Eksploracja hurtowni danych:

1. Cel biznesowy eksploracji danych:

Poszukiwanie który z kryteriów ma największą wagę przy klastrowaniu za grupami:

a. „severity_type”,
b. „intersection_id”

2. Typ predykcji:

Selekecja najwazniejszych atrybutów dla wyznaczenia który z nich jest najważniejszy przy klastrowaniu na różne grupy.

3. Wybrana metoda:

Będziemy korzystali z algorytmu losowych lasów. Korzystamy same z nich a nie z drzew decyzyjnych, z powodu że losowe lasy decyzyjne korygują nawyk drzew decyzyjnych do nadmiernego dopasowania do ich zbioru treningowego.

Losowe lasy mają wiele zalet:

- Niewrażliwość na skalowanie (i ogólnie na wszelkie przekształcenia monotoniczne) wartości cech.
- Zarówno cechy ciągłe, jak i dyskretne są przetwarzane równie dobrze. Istnieją metody konstruowania drzew z danych z brakującymi wartościami cech.
- Istnieją metody szacowania istotności poszczególnych cech w modelu.
- Wewnętrzna ocena zdolności modelu do uogólniania (test na niewyselekcjonowanych próbach).
- Wysoka równoległość i skalowalność .

Oprócz tych wszystkich zalet, mamy jeszcze jedną bardzo ważliwą przyczynę – z 9 metod które są dostępne w „SSAS”, w naszym przypadku drzewa decyzyjne są najlepszym wyborem, a żeby otrzymać lepsze wyniki – wykorzystujemy lasy losowe.

4. Wykonanie analizy:

Konfiguracja losowego lasu:

- Liczba_drzew: 50
- Maksymalna głębokość: 10
- Główny atrybut: Severity_type
- Kryterium: Współczynnik wzmocnienia
- Strategia głosowania: wotum zaufania

Eksperyment 1 („severity_type”):

W pierwszym przypadku będziemy rozglądać następne pola naszej hurtowni:

- Severity_type
- Pedestrian_id
- Intersection_id

- Day_of_week

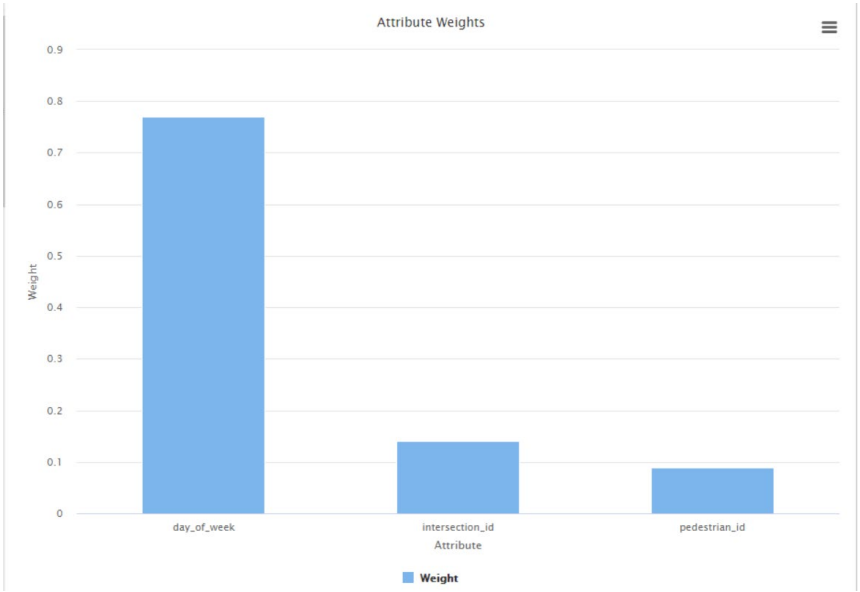
Gdzie „Severity_type” – pole za które będzie zmianiać.

Wykorzystujemy same Pedestrian_id oraz Intersection_id z powodu że tabela „Crash” już zawiera te atrybuty a mają one tylko 2 stany: „T” czy „F”, dlatego żeby nie robić złączeń z innymi tabelamy wykorzystujemy atrybuty które już są w tabeli faktu.

Wyniki z programu:

Wagi atrybutów:

attribute	weight
pedestrian_id	0.089
intersection_id	0.141
day_of_week	0.770



Jedno z 50 drzew którę będziemy analizować:

```
intersection_id = F
|   pedestrian_id = F
|   |   day_of_week > 6.500: minor_injury {property_damage=5360, serious_injury=6237,
minor_injury=7829, fatality=521}
|   |   day_of_week ≤ 6.500
|   |   |   day_of_week > 5.500: minor_injury {property_damage=7169, serious_injury=6413,
minor_injury=8918, fatality=561}
|   |   |   day_of_week ≤ 5.500
|   |   |   |   day_of_week > 1.500
|   |   |   |   |   day_of_week > 3.500
|   |   |   |   |   |   day_of_week > 4.500: minor_injury {property_damage=9783,
serious_injury=5977, minor_injury=10167, fatality=479}
|   |   |   |   |   |   day_of_week ≤ 4.500: minor_injury {property_damage=9131,
serious_injury=5516, minor_injury=9733, fatality=418}
|   |   |   |   |   day_of_week ≤ 3.500
|   |   |   |   |   |   day_of_week > 2.500: minor_injury {property_damage=8816,
serious_injury=5117, minor_injury=9584, fatality=358}
|   |   |   |   |   |   day_of_week ≤ 2.500: minor_injury {property_damage=8273,
serious_injury=4904, minor_injury=9434, fatality=377}
|   |   |   |   day_of_week ≤ 1.500: minor_injury {property_damage=7633, serious_injury=4972,
minor_injury=8665, fatality=409}
|   pedestrian_id = T
|   |   day_of_week > 1.500
|   |   |   day_of_week > 6.500: property_damage {property_damage=62, serious_injury=50,
minor_injury=46, fatality=7}
|   |   |   day_of_week ≤ 6.500
|   |   |   |   day_of_week > 5.500: property_damage {property_damage=106, serious_injury=68,
minor_injury=48, fatality=15}
|   |   |   |   day_of_week ≤ 5.500
|   |   |   |   |   day_of_week > 2.500
|   |   |   |   |   |   day_of_week > 4.500: property_damage {property_damage=143,
serious_injury=82, minor_injury=97, fatality=9}
|   |   |   |   |   |   day_of_week ≤ 4.500
```

```
|   |   |   |   |   |   day_of_week > 3.500: property_damage {property_damage=108,
serious_injury=66, minor_injury=56, fatality=7}
|   |   |   |   |   |   day_of_week ≤ 3.500: property_damage {property_damage=116,
serious_injury=77, minor_injury=61, fatality=7}
|   |   |   |   |   day_of_week ≤ 2.500: property_damage {property_damage=104, serious_injury=70,
minor_injury=59, fatality=14}
|   |   day_of_week ≤ 1.500: property_damage {property_damage=115, serious_injury=54,
minor_injury=49, fatality=11}
intersection_id = T
|   pedestrian_id = F
|   |   day_of_week > 6.500: minor_injury {property_damage=5404, serious_injury=3909,
minor_injury=7462, fatality=129}
|   |   day_of_week ≤ 6.500
|   |   |   day_of_week > 5.500: minor_injury {property_damage=7734, serious_injury=4608,
minor_injury=9209, fatality=190}
|   |   |   day_of_week ≤ 5.500
|   |   |   |   day_of_week > 3.500
|   |   |   |   |   day_of_week > 4.500: minor_injury {property_damage=10763,
serious_injury=4884, minor_injury=12367, fatality=199}
|   |   |   |   |   day_of_week ≤ 4.500: minor_injury {property_damage=10501,
serious_injury=5048, minor_injury=12022, fatality=207}
|   |   |   |   day_of_week ≤ 3.500
|   |   |   |   |   day_of_week > 1.500
|   |   |   |   |   |   day_of_week > 2.500: minor_injury {property_damage=10064,
serious_injury=4436, minor_injury=11713, fatality=158}
|   |   |   |   |   |   day_of_week ≤ 2.500: minor_injury {property_damage=9871,
serious_injury=4550, minor_injury=11479, fatality=159}
|   |   |   |   |   day_of_week ≤ 1.500: minor_injury {property_damage=8959, serious_injury=4314,
minor_injury=10672, fatality=137}
|   pedestrian_id = T
|   |   day_of_week > 3.500
|   |   |   day_of_week > 5.500
|   |   |   |   day_of_week > 6.500: property_damage {property_damage=42, serious_injury=26,
minor_injury=25, fatality=0}
|   |   |   |   day_of_week ≤ 6.500: property_damage {property_damage=40, serious_injury=28,
minor_injury=24, fatality=0}
|   |   |   day_of_week ≤ 5.500
|   |   |   |   day_of_week > 4.500: property_damage {property_damage=59, serious_injury=37,
minor_injury=42, fatality=3}
|   |   |   |   day_of_week ≤ 4.500: minor_injury {property_damage=46, serious_injury=27,
minor_injury=49, fatality=5}
|   |   day_of_week ≤ 3.500
|   |   |   day_of_week > 2.500: property_damage {property_damage=54, serious_injury=19,
minor_injury=16, fatality=0}
|   |   |   day_of_week ≤ 2.500
|   |   |   |   day_of_week > 1.500: property_damage {property_damage=69, serious_injury=25,
minor_injury=35, fatality=1}
|   |   |   |   day_of_week ≤ 1.500: property_damage {property_damage=52, serious_injury=24,
minor_injury=26, fatality=0}
```

W naszym zbiorze danych mamy:

	[Liczba rekordów w fakcie]
1	336383

I mamy taki podział severety na tą liczbę:

	severity_type	[Liczba rekordów]
1	fatality	4388
2	minor_injury	139850
3	property_damage	120739
4	serious_injury	71406

Z tego możemy dowiedzieć że zawszę najmniej będzie wypadków z „severity_type = fatality” oraz będzie nie dużo wypadkó z „serious_injury”. To z tego już rozumiemy, że w większości wypadków będziemy mieć wypadki z „minory_injury” czy z „property_damage”.

Legenda:

Niebieskim kolorem – odznaczone wypadki z „Property damage”

Żółtym kolorem – odznaczone wypadki z „Minory injury”

Zielonym kolorem – odznaczone wypadki z „Serious injury”

Czerwonym kolorem – odznaczone wypadki z „Fatality”

Im większa kolumna pionowo – im więcej rekordów zostało przyznaczono do tego wypadku.

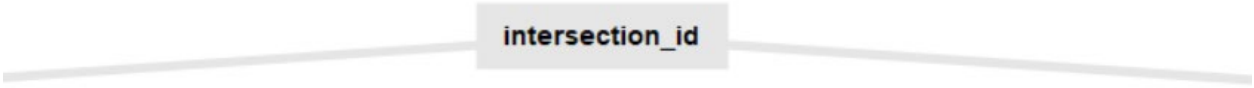
Im większa kolumna poziomowo – im więcej rekordów zostało pryznaczono temu koloru(surowości)

Rozpoczniemy analizę:

Mamy takie kroki w naszym drzewie:

Intersection_id[T|F] - > Pedestrian_id[T|F] - > Day_of_week[1-7]

Jak można zobaczyć wyżej na początku sprawdzamy, czy wypadek był na skrzerzowaniu czy nie:



Od grubości zależy ile rekordów idzie do którejś z stron, w naszym przypadku:

	intersection_id	[Liczba rekordów]
1	T	171727
2	F	164656

Możemy zobaczyć że rozdzieliło nam nasze wszystkie rekordy na poł do każdego z wariantów.

Dalej mamy rozdzielenie za pedestrian_id. Z powodu że mamy dużo mniej rekordów gdzie uczestniczą piesze – zawsze będzie więcej rekordów gdzie „pedestrian_id = F”:

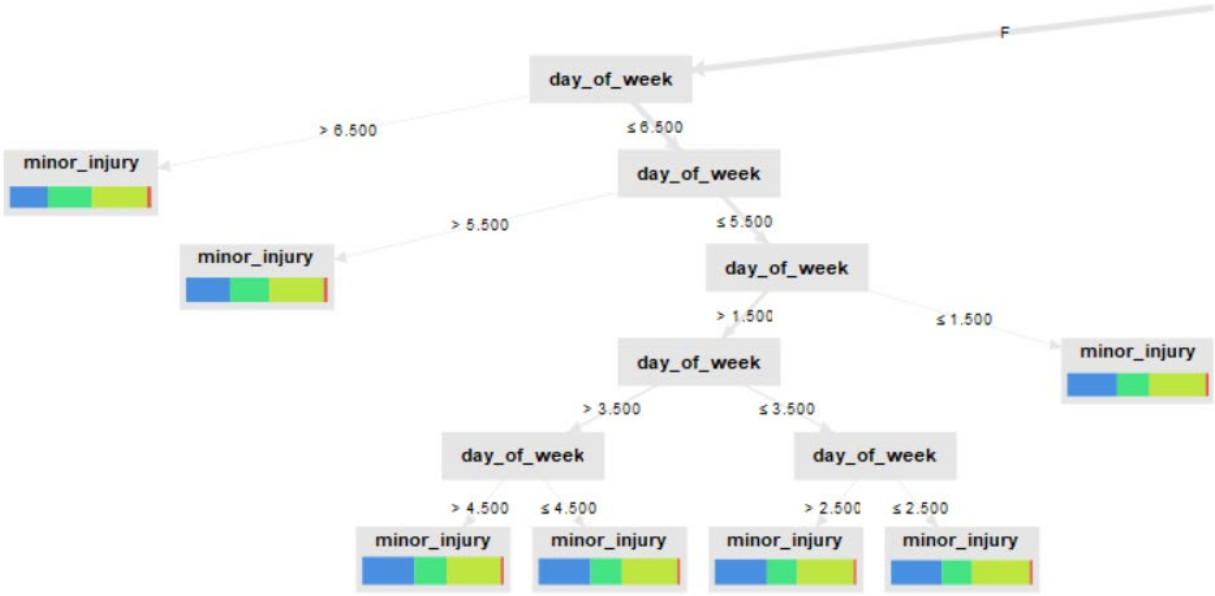
	pedestrian_id	[Liczba rekordów]
1	T	2433
2	F	333950

Zobaczyć to można na zdjęciu wyżej, rekordów z „pedestrian_id = T” mniej niż 1%.

Zaraz rozważamy gałąź: intersection(F) - > pedestrian_id(F)

```
intersection_id = F
|   pedestrian_id = F
|   |   day_of_week > 6.500: minor_injury {property_damage=5360, serious_injury=6237,
minor_injury=7829, fatality=521}
|   |   day_of_week ≤ 6.500
|   |   |   day_of_week > 5.500: minor_injury {property_damage=7169, serious_injury=6413,
minor_injury=8918, fatality=561}
|   |   |   day_of_week ≤ 5.500
|   |   |   |   day_of_week > 1.500
|   |   |   |   |   day_of_week > 3.500
|   |   |   |   |   |   day_of_week > 4.500: minor_injury {property_damage=9783,
serious_injury=5977, minor_injury=10167, fatality=479}
```

```
|   |   |   |   |   |   day_of_week ≤ 4.500: minor_injury {property_damage=9131,
serious_injury=5516, minor_injury=9733, fatality=418}
|   |   |   |   |   day_of_week ≤ 3.500
|   |   |   |   |   |   day_of_week > 2.500: minor_injury {property_damage=8816,
serious_injury=5117, minor_injury=9584, fatality=358}
|   |   |   |   |   |   day_of_week ≤ 2.500: minor_injury {property_damage=8273,
serious_injury=4904, minor_injury=9434, fatality=377}
|   |   |   |   day_of_week ≤ 1.500: minor_injury {property_damage=7633, serious_injury=4972,
minor_injury=8665, fatality=409}
```



Najciekawszę tu jest to, że dni od 5.5 (Połowa piątku) i do niedzieli włącznie mają najwięcej wypadków z surowością „serious_injury”:

- Bezwzględna liczba wypadków z surowością „serious_injury”, z lewa na prawo(ser – „serious_injury”, fat – „fatality”)

1. Ser = 0.312, fat = 0.026
2. Ser = 0.278, fat = 0.024
3. Ser = 0.226, fat = 0.018
4. Ser = 0.222, fat = 0.0168
5. Ser = 0.213, fat = 0.0163
6. Ser = 0.229, fat = 0.018

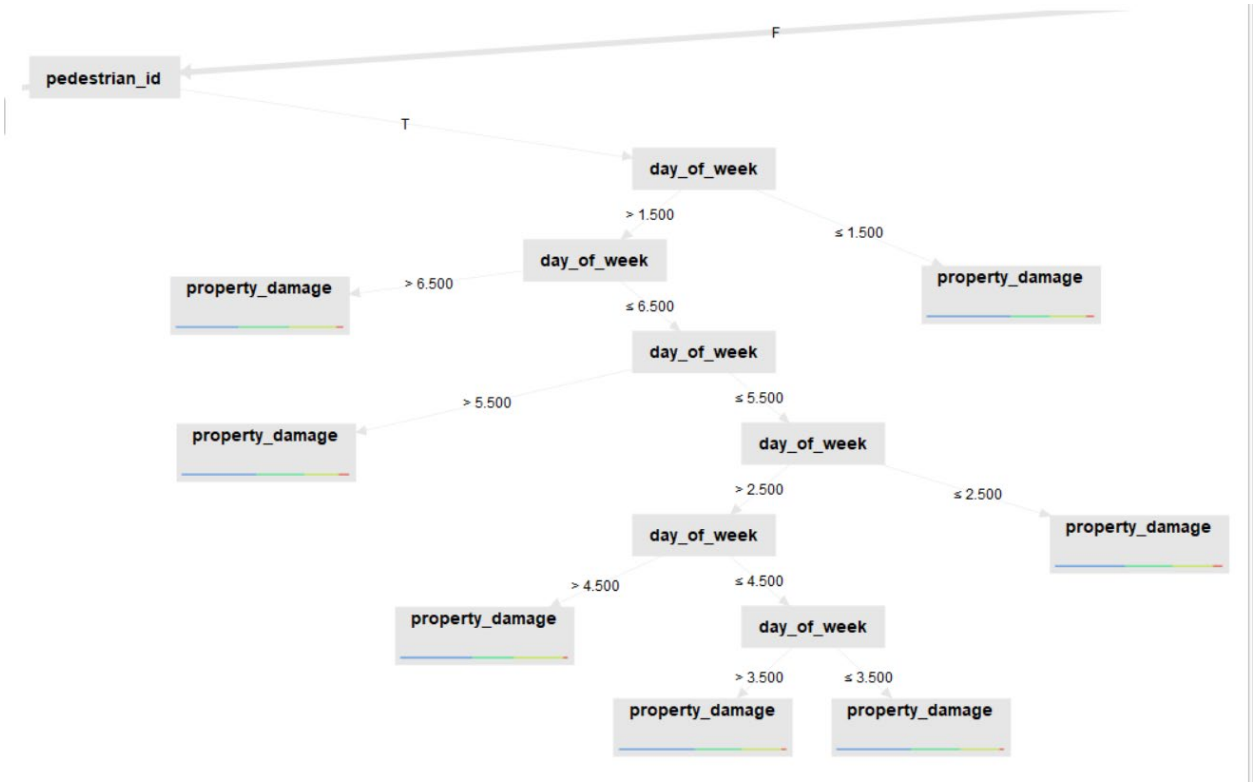
Z tego już możemy stwierdzić że bezwzględnie wypadki bez uczestnictwa pieszych, oraz nie na skrzyrzowaniu najbardziej niebiezpieczni od południa piątku i do niedzieli, dlatego że mają więcej wypadkó z surowościamy „serious_injury” oraz „fatality”.

Ale dlaczego tak jest? Moim zdaniem jedną z przyczyn może być to, że wiele ludzi w weekendy wyjeżdżają za miasto, tam oni na jadą z większą prędkością, niestety za takich umow wyadki mogą być bardziej niebiezpieczne.

Dalej rozważamy gałąź: intersection(F) - > pedestrian_id(T)

```
intersection_id = F
|   pedestrian_id = T
|   |   day_of_week > 1.500
|   |   |   day_of_week > 6.500: property_damage {property_damage=62, serious_injury=50,
minor_injury=46, fatality=7}
|   |   |   day_of_week ≤ 6.500
```

```
|   |   |   |   day_of_week > 5.500: property_damage {property_damage=106, serious_injury=68,
minor_injury=48, fatality=15}
|   |   |   |   day_of_week ≤ 5.500
|   |   |   |   |   day_of_week > 2.500
|   |   |   |   |   |   day_of_week > 4.500: property_damage {property_damage=143,
serious_injury=82, minor_injury=97, fatality=9}
|   |   |   |   |   |   day_of_week ≤ 4.500
|   |   |   |   |   |   |   day_of_week > 3.500: property_damage {property_damage=108,
serious_injury=66, minor_injury=56, fatality=7}
|   |   |   |   |   |   |   day_of_week ≤ 3.500: property_damage {property_damage=116,
serious_injury=77, minor_injury=61, fatality=7}
|   |   |   |   |   day_of_week ≤ 2.500: property_damage {property_damage=104, serious_injury=70,
minor_injury=59, fatality=14}
|   |   day_of_week ≤ 1.500: property_damage {property_damage=115, serious_injury=54,
minor_injury=49, fatality=11}
```



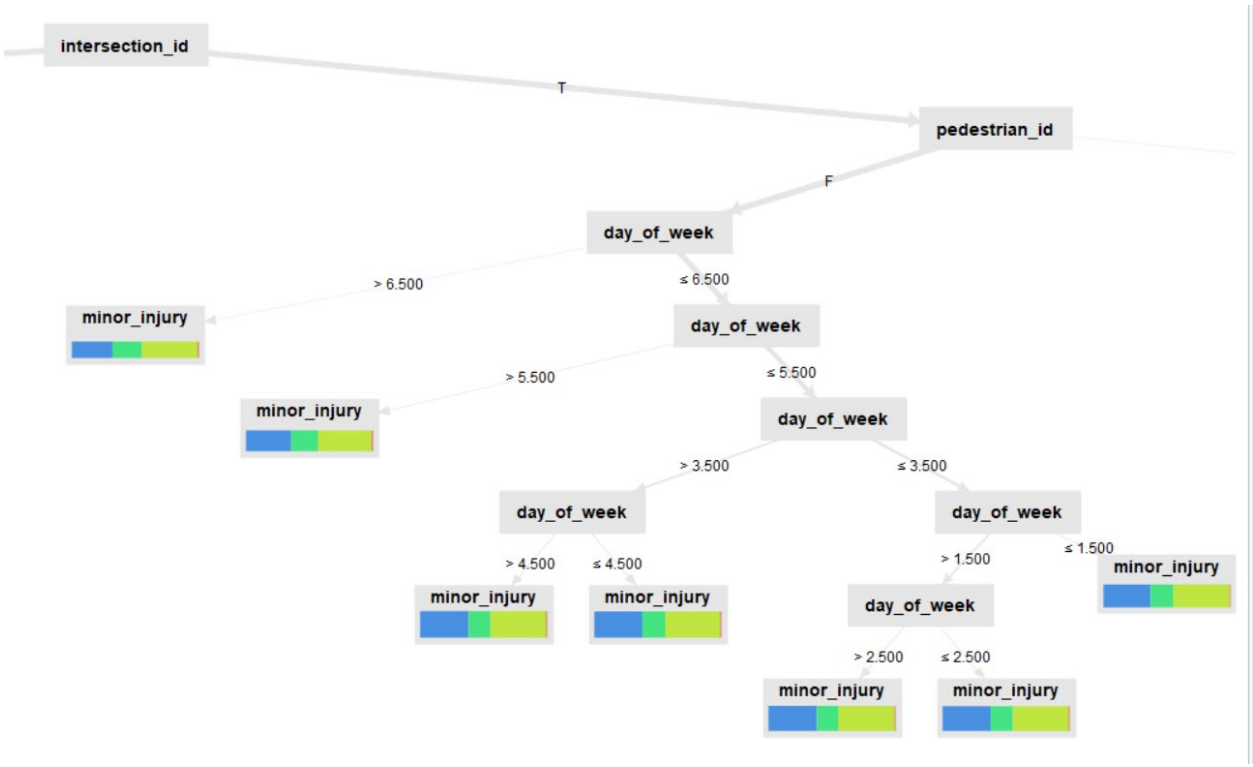
Już od razu jest widoczne że kolumny pionowo są dużo mniejsze – to jak już było powiedziano, jest z powodu że przepisano tym wypadkam mniej rekordów.

Tutaj wszędzie jest więcej wypadków z „Property damage”, odwrotnie do przerzłego zdjęcia, gdzie wszędzie więcej „minory injury”.

Zaraz rozważamy gałąź: intersection(T) - > pedestrian_id(F):

```
intersection_id = T
|   pedestrian_id = F
|   |   day_of_week > 6.500: minor_injury {property_damage=5404, serious_injury=3909,
minor_injury=7462, fatality=129}
|   |   day_of_week ≤ 6.500
|   |   |   day_of_week > 5.500: minor_injury {property_damage=7734, serious_injury=4608,
minor_injury=9209, fatality=190}
|   |   |   day_of_week ≤ 5.500
|   |   |   |   day_of_week > 3.500
|   |   |   |   |   day_of_week > 4.500: minor_injury {property_damage=10763,
serious_injury=4884, minor_injury=12367, fatality=199}
```

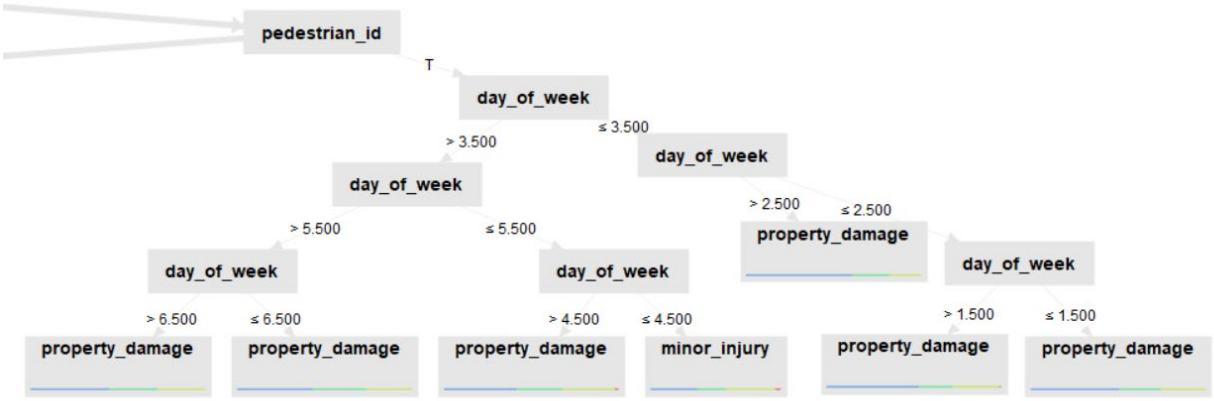
```
|   |   |   |   |   day_of_week ≤ 4.500: minor_injury {property_damage=10501,
serious_injury=5048, minor_injury=12022, fatality=207}
|   |   |   |   day_of_week ≤ 3.500
|   |   |   |   |   day_of_week > 1.500
|   |   |   |   |   |   day_of_week > 2.500: minor_injury {property_damage=10064,
serious_injury=4436, minor_injury=11713, fatality=158}
|   |   |   |   |   |   day_of_week ≤ 2.500: minor_injury {property_damage=9871,
serious_injury=4550, minor_injury=11479, fatality=159}
|   |   |   |   |   day_of_week ≤ 1.500: minor_injury {property_damage=8959, serious_injury=4314,
minor_injury=10672, fatality=137}
```



Tutaj mamy bardzo podobne wyniki do gałęzi intersection(F) - > pedestrian_id(T), tak samo mamy tutaj, że w okres od drugiej połowy piątku i do niedzieli to jest najbardziej niebiezpieczny czas do podrorzy na drogie.

Dalej rozważamy gałąź: intersection(T) - > pedestrian_id(T):

```
Intersection_id = T
|   pedestrian_id = T
|   |   day_of_week > 3.500
|   |   |   day_of_week > 5.500
|   |   |   |   day_of_week > 6.500: property_damage {property_damage=42, serious_injury=26,
minor_injury=25, fatality=0}
|   |   |   |   day_of_week ≤ 6.500: property_damage {property_damage=40, serious_injury=28,
minor_injury=24, fatality=0}
|   |   |   day_of_week ≤ 5.500
|   |   |   |   day_of_week > 4.500: property_damage {property_damage=59, serious_injury=37,
minor_injury=42, fatality=3}
|   |   |   |   day_of_week ≤ 4.500: minor_injury {property_damage=46, serious_injury=27,
minor_injury=49, fatality=5}
|   |   day_of_week ≤ 3.500
|   |   |   day_of_week > 2.500: property_damage {property_damage=54, serious_injury=19,
minor_injury=16, fatality=0}
|   |   |   day_of_week ≤ 2.500
|   |   |   |   day_of_week > 1.500: property_damage {property_damage=69, serious_injury=25,
minor_injury=35, fatality=1}
|   |   |   |   day_of_week ≤ 1.500: property_damage {property_damage=52, serious_injury=24,
minor_injury=26, fatality=0}
```



Z wyników wyżej widocznie jest że najwięcej wypadków z „property_damage” mamy we wtorki w drugiej połowie dnia. I co dla mnie jest najbardziej interesującę w tych wynikach – że w okresie od środy i do drugiej połowy czwartku mamy najwięcej wypadków same z surowością „minory_injury”.

Wnioski z tego eksperumentu 1:

Za pomocą lasa losowego otrzymaliśmy bardzo ważną i ciekawą informację – w soboty i niedzieli mamy najwięcej wypadków z ciężkiej surowością, największe w soboty i niedzieli wypadków z surowością „serious_injury” oraz „fatality”. Przyjęli że to może być zpowodu, że wiele ludzi w weekendy wyjeżdżają za miasto, tam oni na jadą z większą prędkością, niestety za takich umow wyadki mogą być bardziej niebiezpieczne.

Eksperyment 2 („intersection_id”):

W drugim przypadku będziemy rozglądać następne pola naszej hurtowni:

- Severity_type
- Pedestrian_id
- Intersection_id
- Day_of_week

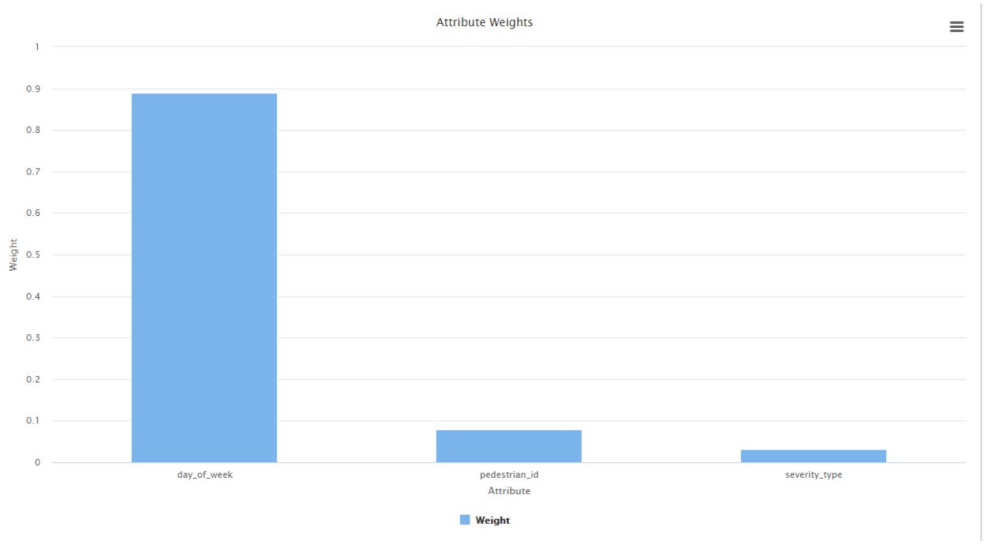
Gdzie „intersection_id” – pole za które będzie zmianiać.

Wykorzystujemy same Pedestrian_id oraz Intersection_id z powodu że tabela „Crash” już zawiera te atrybuty a mają one tylko 2 stany: „T” czy „F”, dlatego żeby nie robić złączeń z innymi tabelamy wykorzystujemy atrybuty które już są w tabeli faktu.

Wyniki lasa losowego:

Wagi:

attribute	weight
pedestrian_id	0.080
severity_type	0.032
day_of_week	0.889



Wyniki jednego z 50 drzew:

```
severity_type = fatality
|   pedestrian_id = F
|   |   day_of_week > 5.500
|   |   |   day_of_week > 6.500: F {T=145, F=531}
|   |   |   day_of_week ≤ 6.500: F {T=162, F=568}
|   |   day_of_week ≤ 5.500
|   |   |   day_of_week > 1.500
|   |   |   |   day_of_week > 2.500
|   |   |   |   |   day_of_week > 4.500: F {T=203, F=487}
|   |   |   |   |   day_of_week ≤ 4.500
|   |   |   |   |   |   day_of_week > 3.500: F {T=168, F=420}
|   |   |   |   |   |   day_of_week ≤ 3.500: F {T=154, F=371}
|   |   |   |   day_of_week ≤ 2.500: F {T=161, F=360}
|   |   |   day_of_week ≤ 1.500: F {T=210, F=432}
|   pedestrian_id = T
|   |   day_of_week > 1.500
|   |   |   day_of_week > 2.500
|   |   |   |   day_of_week > 3.500
|   |   |   |   |   day_of_week > 4.500
|   |   |   |   |   |   day_of_week > 6.500: F {T=1, F=8}
|   |   |   |   |   |   day_of_week ≤ 6.500: F {T=0, F=28}
|   |   |   |   |   day_of_week ≤ 4.500: F {T=2, F=16}
|   |   |   |   day_of_week ≤ 3.500: F {T=0, F=6}
|   |   |   day_of_week ≤ 2.500: F {T=1, F=6}
|   |   day_of_week ≤ 1.500: F {T=2, F=7}
severity_type = minor_injury
|   pedestrian_id = F
|   |   day_of_week > 5.500
|   |   |   day_of_week > 6.500: F {T=7291, F=7682}
|   |   |   day_of_week ≤ 6.500: T {T=9220, F=9069}
|   |   day_of_week ≤ 5.500
|   |   |   day_of_week > 4.500: T {T=12359, F=10325}
|   |   |   day_of_week ≤ 4.500
|   |   |   |   day_of_week > 3.500: T {T=12066, F=9771}
|   |   |   |   day_of_week ≤ 3.500
|   |   |   |   |   day_of_week > 2.500: T {T=11650, F=9701}
|   |   |   |   |   day_of_week ≤ 2.500
|   |   |   |   |   |   day_of_week > 1.500: T {T=11585, F=9515}
|   |   |   |   |   |   day_of_week ≤ 1.500: T {T=10435, F=8673}
|   pedestrian_id = T
|   |   day_of_week > 1.500
|   |   |   day_of_week > 5.500
|   |   |   |   day_of_week > 6.500: F {T=25, F=36}
|   |   |   |   day_of_week ≤ 6.500: F {T=24, F=36}
|   |   |   day_of_week ≤ 5.500
|   |   |   |   day_of_week > 4.500: F {T=33, F=92}
|   |   |   |   day_of_week ≤ 4.500
|   |   |   |   |   day_of_week > 3.500: F {T=35, F=52}
|   |   |   |   |   day_of_week ≤ 3.500
|   |   |   |   |   |   day_of_week > 2.500: F {T=19, F=57}
|   |   |   |   |   |   day_of_week ≤ 2.500: F {T=30, F=59}
|   |   day_of_week ≤ 1.500: F {T=37, F=40}
severity_type = property_damage
|   pedestrian_id = F
```

```
|   |   day_of_week > 6.500: T {T=5383, F=5380}
|   |   day_of_week ≤ 6.500
|   |   |   day_of_week > 4.500
|   |   |   |   day_of_week > 5.500: T {T=7767, F=7073}
|   |   |   |   day_of_week ≤ 5.500: T {T=10858, F=9877}
|   |   |   day_of_week ≤ 4.500
|   |   |   |   day_of_week > 1.500
|   |   |   |   |   day_of_week > 3.500: T {T=10500, F=9059}
|   |   |   |   |   day_of_week ≤ 3.500
|   |   |   |   |   |   day_of_week > 2.500: T {T=10128, F=8653}
|   |   |   |   |   |   day_of_week ≤ 2.500: T {T=9932, F=8490}
|   |   |   |   day_of_week ≤ 1.500: T {T=8695, F=7785}
|   pedestrian_id = T
|   |   day_of_week > 4.500
|   |   |   day_of_week > 5.500
|   |   |   |   day_of_week > 6.500: F {T=27, F=61}
|   |   |   |   day_of_week ≤ 6.500: F {T=41, F=99}
|   |   |   day_of_week ≤ 5.500: F {T=61, F=135}
|   |   day_of_week ≤ 4.500
|   |   |   day_of_week > 3.500: F {T=52, F=103}
|   |   |   day_of_week ≤ 3.500
|   |   |   |   day_of_week > 2.500: F {T=60, F=105}
|   |   |   |   day_of_week ≤ 2.500
|   |   |   |   |   day_of_week > 1.500: F {T=62, F=112}
|   |   |   |   |   day_of_week ≤ 1.500: F {T=54, F=95}
severity_type = serious_injury
|   pedestrian_id = F
|   |   day_of_week > 5.500
|   |   |   day_of_week > 6.500: F {T=3854, F=6064}
|   |   |   day_of_week ≤ 6.500: F {T=4531, F=6434}
|   |   day_of_week ≤ 5.500
|   |   |   day_of_week > 1.500
|   |   |   |   day_of_week > 4.500: F {T=4999, F=5907}
|   |   |   |   day_of_week ≤ 4.500
|   |   |   |   |   day_of_week > 3.500: F {T=5028, F=5462}
|   |   |   |   |   day_of_week ≤ 3.500
|   |   |   |   |   |   day_of_week > 2.500: F {T=4602, F=5251}
|   |   |   |   |   |   day_of_week ≤ 2.500: F {T=4507, F=4913}
|   |   |   day_of_week ≤ 1.500: F {T=4184, F=5015}
|   pedestrian_id = T
|   |   day_of_week > 1.500
|   |   |   day_of_week > 6.500: F {T=12, F=51}
|   |   |   day_of_week ≤ 6.500
|   |   |   |   day_of_week > 2.500
|   |   |   |   |   day_of_week > 4.500
|   |   |   |   |   |   day_of_week > 5.500: F {T=27, F=68}
|   |   |   |   |   |   day_of_week ≤ 5.500: F {T=24, F=84}
|   |   |   |   |   day_of_week ≤ 4.500
|   |   |   |   |   |   day_of_week > 3.500: F {T=32, F=63}
|   |   |   |   |   |   day_of_week ≤ 3.500: F {T=21, F=64}
|   |   |   |   day_of_week ≤ 2.500: F {T=14, F=56}
|   |   day_of_week ≤ 1.500: F {T=43, F=60}
```

Jak tu już jest widocznie mamy tu dużo więcej tych gałęziej, większość z których nie dają nam jakiś interesujących wyników, dlatego będziemy rozważać tylko kilka różnych gałęzi.

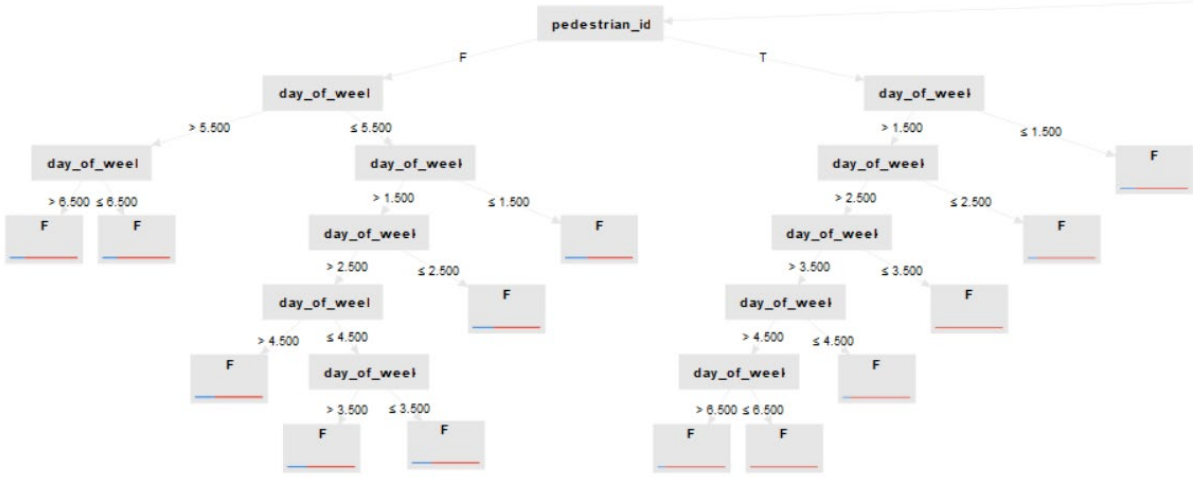
Schemat drzewa w tym przypadku:

Severity_type[property_damage|serioues_injury|minory_injury|fatality] - > pedestrian_id[T|F] -> day_of_week[1-7]

Rozglądamy gałąź Severity_type (fatality) :

```
severity_type = fatality
|   pedestrian_id = F
|   |   day_of_week > 5.500
|   |   |   day_of_week > 6.500: F {T=145, F=531}
|   |   |   day_of_week ≤ 6.500: F {T=162, F=568}
|   |   day_of_week ≤ 5.500
|   |   |   day_of_week > 1.500
|   |   |   |   day_of_week > 2.500
|   |   |   |   |   day_of_week > 4.500: F {T=203, F=487}
|   |   |   |   |   day_of_week ≤ 4.500
|   |   |   |   |   |   day_of_week > 3.500: F {T=168, F=420}
```

```
|   |   |   |   |   |   day_of_week ≤ 3.500: F {T=154, F=371}
|   |   |   |   day_of_week ≤ 2.500: F {T=161, F=360}
|   |   |   day_of_week ≤ 1.500: F {T=210, F=432}
|   pedestrian_id = T
|   |   day_of_week > 1.500
|   |   |   day_of_week > 2.500
|   |   |   |   day_of_week > 3.500
|   |   |   |   |   day_of_week > 4.500
|   |   |   |   |   |   day_of_week > 6.500: F {T=1, F=8}
|   |   |   |   |   |   day_of_week ≤ 6.500: F {T=0, F=28}
|   |   |   |   |   day_of_week ≤ 4.500: F {T=2, F=16}
|   |   |   |   day_of_week ≤ 3.500: F {T=0, F=6}
|   |   |   day_of_week ≤ 2.500: F {T=1, F=6}
|   |   day_of_week ≤ 1.500: F {T=2, F=7}
```



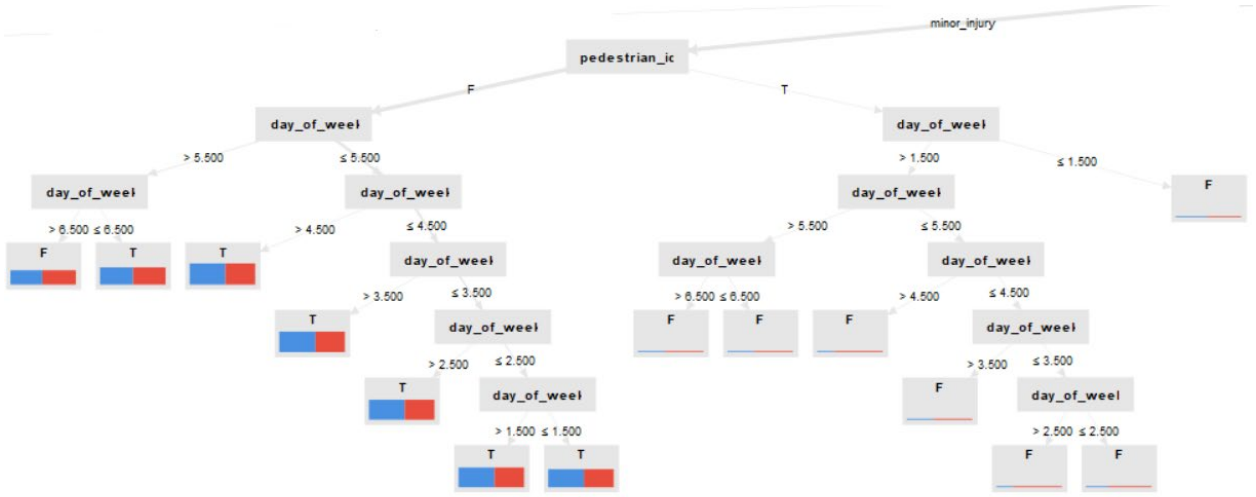
W porównywaniu z innymi surowościami, tutaj w większości wypadki zdarzają na odcinkach drogi bez skrzyrzowania. I co jest bardzo straszne, ale interesujące – jak jest widoczne ze zdjęcia wyżej, prawie wszystkie wypadki w których uczestniczyli piesze i surowość to „fatality”, zdarzali same bez skrzyrzowania, ale jeśli popatrzy na liczbu to zdarzają taki wypadki na ścięście rzadko.

Dlaczego aż tak wiele wypadków z surowością „fatality” zdarzają same bez skrzyrzowania? Według mnie tak jest, dlatego że bez srzyżowania np za miastem kierowcy jadą z większą prędkością i tam szansa fatalnego wypadku jest w kilka razy większa.

Dalej przechodzimy do gałęzi severity_type (minory_injury):

```
severity_type = minor_injury
|   pedestrian_id = F
|   |   day_of_week > 5.500
|   |   |   day_of_week > 6.500: F {T=7291, F=7682}
|   |   |   day_of_week ≤ 6.500: T {T=9220, F=9069}
|   |   day_of_week ≤ 5.500
|   |   |   day_of_week > 4.500: T {T=12359, F=10325}
|   |   |   day_of_week ≤ 4.500
|   |   |   |   day_of_week > 3.500: T {T=12066, F=9771}
|   |   |   |   day_of_week ≤ 3.500
|   |   |   |   |   day_of_week > 2.500: T {T=11650, F=9701}
|   |   |   |   |   day_of_week ≤ 2.500
|   |   |   |   |   |   day_of_week > 1.500: T {T=11585, F=9515}
|   |   |   |   |   |   day_of_week ≤ 1.500: T {T=10435, F=8673}
|   pedestrian_id = T
|   |   day_of_week > 1.500
|   |   |   day_of_week > 5.500
|   |   |   |   day_of_week > 6.500: F {T=25, F=36}
|   |   |   |   day_of_week ≤ 6.500: F {T=24, F=36}
|   |   |   day_of_week ≤ 5.500
```

```
|   |   |   |   day_of_week > 4.500: F {T=33, F=92}
|   |   |   |   day_of_week ≤ 4.500
|   |   |   |   |   day_of_week > 3.500: F {T=35, F=52}
|   |   |   |   |   day_of_week ≤ 3.500
|   |   |   |   |   |   day_of_week > 2.500: F {T=19, F=57}
|   |   |   |   |   |   day_of_week ≤ 2.500: F {T=30, F=59}
|   |   day_of_week ≤ 1.500: F {T=37, F=40}
```



Tutaj już mamy dużo więcej rekordów – dlatego kolumny dłuższe pionowo.

Prawie wszystkie resultaty, mają coś średnie między „True” a „False”, ale co dla mnie jest najbardziej interesujące – popatrzcie, po lewiej stronie jest jedyna litera „F”.

```
day_of_week > 6.500: F {T=7291, F=7682}
```

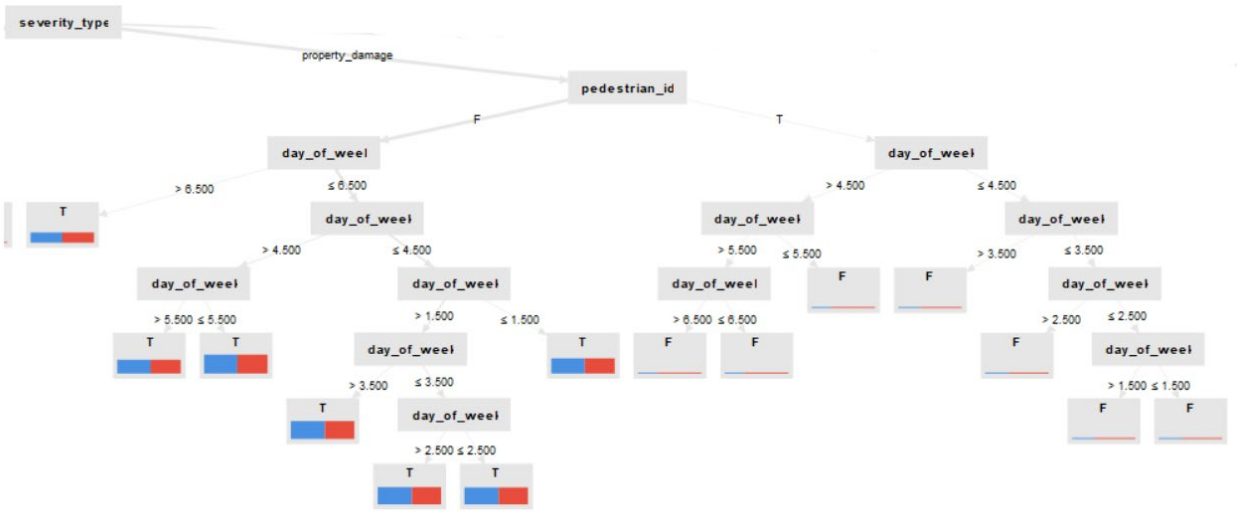
Mamy tam tyle rekordów, to jeszcze trochy i byłoby „T”.

Co jest ciekawe, że tutaj w przypadku z „pedestrian_id = F” mamy bezwzględnie dużo więcej rekordów, z „intersection_id = T”, ale kiedy już mamy „pedestrian_id = T” to wtedy mamy już więcej wypadków z „intersection_id = F”. Wnioski z tego już zrobię na końcu eksperymentu kiedy zobaczymy wszystkie gałęzi.

Następnie, weźmiemy gałąź severity_type(property_damage):

```
severity_type = property_damage
|   pedestrian_id = F
|   |   day_of_week > 6.500: T {T=5383, F=5380}
|   |   day_of_week ≤ 6.500
|   |   |   day_of_week > 4.500
|   |   |   |   day_of_week > 5.500: T {T=7767, F=7073}
|   |   |   |   day_of_week ≤ 5.500: T {T=10858, F=9877}
|   |   |   day_of_week ≤ 4.500
|   |   |   |   day_of_week > 1.500
|   |   |   |   |   day_of_week > 3.500: T {T=10500, F=9059}
|   |   |   |   |   day_of_week ≤ 3.500
|   |   |   |   |   |   day_of_week > 2.500: T {T=10128, F=8653}
|   |   |   |   |   |   day_of_week ≤ 2.500: T {T=9932, F=8490}
|   |   |   day_of_week ≤ 1.500: T {T=8695, F=7785}
|   pedestrian_id = T
|   |   day_of_week > 4.500
|   |   |   day_of_week > 5.500
|   |   |   |   day_of_week > 6.500: F {T=27, F=61}
|   |   |   |   day_of_week ≤ 6.500: F {T=41, F=99}
|   |   |   day_of_week ≤ 5.500: F {T=61, F=135}
|   |   day_of_week ≤ 4.500
|   |   |   day_of_week > 3.500: F {T=52, F=103}
|   |   |   day_of_week ≤ 3.500
|   |   |   |   day_of_week > 2.500: F {T=60, F=105}
|   |   |   |   day_of_week ≤ 2.500
|   |   |   |   |   day_of_week > 1.500: F {T=62, F=112}
```

```
|   |   |   |   day_of_week ≤ 1.500: F {T=54, F=95}
```

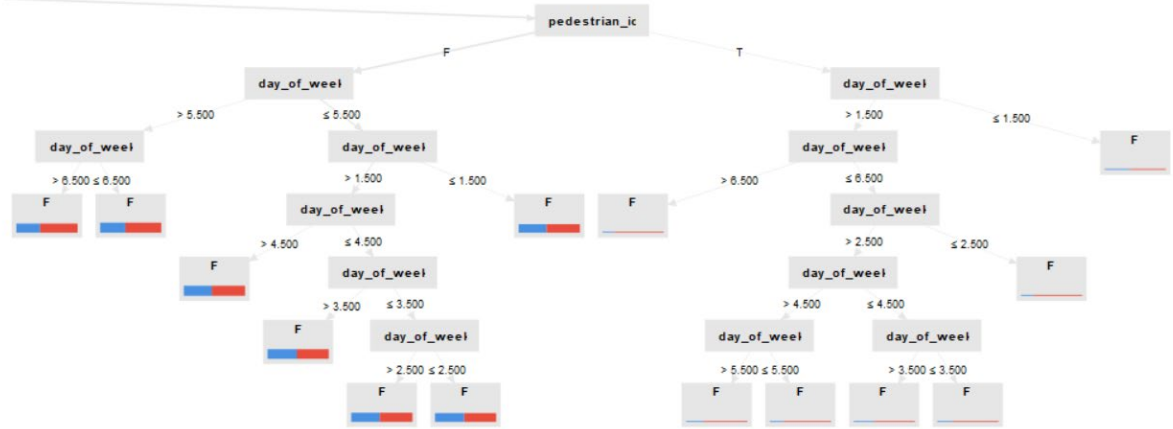


Tutaj widzimy podobne wyniki do gałęzi (severtity_type(property_damage)). Też wszędzie gdzie „pedestrian_id = F” mamy „intersection_id = T”, gdzie mamy „pedestrian_id = T” – to mamy wszędzie „intersection_id = F”. Też już opiszę moją opinię we wniskach o całym eksperymencie po ostatnim „severity_type”.

A ogólnie to aż bardzo podobnie do gałęzi z „severity_type(property_damage)”, dlatego tutaj, nie mamy co więcej analizować.

Rozważymy ostatnnią gałąź z „severity_type(serious_injury)”:

```
severity_type = serious_injury
|   pedestrian_id = F
|   |   day_of_week > 5.500
|   |   |   day_of_week > 6.500: F {T=3854, F=6064}
|   |   |   day_of_week ≤ 6.500: F {T=4531, F=6434}
|   |   day_of_week ≤ 5.500
|   |   |   day_of_week > 1.500
|   |   |   |   day_of_week > 4.500: F {T=4999, F=5907}
|   |   |   |   day_of_week ≤ 4.500
|   |   |   |   |   day_of_week > 3.500: F {T=5028, F=5462}
|   |   |   |   |   day_of_week ≤ 3.500
|   |   |   |   |   |   day_of_week > 2.500: F {T=4602, F=5251}
|   |   |   |   |   |   day_of_week ≤ 2.500: F {T=4507, F=4913}
|   |   |   day_of_week ≤ 1.500: F {T=4184, F=5015}
|   pedestrian_id = T
|   |   day_of_week > 1.500
|   |   |   day_of_week > 6.500: F {T=12, F=51}
|   |   |   day_of_week ≤ 6.500
|   |   |   |   day_of_week > 2.500
|   |   |   |   |   day_of_week > 4.500
|   |   |   |   |   |   day_of_week > 5.500: F {T=27, F=68}
|   |   |   |   |   |   day_of_week ≤ 5.500: F {T=24, F=84}
|   |   |   |   |   day_of_week ≤ 4.500
|   |   |   |   |   |   day_of_week > 3.500: F {T=32, F=63}
|   |   |   |   |   |   day_of_week ≤ 3.500: F {T=21, F=64}
|   |   |   |   day_of_week ≤ 2.500: F {T=14, F=56}
|   |   day_of_week ≤ 1.500: F {T=43, F=60}
```



Tutaj możemy zobaczyć że tak jak i w „severity_type = fatality” „intersection_id = F” i w przypadku „pedestrian_id = F oraz pedestrian_id = T”. Tak tutaj już jest nie taka duża różnica, jak w przypadku z „severity_type = fatality”.

Wnioski z eksperymentu 2:

Moim zdaniem, najważniejszym wnioskiem z tego całego eksperymentu jest to, co mowiłem że powiem na końcu. Jak moglismy zobaczyć ze wszystkich tych wykresów, mamy tylko dwa wykresy gdzie „intersection_id = F” w przypadkach kiedy „pedestrian_id = F lub = T”, i to jest badrzo ważne, z tego możemy zrobić wniosek, że w wypadkach z „severity_type = fatality” są najmniej ważliwy kryterium „pedestrian_id”, ponieważ jak już mówiłem wypadki z „sevity_type = fatality” zdarzają częściej kiedy kierowcy jadą z większą prędkością i wtedy już nie tak ważliwe czy uczesticzył w wypadku pieszy czy nie. To samo ale nie tak mocno mamy w wypadkach z „severity_type = serious_injury”, też nie zależnie od „pedestrian_id” mamy „intersection_id = F”

Wnioski ogólne:

Przeprowadziliśmy dwa eksperymenty eksploracji danych. W obu przypadkach korzystaliśmy z programu „Rapid Miner” oraz robiliśmy to wszystko wykorzystując metody lasów losowych. Udało nam przeprowadzić tej analizy i dowiedzieć dwa duże ważliwych wyniki, które już byli opisane wcześniej we wnioskach eksperementów.