also using queries by keywords.
The company is working to enhance the service with a recommendation algorithm to automatically advertise job offers to users that are deemed good candidates for that job. The algorithm will be trained on historical data on job applications of the last five years, acquired from three US leading companies in the IT sector.
The deal between Lockedin and the three companies has been made with the following constraints:

- the data will contain the lists of applicants (anonymized) and the outcome of the hiring process (hired/not hired), for the following job positions: software engineer, database administrator, software developer, computer network architect, information security analyst;
- for each applicant, the following information is available: education level, university degree (yes/no), list of skills extracted from cv, current pay (numerical value, 0 if currently unemployed);
- the respect of the following constraints to avoid systematic discriminations of the recommendations:
  - Gender: inverse imbalance ratio (IIR) index >= 0.25;
    - the IIR index is computed as: min(class frequencies) / max (class frequencies)
  - Ethnic group: at least 10% for the less represented group;
  - Age: equal cardinality of age groups.

Consider that the protected attributes are encoded with the following values:

- gender: M/F;
- ethnic group: Caucasian, Afroamerican, Asian, other;
- age: 18-25, 26-50, 50+;

Answer concisely to the following questions (<u>give a separate answer for each point</u>).
In doing that, clearly state your own hypotheses, and any other information that you suppose in addition to the provided data, in order to coherently support your line of reasoning.

1) For which protected attribute(s) there is still a risk of systematic discrimination, despite the constraints placed? Explain <u>why</u>. (5 p.)

2) Propose a specific improvement for mitigating <u>one</u> selected issue from previous answer, explaining the rational for the suggestion and persistent limitations (if any). (2.5)

***

B (7,5 points). Briefly describe the main characteristics of the fairness qualitative assessment (2,5p.) Then, take one step of the procedure and explain it in the context of the Job Recruitment Recommender case (5p.).

Indications for a possible solution

A. CASE (7,5points). Mortgage Application Eligibility Score

The following comments focus on the interpretation of the case. It should be used as a guide and not as the only possible solution, which is not unique, and it might be partially dependent on the reasoning presented and the hypotheses made (if valid). The exam, given the available time, requires synthetic (but precise and logically coherent) answers, therefore the length of the analysis reported herein should not be taken as a reference.

1) Risks of systematic discriminations

The protected attributes mostly involved in the case are gender, ethnic group, age. Their relations to the proxy attributes are explained below, under the hypothesis that proxy attributes correlate with the target variable.

Gender

Clear prevalence of men is a distinguishing feature of the IT field bad for the job positions included in the training data. The constraint of IIR index >=0.25 is not a guarantee that such imbalance will not be reflected in the learning process: table below shows that 0.25 for the binary case implies 20% female and 80% men.

Female	Male	IIR
0,1	0,9	0,11
0,2	0,8	0,25
0,3	0,7	0,43
0,4	0,6	0,67
0,5	0,5	1,00

On top of this, systematic discrimination in gender may occur because of presence of proxy attributes in the data, assuming that they are positively correlated to the target variable:

- "list of skills extracted from cv": because prevalence of men in IT education and market will result in prevalence of more technical skills for male candidates;
- "current pay (if present)": because of the salary gap between males and females;

Ethnic group

Another characteristic of the IT sector in US (and in general in the Global North societies) is the predominance of white people. Therefore, the requirement of at least 10% for the less represented group is not enough to rebalance the training set. This might result in systematic advantage of white candidates, in consideration of their relation to "education level" and "university degree," probably better correlated to success in the hiring process and thus to target variable in the training data.

Age

If the latter observation holds, young candidates without university degree education that fall in the category 18-25 and with less skills are disadvantaged, regardless of the balance of the age categories.

Most frequent errors:

- answer is not adherent to the notion of "protected attribute", widely documented in the course;
- the answer does not refer to considerations on the constraints placed, especially those on protected attributes, which are explicit part of the question;
- observations made on proxy variables without explaining why they are correlated with target variable (or without making the hypothesis explicit)
  - as a consequence, under/over representation of specific groups per se are not sufficient explanations of discriminations;
- wrong interpretation of the IIR index;
- missing explanations.

2) Examples of possible improvements:

For a better gender balance (e.g., at least 35% female) in its current binary representation, it would be necessary to require a minimum IIR of 0.47 (approx. 0.50). Please notice that this intervention might not be sufficient to counterbalance the effect of the proxy variables, which depends on the correlations to be found in the data (between proxy variable and protected attribute, and between proxy variable and target variable).

Most frequent errors:

- too abstract/vague answer, without the clear identification of the improvement and of its rational;
- improvement not specific to the discrimination issue;
- explanation not provided;
- identifying only the removal of a feature, repeating what has been written in q1;
- although only one improvement was asked, listing more than one, some of which are wrong/imprecise/incomplete.

B. Question on theory(7,5points)

Check slides and video-lectures.

01URZSM DATA ETHICS AND DATA PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam 16 September 2022

A. CASE (7,5points).

Predicting absenteeism at work

The managers of a large chain of supermarkets in Italy want to contrast the absenteeism at work of the company's cashiers. They want to understand which factors are related to the most frequent absences. Given such a goal, they built a dataset merging the available information on the absences with the demographic data of the cashiers. The obtained dataset contains absences from 1st September 2015 to 1st September 2022. For each absence, the information in the list below is available: variables A, B and C are related to when the absence took place; variables from D to K are related to the person who made the absence; variable L is the target variable. Notice that holidays and sick days do not count as absences, but only work permits.

A) Day of the week (five possible values, from Monday to Friday)
B) Month (12 possible values, from January to December)
C) Season (4 possible values: winter, springtime, summer, autumn)
D) Distance from residence to work (number of kilometers)
E) Current salary (€)
F) Service time (number of years the person is at the company)
G) Education (3 possible values: bachelor degree, master degree, phd)
H) Son (number of children)
I) Smoker (yes/no)
J) Weight (kgs)
K) Height (centimeters)
L) Target variable: number of hours of absence

The company signed a consultancy contract with a data scientist, who is charged with the tasks of building a prediction model and giving advises on its strengths and limitations.

You are that data scientist who signed the consultancy contract.

Answer concisely to the following questions (give a separate answer for each point).
Clearly state your own hypotheses, and any other information that you suppose in addition to the provided information, in order to coherently support your line of reasoning.

1) Which measurements issues, model limitations or data collection issues do you report to the managers? Identify two issues (in total), and explain them. You don't have to suggest remedies. (4 p.)

2) Which risk of systematic discrimination towards a protected attribute do you identify and report to the managers? Identify only one and explain why. (3,5 p.)

***

B (7,5 points). Explain what the Campbell's law and the Goodhart's law assert (also known as reflexivity problem) (4p.), then make an example of real or hypothetical automated decision making system where they could apply (3,5p.)

Indications for a possible solution

A. CASE (7,5points). Predicting absenteeism at work

The following comments focus on the interpretation of the case. It should be used as a guide and not as the only possible solution, which is not unique, and it might be partially dependent on the reasoning presented and the hypotheses made (if valid). The exam, given the available time, requires synthetic (but precise and logically coherent) answers, therefore the length of the analysis reported herein should not be taken as a reference.

1) Examples of

- measurement issues
  - A) Day of the week: large supermarkets can be open 7/7 or 6/7, this variable does not track weekends;
  - G) Education: low levels are excluded, as a consequence it is probable to have majority of missing values due to the common low level of education required for this job;

- model limitation
  - considering the organization of the dataset (each row is an absence), and the target variable "number of hours of absence", the model can predict the longest absences, but not the most frequent absences, which is the goal of the managers;
  - data of years 2020 and 2021 might be considered for a separate model because of probable highest absenteeism due to the difficulties of the hardest pandemic times;

- data collection limitations
  - the current organization of the dataset is by absence (each row is an absence), hence it might probably include very few data about people hired recently in the company;
  - the year is not tracked: absences in the same day and month of different years are not distinguished.

2) Risk of systematic discriminations:

First of all, it has to be considered that the work of cashiers is a typical example (also illustrated during the course) of highly skewed job by gender: thus, the database will be probably filled with more records related to women, ***under the hypothesis that the distribution of absences will reflect such imbalance***. In addition, women's child-care is usually predominant, that would imply ***more and longest absences for women***. These are the necessary premises to make a robust hypothesis on the ***correlation between the target variable and the protected attribute gender***.

With the premise of the hypothesis explained above, proxy variables for gender need to be identified and explained (remember: a **discrimination will occur only if the proxy variables are significant predictors of the target variable)**:

- the variable "current salary" is expected to be a proxy for gender, because of the salary gap between women and men;
- feature K (height) is expected to be highly correlated with gender: globally, the mean height of women is shorter than that of men.
- feature J (weight) is expected to be slightly correlated with gender: on average it is higher for men.

A specific note concerns variable H ("Son", number of children): there is no rational to hypothesize that men cashiers (globally and in that supermarket) have significantly more/less children than women cashiers, hence this variable should not be considered as a proxy for gender. However, if we assume that on average the more

the children the longest and numerous the absences are, such correlation with target variable might also reinforce the systematic discrimination towards gender explained above.

The exam has been designed by adopting the variables of an existing dataset on absenteeism at work: https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

Most common errors

Question 1:

- answering with discrimination issues: the question was not about that;
- as stated during the course and in several past exams' examples, just writing that a variable is useless is not an acceptable answer;
- no explanations given.

Question 2:

- it is not explained why the discrimination would occur (e.g. correlation with target variable);
- not referring to a protected attribute;
- answering with measurement issues: the question was not about that;
- no explanations given.

B. Question on theory(7,5points)

Check slides and video-lectures.

01TXHSM DATA ETHICS AND PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam 14 July 2020

A. CASE (7,5points). The PESO case.

The software PESO (Public Employment Service Optimization) is an algorithmic profiling system built to reach two goals: 1) increase the efficiency of the job consultancy public services in Spain; 2) increase the effectiveness of active programs for the labor market. The system is based on a statistical model of the prospects of job seekers, which classifies job seekers into two categories:

- Group A: job prospects which are either
  - A1 high short-term prospects (high chances of finding a job within 7 months), or
  - A2 low long-term prospects (low chances of finding a job in the next 2 years).
- Group B: mediocre job prospects

Depending on the category in which job seeker falls, they will be offered different support in reintegrating into the job market. Group A1 is made up of persons with a 66% chance of finding work for at least 3 months within the next 7 months. This group will receive less support through active labor market programs as it is assumed that they will likely find work without further training. Group A2 includes persons with a probability < 25% of being hired for at least 6 months within the next 2 years. This category should also receive fewer support measures. The financial focus will therefore be placed on group B, which includes all persons who are not part of the group A (A1+A2).

Please note that score thresholds for the assignment to groups A1 and A2 were chosen by maximizing, respectively, the correct prediction of successful integration ex post and the correct prediction of unsuccessful integration ex post. The statistical model used the following data:

A. Gender: Male / Female
B. Age group: 0-29 / 30–49 / 50+
C. Citizenship: Spanish / EU except Spanish / Non-EU
D. Highest level of education: Grade school / apprenticeship or vocational school / high- or secondary school / university
E. Occupational group: production sector / service sector
F. Regional labor market: the 17 autonomous regions of Spain
G. Days of gainful employment in last 5 years: ≤ 75% / > 75%
H. Assistance from public employment service obtained in the past 5 years: 0 / 1 / 2 / >2

Please answer to the following answers:

1) Which risks of systematic discrimination do you observe? (4p)
2) Identify one measurement issue (1p)
3) State a different goal from the current two that drove the development of the system. Which data and/or which part of the process would you change to better fit the new goal of the system? (2,5p)

B. THEORY (TOT 7,5 points). Answer briefly to the questions:

B1 (5 points). Compute Gini-Simpson Index (normalized) for the data in the following table (3,5p.) and explain how to interpret the result (1,5p.)

Tab1

Class	N_i
African-american	16
Asian	12
Caucasian	22
Hispanic	4
Other	12

Tab2

Class	N_i
African-american	16
Asian	12
Caucasian	22
Hispanic	14
Other	12

Tab 3

Class	N_i
African-american	6
Asian	12
Caucasian	22
Hispanic	4
Other	12

Tab 4

Class	N_i
African-american	16
Asian	12
Caucasian	22
Hispanic	40
Other	12

Tab 5

Class	N_i
African-american	16
Asian	12
Caucasian	12
Hispanic	4
Other	12

B2 (2,5 points). Describe the History and Etymology of the word "Technology".

Exam 14 July 2020 – Indications for a possible solution

The following comments focus on the interpretation of the cases, the data in the table and the bias measure. It should be used as a guide to the exam and not as the solution, which is not unique and especially for case A it is highly dependent on the reasoning presented and the hypotheses made.

A. CASE (7,5points). The PESO case.

The exam was inspired by a real case in Austria:

- https://epub.oeaw.ac.at/ita/ita-dossiers/ita-dossier052en.pdf
- https://www.frontiersin.org/articles/10.3389/fdata.2020.00005/full

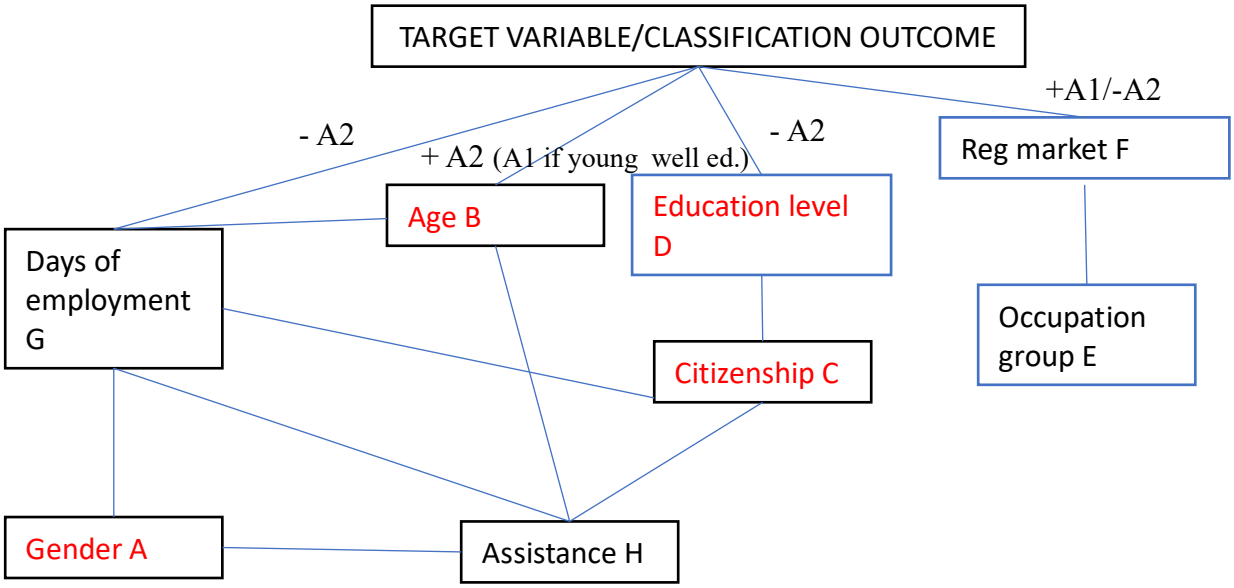
1) Discrimination issues

It can be deduced from the text that two models have been trained, respectively for maximizing true positive rate of successful integration (A1) and true positive rate of unsuccessful integration (A2). All cases not explicitly predicted as successful integration and unsuccessful integration will fall into category B, which implicitly includes low short-term prospects and high long-term prospects.

Since only people who are labeled as category B receive assistance for seeking a job, any hypothesis of systematic correlation of certain groups with **category B** means an **advantage**. On the contrary, any hypothesis toward high correlation with **category A** (A1 high short-term and A2 low long-term) means a **disadvantage**, also via proxy variables, as reported below. Please notice that at the exam only the following combinations of variables were present: 1) AB EH 2) CD FH 3) AC EG 4) BD FG

A. Gender: Male / Female
    - Women, on average, have a more discontinuous career because they carry a higher burden of family care: this might be reflected by a possible correlation with variable G value ≤ 75%
    - Women, on average, have (still) lower presence in the job market, thus might be more likely labeled as A2 and disadvantaged: a correlation with higher values of H is expected.
B. Age group: 0-29 / 30–49 / 50+
    - First category will correlate with lower values of G and H:
        - If well educated (e.g university level), young people will be labeled as A1
        - On average, older people (and mostly with lower level of education) will get A2
C. Citizenship: Spanish / EU except Spanish / Non-EU
    - Because of the characteristics of European flows of immigration, Non-EU people might correlate with lower levels of variable D, G and H
    - Regarding EU people, it would depend on the type of immigration and on their level of education (variable D)
D. Highest level of education: Grade school / apprenticeship or vocational school / high- or secondary school / university
    - The lower the level of education, reasonably the higher the probability to be classified as A2
E. Occupational group: production sector / service sector
    - Highly dependent on the historical period, no hypotheses are reliable here
    - It will be correlated with specific regions (variable F)
F. Regional labor market: the 17 autonomous regions of Spain
    - People living in regions with very stagnant economies might be labeled more often as A2
    - People living in regions with very dynamic economies might be labeled more often as A1
G. Days of gainful employment in last 5 years: ≤ 75% / > 75%
    - Reasonably, lower values will imply label A2, feeding the feedback loop of unemployment
H. Assistance from public employment service obtained in the past 5 years: 0 / 1 / 2 / >2
    - More assistance might be related to pregnancy, poverty, and more general problems that do not allow a person to be active with continuity in the job market: as a consequence, it is

expected more assistance in correspondence of lower values of variable G. For the same motivations, higher correlation with A, B and C might be possible.



2) Measurement issues

A few examples of measurement issues:

1. Given the type of data collected, historical data on non-Spanish people (especially non-EU) might be incomplete or inaccurate, rising issues of data quality for that category.
2. Age category 0-29 might be unbalanced because there will be no data on people 0-16 y.

3) Possible improvements

A different goal for the system could be to maximize the number of matches of job offers and job seekers. The classification system could be turned into a recommender system that uses the following data: i) competences searched/offered; ii) similarity with previous job experiences (if any); iii) desired location; iv) number of working hours per week; v) range of salary.

Results should be presented in random order and not prioritized to minimize the possibility of systematic biases in the rank. An alternative could be a prioritization based on an indicator of need of job: in this case, additional data should be available for the public agency, e.g. economic situation, weeks of non-voluntary last unemployment, total weeks of non-voluntary unemployment.

B. THEORY (TOT 7,5 points). Answer briefly to the questions:

B1 (5 points). Compute Gini-Simpson Index (normalized) for the data in the following table (3,5p.) and explain how to interpret the result (1,5p.)

The higher the index, the higher is the heterogeneity: it means that categories have similar frequencies and risk of propagating bias in a trained model is lower.
In all the tables of the exam, regardless of the specific numbers, Gini-Simpson index is always higher than 0,90. In fact, in all tables at most one or two categories (out of five) have very different frequencies, where "very different" herein means about/more than the double of the median value.

TAB 1			
Class	**n_i**	**f_i**	**⟦f_i⟧ ^2**
African-american	16	0,2424	0,0588
Asian	12	0,1818	0,0331
Caucasian	22	0,3333	0,1111
Hispanic	4	0,0606	0,0037
Other	12	0,1818	0,0331
tot	66		0,2397
m	5		
m/m-1	1,25		
1 - sum[$(f_i)^2$]	0,7603		
GINI NORMALIZED	0,9504		

TAB 2			
Class	**n_i**	**f_i**	**⟦f_i⟧ ^2**
African-american	16	0,2105	0,0443
Asian	12	0,1579	0,0249
Caucasian	22	0,2895	0,0838
Hispanic	14	0,1842	0,0339
Other	12	0,1579	0,0249
tot	76		0,2119
m	5		
m/m-1	1,25		
1 - sum[$(f_i)^2$]	0,7881		
GINI NORMALIZED	0,9851		

TAB 3			
Class	**n_i**	**f_i**	**〚f_i〛^2**
African-american	6	0,1071	0,0115
Asian	12	0,2143	0,0459
Caucasian	22	0,3929	0,1543
Hispanic	4	0,0714	0,0051
Other	12	0,2143	0,0459
tot	56		0,2628
m	5		
m/m-1	1,25		
1 - sum[$(f_i)^2$]	0,7372		
GINI NORMALIZED	0,9216		

TAB 4			
Class	**n_i**	**f_i**	**〚f_i〛^2**
African-american	16	0,1569	0,0246
Asian	12	0,1176	0,0138
Caucasian	22	0,2157	0,0465
Hispanic	40	0,3922	0,1538
Other	12	0,1176	0,0138
tot	102		0,2526
m	5		
m/m-1	1,25		
1 - sum[$(f_i)^2$]	0,7474		
GINI NORMALIZED	0,9343		

TAB 5			
Class	**n_i**	**f_i**	**〚f_i〛^2**
African-american	16	0,2857	0,0816
Asian	12	0,2143	0,0459
Caucasian	12	0,2143	0,0459
Hispanic	4	0,0714	0,0051
Other	12	0,2143	0,0459
tot	56		0,2245
m	5		
m/m-1	1,25		
1 - sum[$(f_i)^2$]	0,7755		
GINI NORMALIZED	0,9694		

01TXHSM DATA ETHICS AND PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam 9 Sept. 2020

A. CASE (7,5points). Predicting high school exams scores.

The government in State X recently used a computer-generated score to replace the graduation exams for secondary education (i.e., the last year) in high schools, canceled due to Covid-19 outbreak.

The score has been computed starting from an estimate given by teachers, who gave their suggested mark based on their knowledge of the students and other information they might have. Marks are: A* (best), A, B, C, D, E (worst), and U (fail). Each student got a suggested mark for each subject.

Those suggested marks were then adjusted by a computer algorithm with the following process:

A. The algorithm made the following transformation: A*=6, A=5, B=4, C=3, D=2, E=1, U=0
B. For each student, the average mark is computed among all the subjects
C. From data of step B, a rank of suggested grades (RankSG) is computed for each age in each school (i.e., in each school, grades for classes of the same year were put together)
D. The school's past performance on 2019 exams were retrieved and a rank made (Rank2019)
E. The final students' grade is made by making RankSG adherent to the distribution of grades in Rank2019: for instance, if a student is halfway down the RankSG, then his/her grade is the grade obtained by the the person halfway down the Rank2019; if a student is in the position corresponding to 25% of the RankSG, then his/her grade is the grade obtained by the person at position 25% in Rank2019; and so on.

The automated adjustment of the suggested marks does not take place for small schools, because data is not enough to get reliable ranks: in this case, RankSG is accepted as is.

The goal of applying such an algorithm was to avoid grade "inflation", because usually teachers are too optimistic about the prospects of their students (several studies support this fact).

An independent audit looked at the algorithm's results. The analysis showed that:

I. half students had their grade lowered in comparison of their mark in last year;
II. the algorithm disproportionately hurted students from working-class and increased the scores of students from private/élite schools;
III. students from families of nationality of State X were less disadvantaged than the students from families with other nationalities

1. Provide possible explanations for the results of the audit: clearly state your own hypotheses, and any other information that you suppose, in order to coherently support your reasoning. (5p)

2. List the measurement issues that you observe in the process and/or data used (2,5p)

B. THEORY (TOT 7,5 points). Answer briefly to the questions:

B1 (2.5 points).

Students in School B received a more advanced instruction which improves the traditional method (which was adapted by School A). The results of the exams (scale of grades: 0-100) are shown in the table below, and it can be observed that both male and female students in School B have higher average scores.

Based on the results, parents of students in School A asked the director of the school to use the advanced teaching techniques. However, the director of School A refuses the request showing that the result is the opposite, i.e. grades on school A are higher than grades on school B.

Please explain why (computations are not mandatory).

CASE 1

	Gender			
	Male		Female	
School	n	Average	n	Average
A	75	83,5	25	79,5
B	25	85	75	81

CASE 2

	Gender			
	Male		Female	
School	n	Average	n	Average
A	40	83	10	80
B	10	85	40	81

CASE 3

	Gender			
	Male		Female	
School	n	Average	n	Average
A	75	94,5	25	89,5
B	25	96	75	91

B2 (5 points). Briefly summarize what are the most prominent socio-technical issues in digital contact tracing for Covid-19.

Exam 9 September 2020 – Indications for a possible solution

A. CASE (7,5points). Predicting high school exams scores.

The following comments focus on the interpretation of the case: they should be used as a guide and not as the solution, which is not unique, and it is highly dependent on the reasoning presented and the hypotheses made.

The exam was based on a real case in UK:

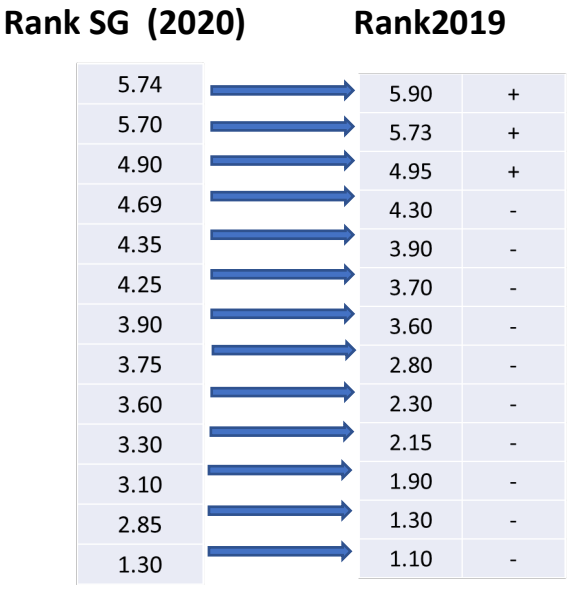
- https://www.technologyreview.com/2020/08/20/1007502/uk-exam-algorithm-cant-fix-broken-system/
- https://www.theguardian.com/commentisfree/2020/aug/19/ditch-the-algorithm-generation-students-a-levels-politics

1) Discrimination issues

The system was anchored to the past performance of a school, rather than to the past performance of an individual. This was the primary cause of the problems revealed by the audit.

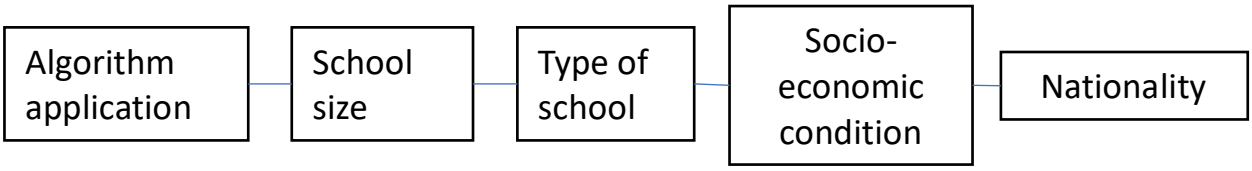
In fact, projecting RanksSG to Rank2019 is equivalent to match current data to the cumulative distribution of past marks. Imagine, as an extreme case, a Rank2019 with all A* and A: whatever is the mark distribution in RankSG, the predicted marks will be all A* and A. On the other extremity, matching RankSG to a Rank2019 with all D and E will make all students being predicted D and E. See the picture for a graphical representation of the problem.

Taking into considerations audit results nr. I (50% of students saw their mark lowered with respect to previous year), it means that Rank2019 in their schools probably had lower marks than RankSG: in practice the system does not take enough into account the variability of performances among students' generations.



Regarding audit result nr. II (decreased scores for working-class students, and increased scores for students from private/élite schools), this is explained by the fact that private/elite high schools are usually smaller and, given the procedure adopted, the automated adjustment of RankSG does not take place and only marks suggested by teachers are considered: as stated in the text, teachers tend to over-estimate grades. The dynamics of marks for schools attended by students from working-class students is explained by the same motivations of audit results nr. I.

Finally, audit result nr. III is connected to the previous one: it is very probable that private/elite schools were – in proportion- attended in majority by students from wealthy families of nationality of State X.



2) Measurement issues

Example of measurements issues are:

- The system ignores both variability within schools (classes might be very different) and between schools (ranks are neither adjusted nor compared with aggregated national results)
- Transforming data from ordinal to ratio scale means that distance between grades is the same for all pairs of grades, that might not be true: for example, the distance between A* and A might not equal to the distance between A and B.
- The fact the student performance may vary during the time is not taken into account.
- Student effective past performance is never taken into account.

B1 (2.5 points).

The phenomenon is due to Simpson's paradox. Overall average without gender distinction:
Case 1 - Total Average A: 82,5 > Total average B: 82
Case 2 - Total Average A: 82,4 > Total average B: 81,8
Case 3 - Total Average A: 93,3 > Total average B: 92,3

01URZSM DATA ETHICS AND DATA PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam 7 February 2022

A. CASE (7,5points). Coronavirus Relief Package

The Coronavirus Relief Package is a program of the U.S. Department of Health and Human Services that aims at reimbursing hospitals for part of the extra expenses or lost revenues due to the Covid-19 outbreak.

Relief funding is allocated according to an algorithm that combines the following parameters:

A) Hospital past revenue (the higher the revenue, the higher the disbursement), measured as:
   I) Hospital Operating Margin, i.e., the difference between revenues and costs related to patient care, as a proportion of the revenues (computed on a yearly base);
   II) Days of Cash On Hand. i.e., the number of days that an organization can continue to pay its operating expenses, given the amount of cash available (computed at the end of the year);

B) Hospital type (the larger the hospital, the higher the disbursement), measured as:
   I) smaller rural hospital vs large urban hospital (if located in a city with more than 1 million inhabitants);
   II) low volume hospital vs high volume hospital (depending on whether it reported an average number of daily new hospitalizations higher than 5.000 in the previous year);

C) COVID-19 burden (the higher the burden, the higher the disbursement), measured as:
   I) cumulative deaths in hospital due to covid-19 (computed on a yearly base).
   II) cumulative hospitalizations due to covid-19 (computed on a yearly base).

The first tranche of reimbursements has been given in April 2021 with data from 2020. A second tranche will be given in April 2022, based on data from 2021. However, the U.S. Department of Health and Human Services declared that the algorithm will be changed, because it has been found that the funding allocation had a disparate impact on Black population. In fact, the audit of the reimbursements given in 2021 at county-level, showed that among counties with majority of Black population received less funding.

N.B.: the exam was provided in 2 versions, according to the following combinations:

- Version 1: Variables A-I, B-I, C-I
- Version 2: Variables A-II, B-II, C-I I

Please briefly answer to the following questions (give a separate answer for each question):

1) Provide possible explanations for the results of the audit: clearly state your own hypotheses, and any other information that you suppose in addition to the provided data, in order to coherently support your reasoning (3p.)

2) Which measurement issues do you observe? (3p.)

3) Which data and/or which part of the process would you change to make the impact of the system less harmful in 2022?(1,5p.)

B (7,5 points). Briefly describe the COMPAS case, clearly indicating the discrimination issue and the cause(s).

Indications for a possible solution

A. CASE (7,5points). Coronavirus Relief Package

The following comments focus on the interpretation of the case. It should be used as a guide and not as the only possible solution, which is not unique, and it might be partially dependent on the reasoning presented and the hypotheses made (if valid). The exam, given the available time, requires synthetic (but precise and logically coherent) answers, therefore the length of the analysis reported herein should not be taken as a reference.

Variables were designed to be equivalent when assigned to each of the two exam versions, because they require very similar reasoning: the two exam versions are equivalent in terms of difficulty and evaluation.

The exam is based on a real case:

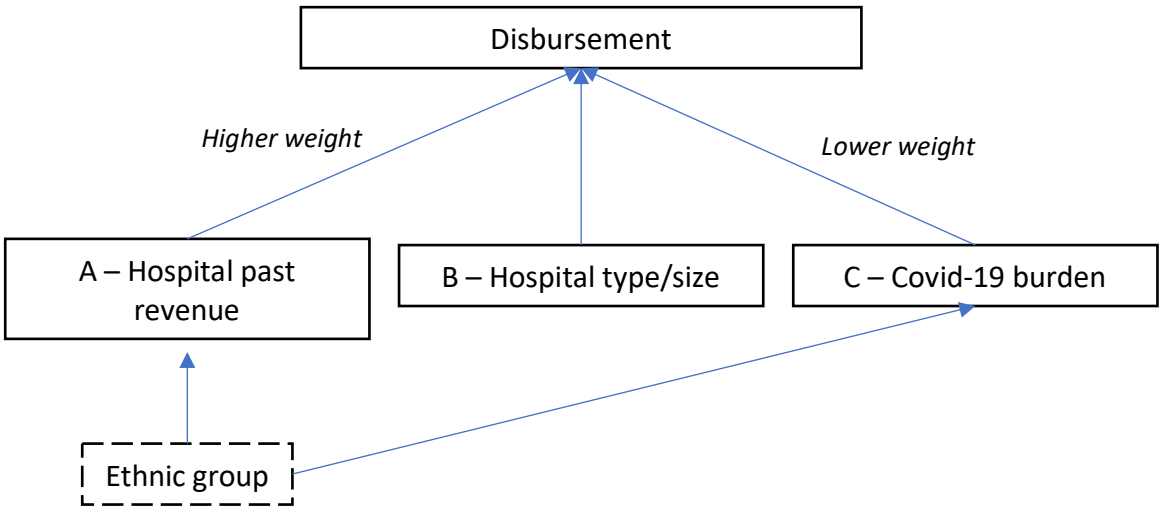
- https://www.reuters.com/business/healthcare-pharmaceuticals/insight-wealthy-hospitals-rake-us-disaster-aid-covid-19-costs-2020-12-29/
- https://www.aljazeera.com/economy/2020/12/29/wealthy-hospitals-tap-into-us-covid-relief-funds-to-cover-costs
- https://jamanetwork.com/journals/jama/fullarticle/2770395

1) Comments on the results of the audit

Given the results of the audit, it is very probable that the discrimination factor is variable A past revenue: counties with more indigent populations generate lower revenues, due to under- or no-insurance and subsequent lower level of medical treatment. Therefore, hospitals caring for them may receive less economical relief despite confronting a greater burden of COVID-19. Usually, the proportion of non-white people among the indigents in US is higher, and they are especially black people.

According to the data available, we cannot assume variable B hospital type to be a proxy variable for ethnic group.

Finally, as far as variable C is concerned, reports on the outbreak in U.S. highlighted that black people were affected much more by the disease than other ethnic groups: given the results of the audit, we deduce that the weight given to this parameter was much lower that the weight of past revenue. In general, all over the world, lower socio-economic status was highly correlated with impact of the virus in terms of deaths and serious illness.



2) Examples of measurements issues

- Using past revenues is problematic in this context because it is influenced by intensity of medicaments uses, machineries used, local market dynamics.
- The thresholds used for determining the hospital are not normalized by county population and they cannot be applied in a comparable way through the whole U.S. territory.
- COVID-19 burden measurements are not normalized by county population and they cannot be applied in a comparable way through the whole U.S. territory.

3) System improvement

Given the goal of the funding allocation program, the U.S. Department of Health and Human Services should consider aligning funding with measures of need rather than revenue, which would increase both equity and economic efficiency. Thus, beyond removing the past revenue parameter, we deem necessary to include indicators for need. The numerousness of the hospital covid-staff in relation to the population served would be an indicator of the level of organizational stress that the hospital faced. The number of non–COVID-19 deaths or non–COVID-19 serious hospitalization would also help in keeping into account the level of "ordinary" work that the hospital has/had to face during the pandemic. Finally, the adequateness of the hospital in terms of machineries and competences necessary to face the emergency is a very important factor, although not easily quantifiable.

B. Question on technology and Society (7,5points)

Check slides and video-lectures.

01TXHSM DATA ETHICS AND PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam 29 June 2020

A. CASE (7,5points). Comment on the J-Crime Prevention case.

The tool J-Crime Prevention (J is for Juvenile) provides judges with a risk indicator on young defendants (i.e. less than 18 years old). The tool is consulted by the judges who will decide which sentence to apply (prison, parole, social work, etc). The indicator estimates the probability that the person will commit a new crime within the next three years, and it labels him/her as at risk of recidivism or not. The tool has been adopted by a certain number of courts in Italy.

An independent audit has been carried out to check whether systematic discriminations occur. The researchers who made the audit used a metric called demographic disparity (DD), which is computed as PP_i / PP_r where PP is probability of being predicted positive for recidivism, respectively for members of group *i* and *r*. PP is computed as the number of defendants of the same group predicted positive for recidivism divided by the number of all defendants of that group. For example, $DD_i = 1$ means that persons of group *i* are as likely to be classified as recidivist as persons group *r*. $DD_i = 2$ means that persons of group *i* are twice as likely to be classified as recidivist as persons of group *r*. The results of the audit are: $DD_{foreigner} = 1.90$ (reference group: Italian)

The company that developed J-Crime Prevention declared that its algorithm does not use neither gender nor nationality in its model, but only the following information:

a) *Previous violent offences (yes/no)*: whether the person committed at least one violent crime previously
b) *Previous non-violent offences (yes/no)*: whether the person committed at least one non-violent crime previously
c) *Criminal parent/caregiver (yes/no)*: whether who took care of the defendant in his/her childhood had committed a crime
d) *Poor school achievement (yes/no)*: whether performance at school has been poor in at least 3 years
e) *Poor compliance (yes/no)*: whether defendants lived most of his/her life in poverty

Neither the source code nor the documentation of the algorithm implementing the model are public, hence it is not possible to know how the variables were used (e.g., their weights).

Please provide possible explanations of the discrimination reported in the audit of J-Crime Prevention. Clearly state your own hypotheses, and any other information that you suppose, in order to coherently support your reasoning.

THEORY (TOT 7,5 points). Answer briefly to the questions:

B1 (3,5 points). The following contingency table refers to results of a classifier that outputs:

- R=1 when a post on Facebook is deemed offensive and is obscured
- R=0 when a post on Facebook is not deemed offensive

The target variable is Y and it indicates when the post was truly offensive (Y=1) or not (Y=0).

Option 1

a. Comment the data in this particular context (hate speech in social networks) (1,5 points)
b. After a change in the algorithm, the table is updated: 40 more cases are added to the cell {Female, R=1 | Y=1} and 40 more cases are added to the cell {Male, R=0 | Y=0}. What will change in terms of reported fairness? Why? (2 points)

	Y=1			Y=0			
Gender	R=1	R=0	tot	R=1	R=0	tot	TOT
Female	20	40	60	20	40	60	120
Male	60	40	100	60	40	100	200
tot	80	80	160	80	80	160	320

Option 1-independence
Fairness in terms of independence is reported below:
R = 1, Female 0,33 - R = 1, Male 0,60

Option 1-separation
Fairness in terms of separation is reported below:
TP Female 0,33 - TP Male 0,60 - FP Female 0,33 - FP Male 0,60

Option 2
a. Comment the data in this particular context (hate speech in social networks) (1,5 points)
b. After a change in the algorithm, the table is updated: 40 more cases are added to the cell {Female, Y=1 | R=1} and 40 more cases are added to the cell {Male, Y=0 | R=0}. What will change in terms of reported fairness? Why? (2 points)

	R=1			R=0			
Gender	Y=1	Y=0	tot	Y=1	Y=0	tot	TOT
Female	20	20	40	40	40	80	120
Male	60	60	120	40	40	80	200
tot	80	80	160	80	80	160	320

Fairness in terms of sufficiency is reported below:
PPV Female 0,50 - PPV Male 0,50 - FOR Female 0,50 - FOR Male 0,50

B2 (2,5 points). Comment the following sentence: "Technology is shaped by existing power relations".

B3 (1,5 points). What are main allegations to Facebook for its advertising platform?

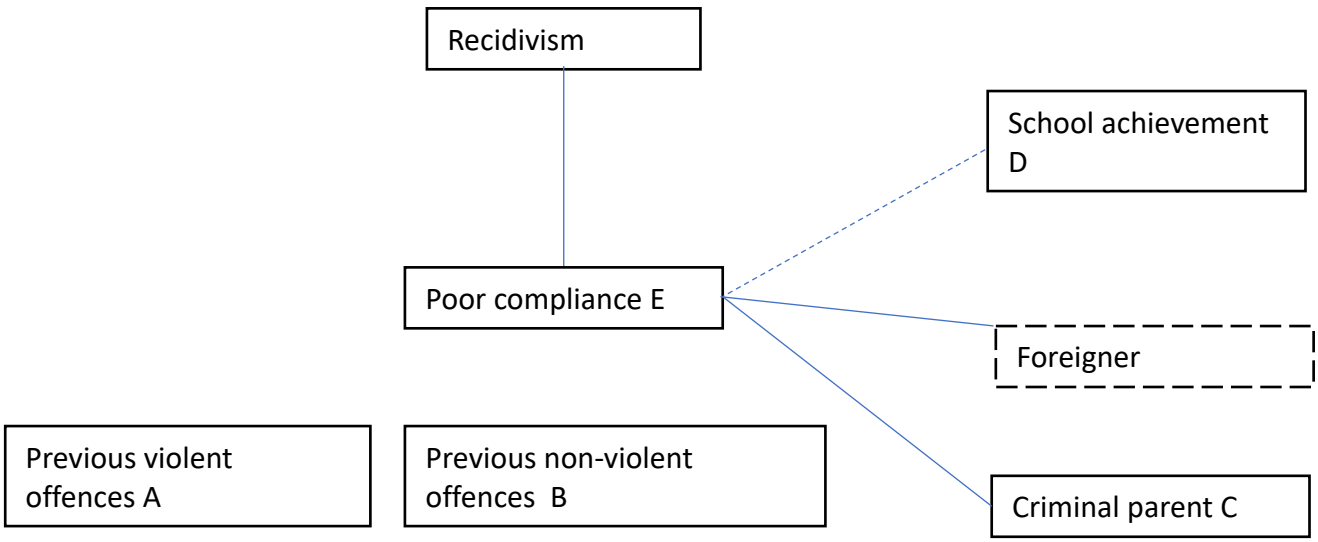
Exam 29 June 2020 – Indications for points A and B1

The following comments and analysis are not exhaustive and they focus mainly on the interpretation of the case and the fairness criteria. It should be used as a guide to the exam and not as the solution, which is not unique.

A. CASE (7,5points). Comment on the J-Crime Prevention case.

Herein we report an analysis for each variable used in the model. We assume, from the information given in the text, that the target variable is recidivism, measured as re-arrests.

- Variables A and B (A for violent crime and B for not violent crimes): since this is a dataset about young defendants, it is expected that for few of them the variables assume value 1. Consequently, we deem not sufficiently reliable the hypothesis that variables *previous violent offences* and *previous non-violent offences* have an effect on the classification of recidivism.
- Variable E (poor compliance): it is known that, on average, wealth of immigrants in Italy is lower than for Italians, and that in percentage more immigrants live in more difficult economic situations[1]. In terms of probability, it is expected that P{living in poverty | being foreigner} > P{living in poverty | being Italian}. Having set it, we also expect, for historical reasons, that poverty is a cause for criminality, including recidivism (e.g., simplifying a lot: if you make a crime because of poverty and you get punished, after punishment you are still poor, or even poorer).
- Variable C (criminal parent): in line with the previous explanation, poverty is a factor for criminality of defendants' parents/care givers.
- Variable D: In addition to the previous observations, poverty can be also connected to bad school performances, but with less strengths than other associations (available empirical evidence can support the hypothesis). Therefore, we hypothesize also for this variable a possible indirect effect on criminality and recidivism.



[1] In case you want to know more about this fact and others related to poverty in Italy, check the official report on poverty in Italy by the Italian Institute of Statistics https://www.istat.it/it/files/2022/06/Report_Povertà_2021_14-06.pdf

B. THEORY (TOT 7,5 points).

B1 (3,5 points).

Independence . b1.a)Looking at fairness in terms of independence, posts from men are more frequently tagged as offensive and removed. Since independence does not contain information on error rates, it is impossible to know whether this is because men are truly more aggressive with their language, or they use expressions that contain words flagged as offensive. b2.a Only one of the two increments will affect independence, i.e. the correctly predicted females' offensive posts, in such a way that independence will be respected: the probability that a post is labeled as offensive will be 50% for both groups.

Separation . b1.a Results show a difference of 27% between women and men both in terms of true positive than in terms of false positive, and always with higher percentage for men. This implies that men receive in general more posts removal, both correctly and not. b2.a The second situation, caused by the two changes, will make separation close to be perfectly satisfied. In fact, an increment of true positive for female will fill the gap with men, while the increment in true negative for men will increase the negative base rate for the category that will reach 140, which will make the rate of false positive lower, thus reducing the gap.
Sufficiency . b1.a The data shows that the classifier is perfectly calibrated: given that a post has been classified as offensive, it has the same probability of being made by a man or a woman. The same applies for wrong negative classifications: picking a post labeled as inoffensive, the probability of being made by a man or female is exactly 50%. b2.a Calibration won't hold anymore in the new situation: the positive predictive value will be higher for female because of the increment of true positives, while the increment of true negatives for men will lower their FOR.

***
Possible explanation: probably the algorithm has been re-tuned to take better into account the language used by women, but the new words included might have been used also by men in non-offensive posts.
For the sake of completeness, we report the changed contingency tables:

	Y=1			Y=0			
Gender	R=1	R=0	tot	R=1	R=0	tot	TOT
Female	60	40	100	20	40	60	160
Male	60	40	100	60	80	140	240
tot	120	80	200	80	120	200	400

	R=1			R=0			
Gender	Y=1	Y=0	tot	Y=1	Y=0	tot	TOT
Female	60	20	80	40	40	80	160
Male	60	60	120	40	80	120	240
tot	120	80	200	80	120	200	400

01URZSM DATA ETHICS AND DATA PROTECTION

COMPUTER SCIENCE PART (DATA ETHICS)

Exam 30 June 2023

A. CASE (7,5points). PowerSchool.

PowerSchool is a US-based company that provides software for lower secondary school (equivalent to "scuola media" in Italy) and it holds data on more than 45 million children. Some of this data is used to create risk-assessment scores aimed at predicting students' failures. It has been found that the most relevant predictors for high-risk scores were the following variables:

- Attendance in current semester: number of days
- Attendance in last year: number of days
- Behavior in current semester: number of referrals
- Behavior in last year: number of referrals
- Past graduation grade: score from 1 to 5
- Limited English Proficiency: yes/no, indicating whether English is not the primary language of the child and he/she have difficulty communicating effectively in English
- Free lunch: yes/no, whether the child has right to free lunches at school due to low income of his/her family
- Failure (target variable): yes/no

An audit for checking the tool against discrimination towards some protected attributes was performed with predictions on 2000 children. Overall, 159 failures were predicted, divided within the following subgroups:

		Total number	**Nr. Predicted at high risk**
Gender	Male	900	71
	Female	1100	88
Race	White	1200	72
	Black	240	48
	Hispanic/latino	390	28
	Asian	140	10
	American Indian	20	1
	Pacific Islander	10	0

1) **Analyze and comment the results of the audit with a fairness criterion at your choice: towards which category a relevant disparate impact of the prediction tool occurred? Why? (5 p.).**
2) **How should the table look like if results were fairer? Identify which values of the table you would change and explain the modification(s). Be sure all the numbers will be consistent with the change introduced. (2,5 p.)**

N.B. Answer concisely to the questions: give a separate answer for each point.
In doing that, clearly state your own hypotheses, and any other information that you suppose in addition to the provided data, to coherently support your line of reasoning.

***

B (7,5 points). Define the three approaches of normative ethics (3p.). Then identify which of the following three elements of the list belong to which approach, explaining why (4,5p.):
1. Fairness criteria computation; 2. Fairness qualitative assessment; 3. ACM code of ethics.

Indications for a possible solution

A. CASE (7,5points). PowerSchool.

The following comments focus on the interpretation of the case. It should be used as a guide and not as the unique solution, which might be partially dependent on the reasoning presented and the hypotheses made (if valid). The exam, given the available time, requires synthetic (but precise and logically coherent) answers, therefore the length of the analysis reported herein should not be taken as a reference.

The exam test is based on a real tool (link 1) with the problem highlighted in link 2:

1) https://www.documentcloud.org/documents/21175358-overview-presentation-school-district-u-46
2) https://www.documentcloud.org/documents/21175361-variable-weights-school-district-u-46

Answer to question 1)

The only applicable fairness criterion is Independence, because the table contains only data about high-risk predictions (R=1): to compute it, it is sufficient to compute the ratio Nr. Predicted at high risk/ Total number.

		Total number	**Nr. Predicted at high risk**	**P(R=1)**
Gender	Male	900	71	8%
	Female	1100	88	8%
Race	White	1200	72	6%
	Black	240	48	20%
	Hispanic/latino	390	28	7%
	Asian	140	10	7%
	American Indian	20	1	5%
	Pacific Islander	10	0	0%

As it can be noticed from the results, Black children are predicted as high risk with more than double percentage than children of other races. Based on the inequality patterns explained in the lectures (slideset demographic disparities), the most relevant proxies is "free lunch": it is a welfare benefit for low-income families, which are composed with higher percentage of black people. Poor school performances in US can also be easily tracked back to difficult socio-economic conditions, therefore also to black category. Results on Pacific islander children can be ignored due to very small sample size.

Answer to question 2)

A fairer prediction would result in a high-risk prediction rate for black like the other races: for example, by changing the number of predicted high risks for black to 20, the ratio will be 8%, which is very similar to the other races. To keep the total numbers consistent, male and females' high risks should be decreased accordingly: for example, respectively to 50 and 81. Table is reported below, with changed values in red.

		Total number	**Nr. Predicted at high risk**
Gender	Male	900	50
	Female	1100	81
Race	White	1200	72
	Black	240	20
	Hispanic/latino	390	28
	Asian	140	10
	American Indian	20	1
	Pacific Islander	10	0

B. Question on theory (7,5points)

Check slides and video-lectures to check definitions and understand the correct mapping, which was: **Fairness criteria computation → consequentialism, Fairness qualitative assessment → virtue ethics, ACM code of ethics → deontology**.

Data Ethics and Protection

Exam_2023-07-14_Part_2

	[redacted]
Iniziato	venerdì, 14 luglio 2023, 15:10
Terminato	venerdì, 14 luglio 2023, 16:00
Tempo impiegato	50 min.
Valutazione	**6,50** su un massimo di 15,00 (**43**%)

Domanda 1

Completo

Punteggio ottenuto 2,00 su 7,50

Soft loans.

Each year, Rotterdam, a city in the Netherlands, provides “soft” loans to approximately 10.000 individuals to support their rent payments, purchase food, and cover essential bills. Soft loans involve small quantity of money (<2.000 euros), do not accumulate interests, and they should be paid back after 5 years. Despite the favorable conditions, a certain number of these beneficiaries is not able to pay back their debt.

In 2022, the city adopted a machine learning model to generate a prediction of insolvency (0/1) for each new applicant to the soft loan program: the loan is denied when a prediction of insolvency is 1. The model was trained with a subset of historical data of past beneficiaries (2015-2019, to exclude the socio-economic effects of the pandemic) on 10.000 individuals, 75% of which were randomly picked from a subset of beneficiaries whose characteristics correspond to the most recurrent statistical patterns of insolvencies in the previous years (called “archetypes”). The “archetypes” used are: “Financially Struggling Single Mother”, “Migrant Worker”, “Lack of advanced educational skills”.
The remaining 25% were randomly selected from the whole set.

The list below contains the variables (and their possible values) used to train the model: the demographic variables were directly retrieved from the city databases; the other variables were extracted from the periodic surveys that beneficiaries of welfare benefits must fill in.

A. Gender: Man / Woman
B. Age: <18 / 18-50 / 50+
C. Children: Yes/No
D. Neighborhood: 15 possible neighborhoods

E. Days lived at current address: Integer value >= 0
F. Proficiency in Dutch language: Yes/No
G. Days with financial difficulties: 0 days / 250 days / 700 days or more
H. Days in current marriage: 0 days / 360 days / 720 days or more
I. Number of advanced competences: 0 / 1 / 2 / 3 or more
J. Verified insolvency (TARGET VARIABLE): 0/1 (=No/Yes)

A risk analysis was carried to determine whether the algorithm could potentially discriminate the groups protected by the Netherlands' General Equal Treatment Act, which prohibits direct and indirect discrimination based on race, gender, nationality, sexual orientation, or marital status. The analysis found that there is no or very low risk of discrimination only for **one** of those protected groups.

1) Which was the group with no/very low risk of systematic discrimination? Why? (5 p.).

N.B. a full analysis of the features or of the protected groups is NOT mandatory

2) Identify and explain one limitation of the whole measurement process. (2,5 p.)

N.B. Answer concisely to the questions: give a separate answer for each point.
In doing that, clearly state your own hypotheses, and any other information that you sup pose in addition to the provided data, to coherently support your line of reasoning.

---

1)

I think sexual orientation here is not considered, I mean third gender or LGBTQ that are single and do not have kids or their marriage is not official... I can say that might be a representation bias because men and women are just there in the gender part. they are now in minority think that not even be specified in the gender options.

2)

I think having children or not is not a good measure maybe the number of children should be mentioned to find their difficulties.

age intervals also better to be in smaller borders in the 18-50 years. and days of being in a marriage is not related to the target variable. maybe people with more years of marriage have more financial issues than younger ones.

Commento:
1) 1/5 vague and not clear answer, missing specific explanation mechanisms

2) 1

Only ONE measurement issue was asked !

- "I think having children or not is not a good measure maybe the number of children should be mentioned to find their difficulties".--> why find their difficulties as a goal of the variable ? -0,25-

-"age intervals also better to be in smaller borders in the 18-50 years." --> why ? -0,5

- "and days of being in a marriage is not related to the target variable" --> vague/not explained -0,5

- "maybe people with more years of marriage have more financial issues than younger ones. " : why ? -0,25

Domanda 2

Completo

Punteggio ottenuto 4,50 su 7,50

What is the main difference between historical bias and measurement bias? (3,5p).
Make one example of historical bias, and one example of measurement bias (4 p.).
N.B. Be synthetic.

---

historical bias is happend from potential space to construct space. is something from the past that now might effect the mind something like talking about afghanistan girls. when the target is related to a protected attribute. or we can think about the bias of gender pay gap between women and men. or the fact that women are more responsible of take care of children. and the target is to measure the time that employees might be absent.

measurement bias is happend from construct space to observed space. when happend when in variables some data are missing or we have percentages and borders that might effect some people in between on the threshhold that might become discriminated. like setting the score for giving a loan to people and one variable is the work experience and the interval is <1 20 points,and 1-4, 40 points and >4 years, 60 points . and you would get more score if you have more experience. now this interval is not ok because between 4 years of experience and 5 years of experience is a big difference of points you get. there should be a better mearement. or for age and when you might open your account. if you are younge age should be normalized.

Commento:

1) 2,5/3,5

some imprecisions in the definitions of m.bias (e.g. proxy mechanism in measurement bias missing)

2) /4

2.1) 2/2 OK

2.2) 0/2 NO