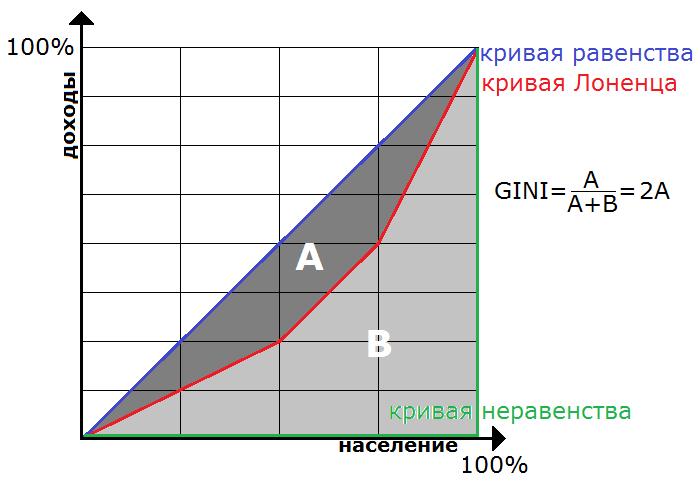
**Знакомьтесь, Джини**

Многие путаются в **коэффициентах Джини**, не понимают, что они бывают разные и для разных задач (и названия у них разные — просто в русском переводе, как всегда, многое схлопывается в один термин).

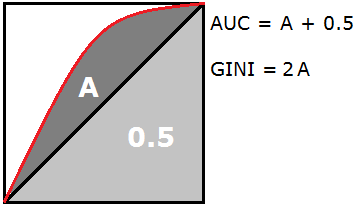
Есть коэффициент/индекс Джини (**Gini coefficient**), который используют при [оценке качества классификации и регрессии](https://www.kaggle.com/wiki/Gini). На [русской странице Wiki](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D1%8D%D1%84%D1%84%D0%B8%D1%86%D0%B8%D0%B5%D0%BD%D1%82_%D0%94%D0%B6%D0%B8%D0%BD%D0%B8) не очень информативно, но вот [на английской](https://en.wikipedia.org/wiki/Gini_coefficient) всё подробно: изначально это был *статистический показатель степени расслоения общества данной страны или региона по отношению к какому-либо изучаемому признаку. Вычисляется как отношение площади фигуры, образованной кривой Лоренца и кривой равенства, к площади треугольника, образованного кривыми равенства и неравенства*. Сейчас поясню.

Допустим, в компании работают 4 человека с суммарным доходом 8000$. Равномерное распределение дохода — это 2000$+2000$+2000$+2000$, неравномерное — 0$+0$+0$+8000$. А как оценить неравномерность, скажем, для случая 1000$+1000$+2000$+4000$? Упорядочим сотрудников по возрастанию дохода. Построим кривую (Лоренца) в координатах [процент населения, процент дохода этого населения] — идём по всем сотрудникам и откладывает точки. Для первого — [25%, 12.5%] — это сколько он составляет процентов от всего штата и сколько процентов составляет его доход, для первого и второго — [50%, 25%] — это сколько они составляют процентов и сколько процентов их доход, для первых трёх — [75%, 50%], для всех — [100%, 100%].

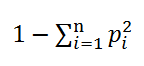
Рис. 1. Вычисление gini с помощью кривой Лоренца

На. Рис. 1. построенная кривая Лоренца показана красным цветом. Кривая Лоренца, которая соответствует равномерному распределению дохода, — синяя диагональ (т.н. кривая равенства). Кривая Лоренца, которая соответствует неравномерному распределению, — зелёная (т.н. кривая неравенства). Вот площадь A, делённая на A+B=0.5, и есть коэффициент Gini.

При оценке качества классификации GINI = 2\*AUCROC-1. Про AUCROC я уже как-то писал. Почему это они так связаны **нигде** подробно не описано. Я нашёл упоминание в работе [Supervised Classification and AUC](https://www.rocq.inria.fr/axis/modulad/Workshop_Franco_Bresilien/programme/niang.pdf" \t "_blank). Там всё логично: если в задаче классификации на два класса 0 и 1 интерпретировать эти числа как доходы. Но чтобы связь была именно GINI = 2\*AUCROC-1, должно быть что-то типа рис. 2 (но ROC-кривая и кривая Лоренца это не одно и то же), кстати в презентации [Credit Scoring and the Optimization concerning Area under the curve](https://alexanderdyakonov.wordpress.com/2015/12/15/%d0%b7%d0%bd%d0%b0%d0%ba%d0%be%d0%bc%d1%8c%d1%82%d0%b5%d1%81%d1%8c-%d0%b4%d0%b6%d0%b8%d0%bd%d0%b8/www.business-school.ed.ac.uk/waf/crc_archive/2013/18.pdf" \t "_blank) такая же картинка.

Рис.2. Связь AUCROC и GINI.

Есть ещё коэффициент/индекс Джини (**Gini impurity**), который используется в решающих деревьях [при выборе расщепления](https://en.wikipedia.org/wiki/Decision_tree_learning#Gini_impurity). Я дал ссылку на английскую Wiki, поскольку русского аналога нет. Он тоже измеряет «равномерность», если p\_i — частоты представителей разных классов в листе дерева, то коэффициент Джини для него равен



Только вот это другая равномерность, никак не связанная с рассмотренной ранее. Для первой нужно два показателя — доход и численность населения с таким доходом, а тут только проценты (частоты). В английской версии на странице Gini coefficient написано «не путать с Gini impurity» и наоборот.

Я не знаю, как лучше переводить **impurity**, скажем, [С.П.Чистяков](https://www.google.ru/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwi7qd2Al9TJAhVLwHIKHcogDykQFggbMAA&url=http%3A%2F%2Fwww.krc.karelia.ru%2Fdoc_download.php%3Fid%3D5876%26table_name%3Dsection%26table_ident%3D957&usg=AFQjCNF7PoXItfEohhBrW6-DrhPbFJXkEw&bvm=bv.109910813,d.bGQ" \t "_blank)переводит как «загрязненность» (на мой взгляд, не очень звучит…).

**Коррадо Джини** (Corrado Gini, 1884), который всё это придумал был итальянским статистиком. Но кроме этого, он известный идеолог фашизмa, написал книгу «Научные основы фашизма». Прожил, кстати, довольно много — 80 лет, видимо, после войны не преследовался. Вот так бывает…