Przewidywanie dolegliwości wypadków samochodowych na podstawie warunków atmosferycznych, rodzaju drogi i zagęszczenia ruchu

Marcin Dolatowski

272680

Spis treści

1 Opisanie problemu

Jako temat projektu na zajęcia laboratoryjne z przedmiotu Metody Systemowe i Decyzyjne postanowiłem opracować klasyfikację dolegliwości wypadków drogowych w zależności od warunków atmosferycznych, rodzaju drogi oraz zagęszczenia ruchu. Celem tego projektu jest zastosowanie metod systemowych i decyzyjnych w analizie danych dotyczących wypadków drogowych w celu identyfikacji czynników wpływających na ich powstawanie oraz określenia stopnia ryzyka związanych z różnymi warunkami drogowymi.

W opracowywanym systemie klasyfikacji będę uwzględniał różnorodne parametry, takie jak typy nawierzchni, intensywność ruchu, która będzie wyrażona poprzez populację stanu, w którym odbył się wypadek, pogoda oraz warunki atmosferyczne. Poprzez analizę tych czynników będę dążył do stworzenia modelu, który pozwoli na przypisanie dolegliwości wypadków do odpowiednich kategorii (od 1 do 4), co ułatwi identyfikację potencjalnych zagrożeń oraz podejmowanie skutecznych działań zapobiegawczych.

Wprowadzenie klasyfikacji opartej na systemowych metodach i danych decyzyjnych może przyczynić się do poprawy bezpieczeństwa drogowego poprzez identyfikację obszarów o podwyższonym ryzyku wypadków oraz skoncentrowanie się na wprowadzeniu odpowiednich środków zaradczych. Ostatecznie, ten projekt ma na celu przyczynienie się do redukcji liczby wypadków drogowych poprzez zastosowanie systemowych podejść w analizie i zarządzaniu ryzykiem.

2 Pozyskanie danych

2.1 Pobranie danych z kaggla

Moją startową bazą danych był zbiór danych ze strony kaggle, który zawierał spis wypadków samochodowych ze Stanów Zjednoczonych wraz z informacjami na ich temat. `https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents` Nie wszystkie dane były potrzebne, więc zredukowałem listę kolumn tylko do tych, które były mi potrzebne.

2.2 Próba pozyskania danych o pogodzie z API

W celu urozmaicenia zdobywania danych planowałem pobrać dane o warunkach atmosferycznych z API, które udostępniało historyczne dane o pogodzie. `https://www.visualcrossing.com/resources/documentation/weather-api/timeline-weather-api/` Niestety limit dziennych zapytań w darmowym planie wynosił 50 zapytań na dzień, co niweczyło moje plany. Postanowiłem więc używać danych o pogodzie z początkowego zbioru danych z kaggla.

2.3 Pozyskiwanie danych na temat rodzaju drogi

W celu pozyskania danych na temat rodzaju drogi użyłem biblioteki osmnx. `https://osmnx.readthedocs.io/en/stable/user-reference.html` Na pod-

stawie szerokości i długości geograficznej w promieniu 2 kilometrów szukałem dróg i jeśli nie udało się znaleźć zwracałem typ drogi jako pusty, ale jeśli udało się coś znaleźć zwracałem rodzaj tej drogi, która była najbliżej startowego punktu.

2.4 Pozyskiwanie danych na temat populacji stanu

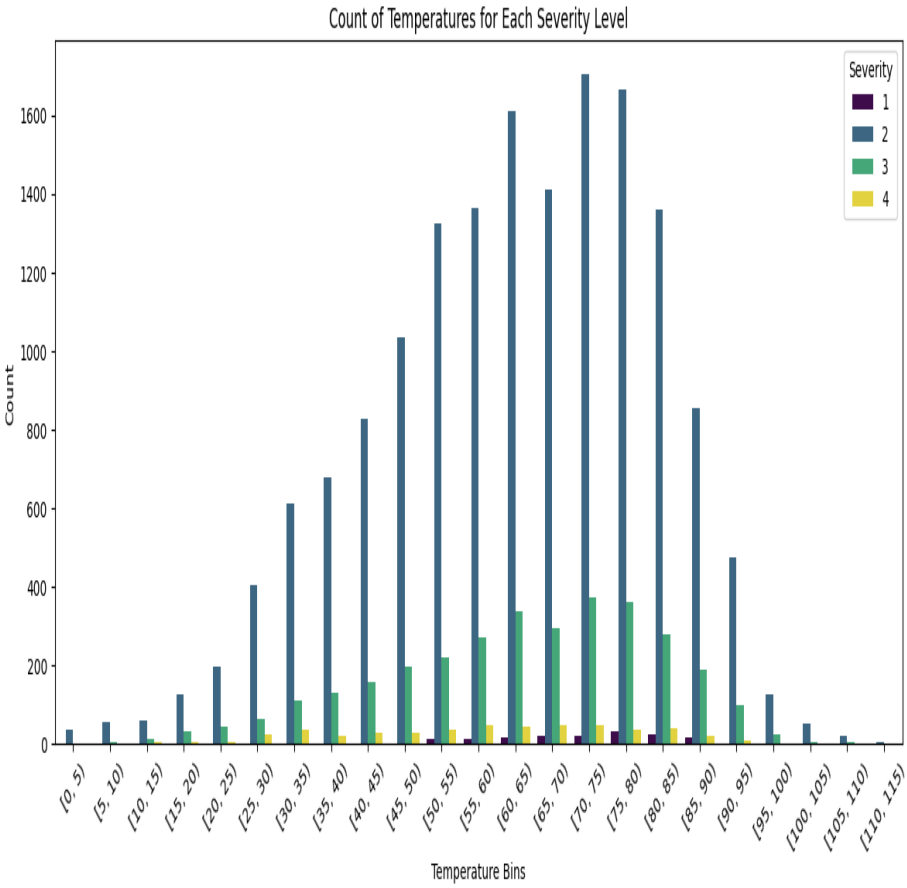
W celu pozyskania danych na temat populacji danego stanu, użyłem pliku txt ze strony `https://www.101computing.net/us-population/`, dzięki któremu na podstawie kolumny state mogłem otrzymać populację danego stanu, która będzie służyła jako wyznacznik zagęszczenia ruchu.

2.5 Końcowa obróbka danych w celu zapisania do pliku csv

Po dodaniu kolumn z rodzajem drogi i populacją, postanowiłem usunąć wszystkie rekordy, które mają jakąś kolumnę z wartością Nan. Ponieważ uzupełnianie danych na temat rodzaju drogi robiłem w kilku iteracjach (ze względu na bardzo długi czas potrzebny na uzupełnienie danych) musiałem łączyć ze sobą pliki csv, do których zapisywałem częściowo uzupełnione rekordy. Z tego powodu zaczęły się tworzyć dodatkowe indeksy, które także usunąłem. Ostatecznie w końcowym pliku, na którym pracowałem zostało ponad 20000 rekordów z kolumnami: Severity, Temperature(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Speed(mph), Road_Type, Population.

3 Analiza danych

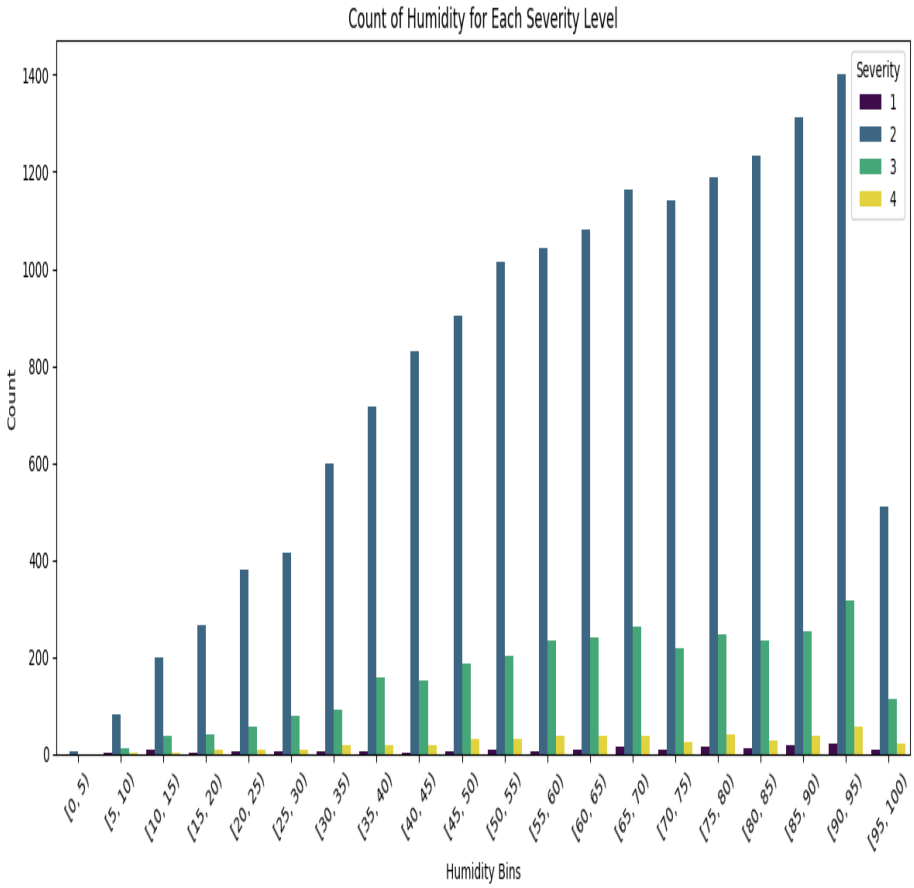
3.1 Zależność severity od temperature



Rysunek 1: Wykres pokazujący zależność severity od temperature

W celu pokazania relacji między temperaturą, a dolegliwością postanowiłem stworzyć wykres, który ukazuje liczbę wystąpień danej dolegliwości dla temperatur. Niestety pokazanie liczności dolegliwości dla każdej temperatury negatywnie wpływało na czytelność wykres, dlatego zdecydowałem się podzielić temperatury na przedziały, tak jak można to zaobserwować na wykresie. Na wykresie można zauważyć, że najwięcej zdarzeń występuje w przedziałach temperatur od 45 do 80 stopni Fahrenheita. Nasilenie poziomu 2 dolegliwości dominuje we wszystkich przedziałach temperatur, widać, że im wyższa temperatura, tym mniejsza liczba zdarzeń o wyższym poziomie nasilenia (3 i 4).

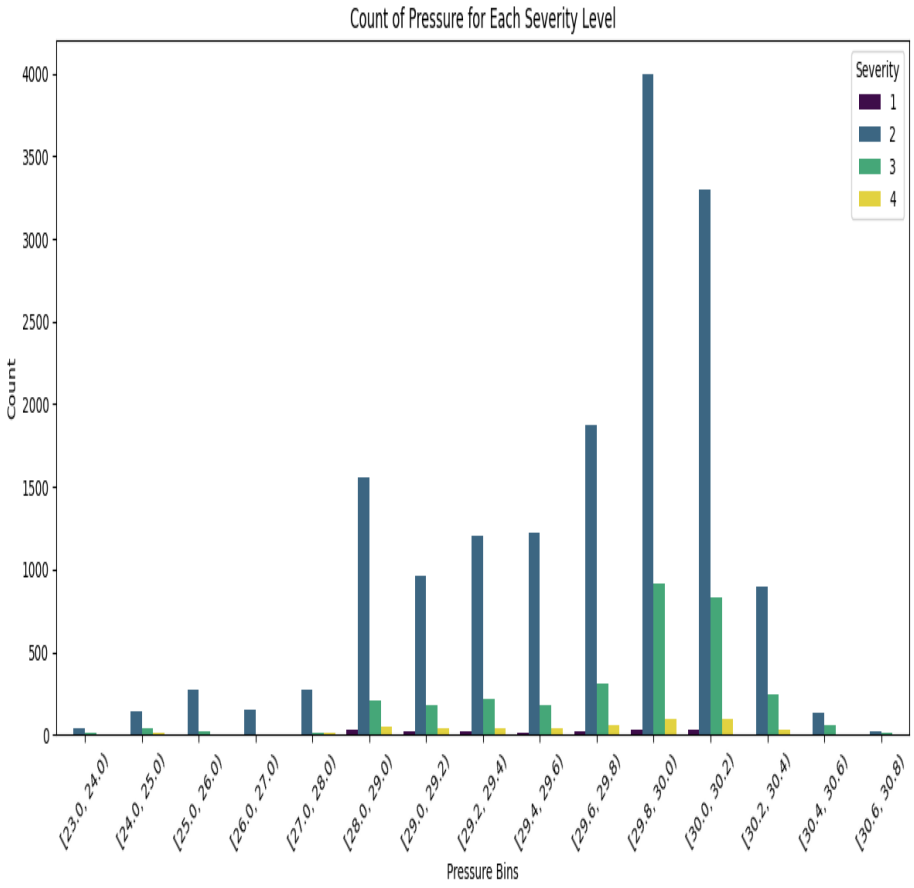
3.2 Zależność severity od humidity



Rysunek 2: Wykres pokazujący zależność severity od humidity

Podobnie jak w przypadku temperatury nie było możliwe przedstawienie liczności severity z osobna dla każdej wilgotności, dlatego tak jak w poprzednim przykładzie zdecydowałem się na pogrupowanie wartości wilgotności. Z wykresu można odczytać, że po raz kolejny dominowały zdarzenia z 2 poziomem dolegliwości, można zauważyć, że doszło do największej ilości zdarzeń w przedziale 90-95 stopni wilgotności.

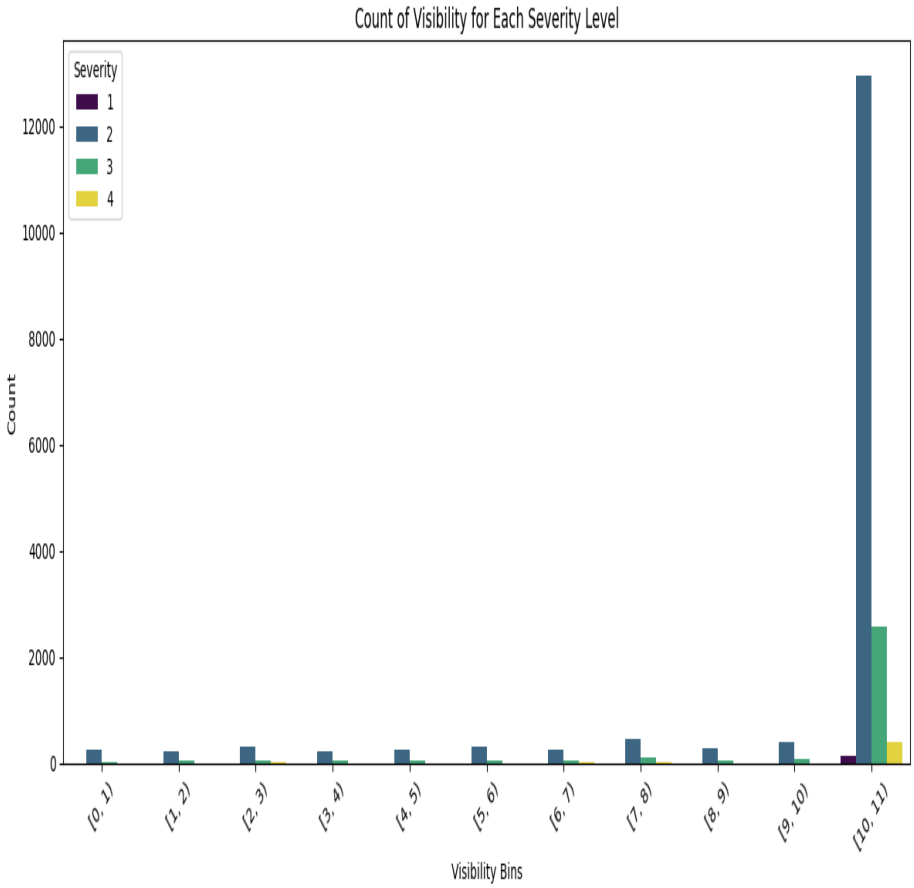
3.3 Zależność severity od pressure



Rysunek 3: Wykres pokazujący zależność severity od pressure

Tak samo jak w poprzednich przypadkach pogrupowałem wartości ciśnienia przedziałami, aby zwiększyć czytelność wykresu. Ponieważ najwięcej wartości występowało na przedziale 29-31, zdecydowałem się na zastosowanie dwóch typów podziałów, od 23 do 29 zastosowałem przedziały co 1 jednostkę, a od 29 do 31 przedziały co 0.2 jednostki. Dominują zdarzenia z 2 poziomem dolegliwości, a drugie co do częstotliwości wystąpień są zdarzenia z 3 stopniem nasilenia.

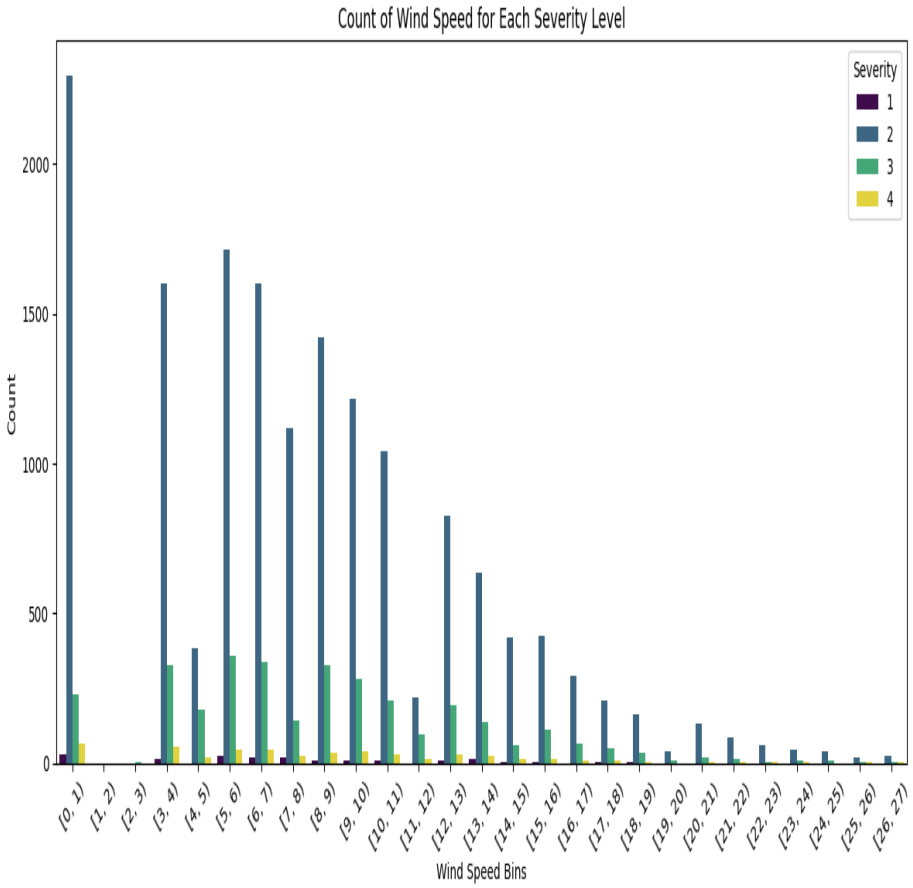
3.4 Zależność severity od visibility



Rysunek 4: Wykres pokazujący zależność severity od visibility

W tym przypadku także zdecydowałem się na pogrupowanie wartości widoczności, aby zwiększyć przejrzystość wykresu. Zdecydowana większość zdarzeń występuje przy widoczności 10-11 mil. Podobnie jak w poprzednich wykresach poziom nasilenia 2 jest najbardziej powszechny.

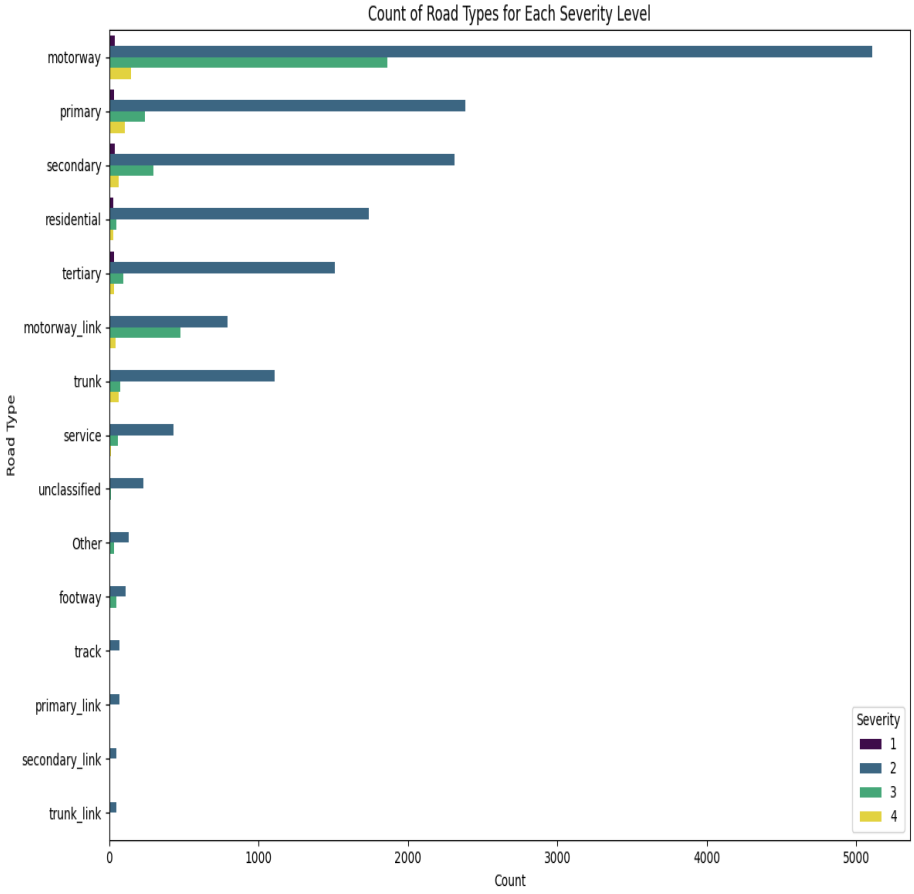
3.5 Zależność severity od wind speed



Rysunek 5: Wykres pokazujący zależność severity od wind speed

W tym przypadku także zdecydowałem się na pogrupowanie wartości widoczności, aby zwiększyć przejrzystość wykresu. Można zauważyć, że najwięcej zdarzeń występuje przy niskich prędkościach wiatru (0-1mph), a następnie liczba zdarzeń maleje wraz ze wzrostem prędkości wiatru. Poziom nasilenia 2 jest również dominujący we wszystkich przedziałach prędkości wiatru. Przy wyższych prędkościach wiatru zdarzenia o wyższym poziomie nasilenia (3 i 4) są mniej liczne.

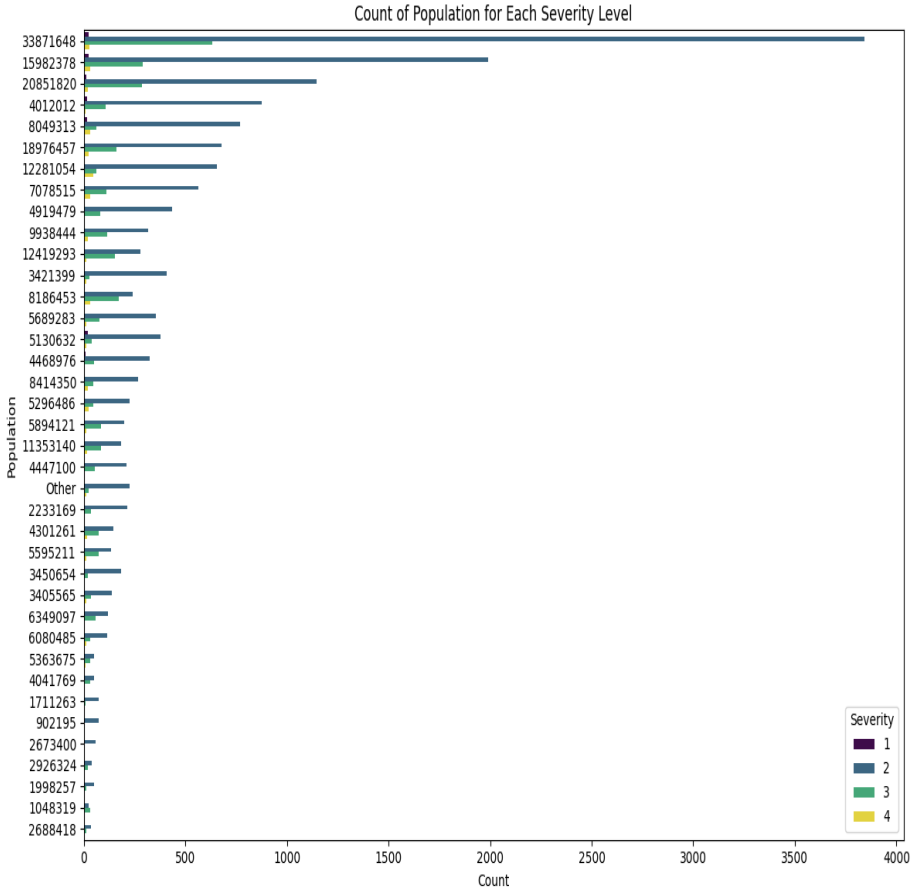
3.6 Zależność severity od road type



Rysunek 6: Wykres pokazujący zależność severity od road type

W przypadku tego wykresu, aby zwiększyć przejrzystość, postanowiłem typy dróg, których liczba wystąpień nie przekracza 50 zaliczyć do zbioru other. Z wykresu można odczytać, że największa część wypadków zdarzyła się na drodze typu motorway, czyli na autostradzie, co wydaję się być logiczne. Po raz kolejny największa cześć wypadków miała stopień dolegliwości 2, ale na autostradzie znaczna część była także 3 stopnia dolegliwości.

3.7 Zależność severity od population



Rysunek 7: Wykres pokazujący zależność severity od population

Podobnie jak w powyższym przypadku postanowiłem populacje, których liczba wystąpień nie przekracza 50 zaliczyć do zbioru other, dzięki czemu przejrzystość wykresu znacznie się poprawiła. Z wykresu możemy odczytać, że największa część wypadków odbyła się w stanach, których populacja jest największa, co za tym idzie zagęszczenie ruchu też (tak zostało przyjęte na początku raportu).

4 Eksperymenty i wyniki klasyfikacji

4.1 Przygotowanie danych

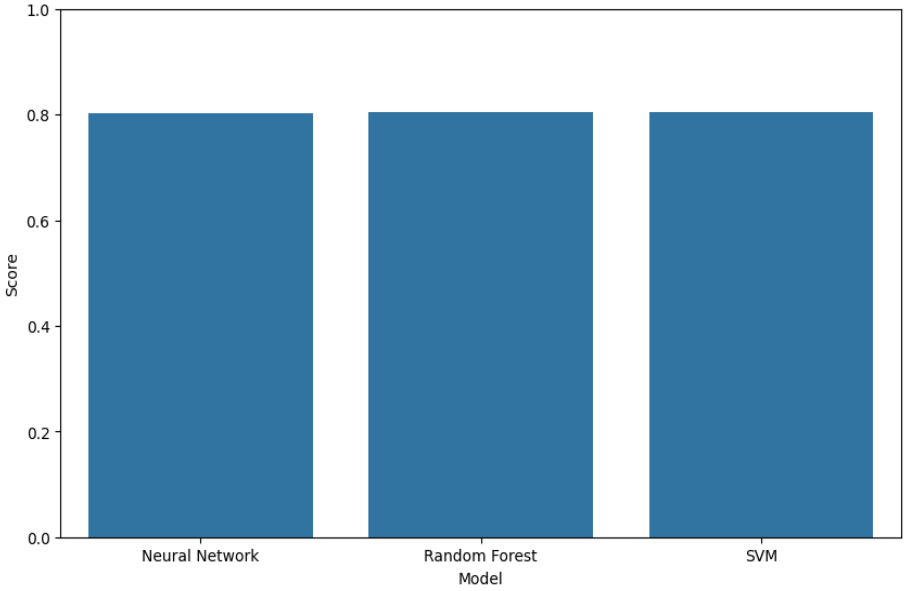
Aby poprawnie przeprowadzić testy modeli klasyfikacji potrzebne było odpowiednie przygotowanie danych. Podzieliłem dane na zbiór danych treningowych i testowych w proporcji odpowiednio 80 proc. do 20 proc. Kolumnę Road_Type z racji tego, że była stringiem zakodowałem, używając funkcji oferowanej przez bibliotekę pandas.

4.2 Przygotowanie modeli

Przed przystąpieniem do tworzenia modeli stworzyłem funkcję, która ma na celu znalezienie najlepszych parametrów dla każdego z modeli. Następne przystąpiłem do tworzenia modeli klasyfikacyjnych - maszyny wektorów nośnych, używając klasy SVC z biblioteki scikit-learn, lasu drzew losowych, także używając biblioteki scikit-learn oraz sieci neuronowych przy użyciu klasy Sequential z biblioteki Keras. W celu zbadania dokładności modeli SVM i Random Forest użyłem funkcji cross_val_score. Do zbadania modelu neural networks niestety nie mogłem użyć tej funkcji, ponieważ nie mogłem użyć KerasClassifiera(napotkałem problemy, których nie mogłem rozwiązać), który umożliwia przekazanie tej funkcji modelu sieci neuronowych. Tak więc użyłem metody fit na tym modelu, która zwraca historię wyników każdej iteracji i z niej obliczyłem średnią.

4.3 Wyniki testów

Jak można zauważyć najlepszy wynik osiągnęła maszyna wektorów nośnych, drugi wynik uzyskał las drzew, a najgorzej wypadły sieci neuronowe. Wszystkie modele osiągnęły wynik dokładności powyżej 80 proc. co uważam za dobry wynik, oznacza to, że mając do dyspozycji dane pogodowe, rodzaj drogi, a także zagęszczenie ruchu drogowego jesteśmy w stanie przewidzieć dolegliwość wypadku z 80 proc. pewnością.



Rysunek 8: Wyniki modeli w postaci wykresu

Model	Score
Neural Network	0.8020637631416321
Random Forest	0.8035647279549719
SVM	0.8054409005628518

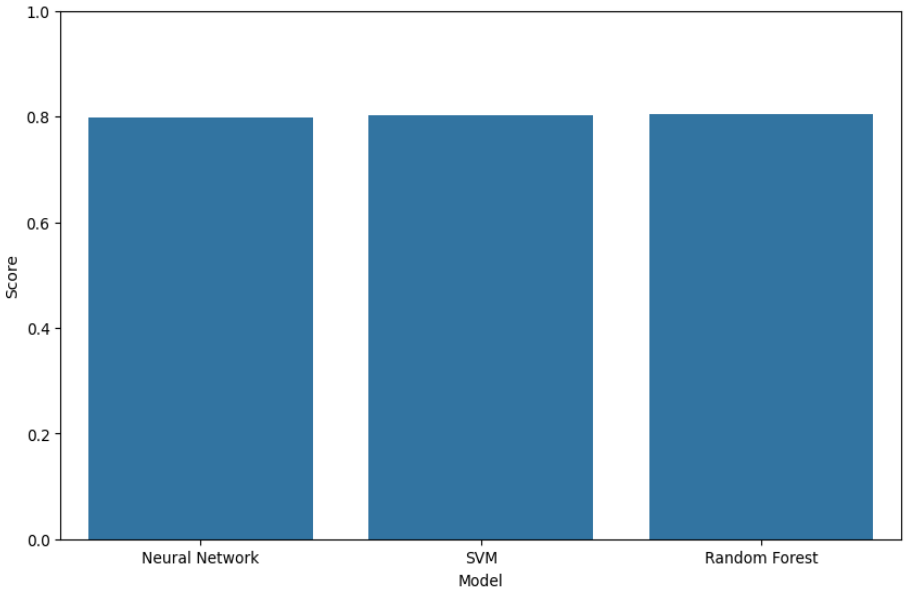
Rysunek 9: Wyniki modeli w postaci tabeli

4.4 Badanie wpływu poszczególnych kolumn na dokładność

W celu weryfikacji, które kolumny danych wpływają na dokładność modelu postanowiłem sprawdzać dokładności modeli usuwając poszczególne kolumny z danych treningowych.

4.5 Usunięcie kolumny population

Jak można zaobserwować dokładność modeli SVM i lasu losowych nie spadła znacząco, model sieci neuronowych spadł minimalne poniżej 80 proc., z tych wyników można wywnioskować, że kolumna population wpływa pozytywnie na dokładność modeli, lecz nie w stopniu bardzo znaczącym.



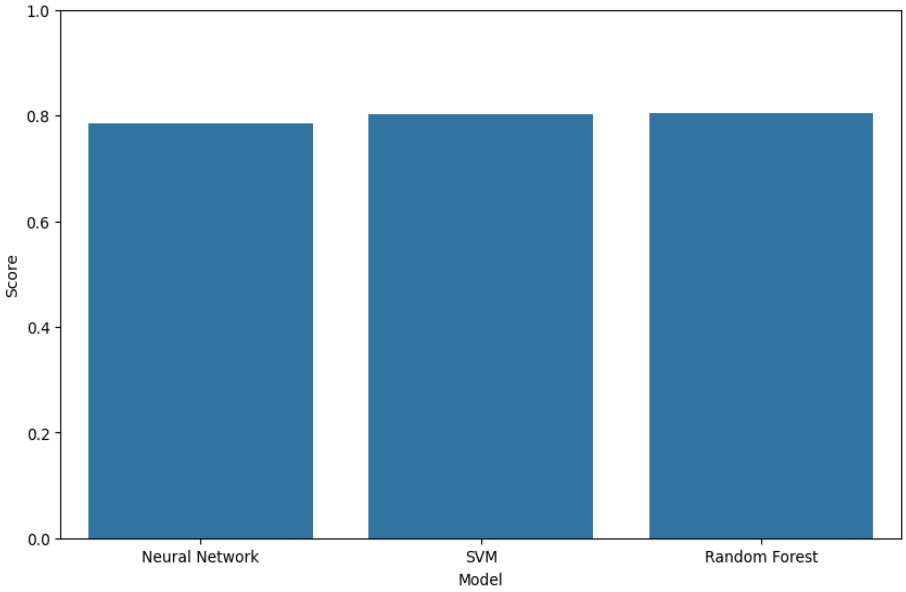
Rysunek 10: Wykres dokładności modeli bez kolumny population

Model	Score
Neural Network	0.7987570321559906
SVM	0.8030644152595373
Random Forest	0.8038148843026892

Rysunek 11: Tabela dokładności modeli bez kolumny population

4.6 Usunięcie kolumny humidity

Jak można zaobserwować dokładność maszyny wektorów nośnych i lasu drzew losowych nie spadła znacząco, model sieci neuronowych spadł poniżej 80 proc. i to o 1 punkt procentowy bardziej niż w przypadku usunięcia kolumny population. Oznacza to, że kolumna humidity odgrywa istotną rolę tylko w przypadku sieci neuronowych.



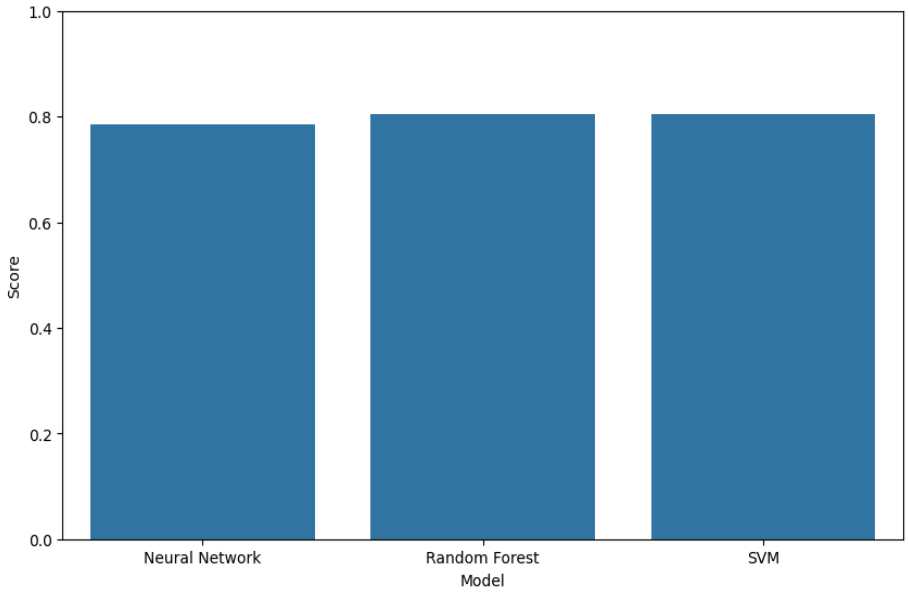
Rysunek 12: Wykres dokładności modeli bez kolumny humidity

Model	Score
Neural Network	0.784624739466235
SVM	0.8031894934333959
Random Forest	0.8038774233896184

Rysunek 13: Tabela dokładności modeli bez kolumny humidity

4.7 Usunięcie kolumny pressure

W przypadku usunięcia kolumny pressure można zauważyć, że jest to bardzo podobna sytuacja, co w przypadku humidity. Tylko model sieci neuronowych odnotował znaczący spadek dokładności.



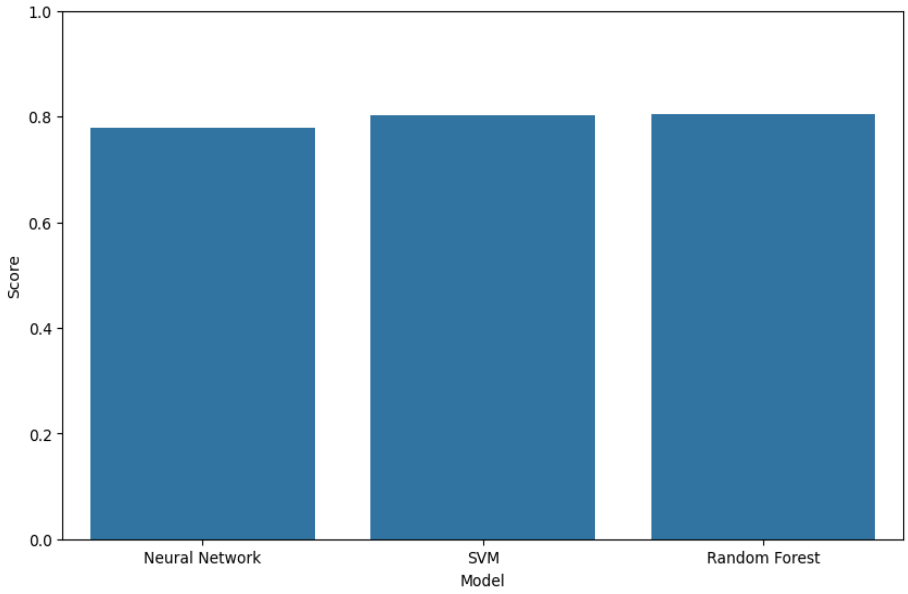
Rysunek 14: Wykres dokładności modeli bez kolumny pressure

Model	Score
Neural Network	0.7840978482179344
Random Forest	0.8035647279549719
SVM	0.8037523452157599

Rysunek 15: Tabela dokładności modeli bez kolumny pressure

4.8 Usunięcie kolumny road type

Jak można zaobserwować dokładność modeli SVM i lasu drzew losowych nie spadła znacząco, model sieci neuronowych spadł niestety poniżej 80 proc. W tym przypadku sieci neuronowe odnotowały jak dotąd największy spadek dokładności.



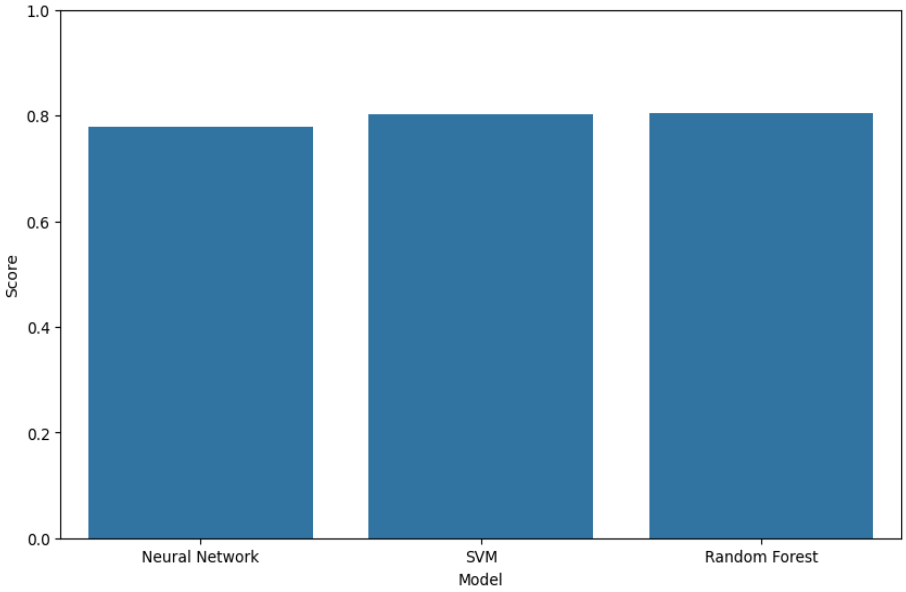
Rysunek 16: Wykres dokładności modeli bez kolumny road type

Model	Score
Neural Network	0.7782864036969841
SVM	0.8020012507817386
Random Forest	0.8036272670419011

Rysunek 17: Tabela dokładności modeli bez kolumny road type

4.9 Usunięcie kolumny wind speed

Tak samo jak w przypadku kolumny road type znaczący spadek dokładności odnotował tylko model sieci neuronowych.



Rysunek 18: Wykres dokładności modeli bez kolumny wind speed

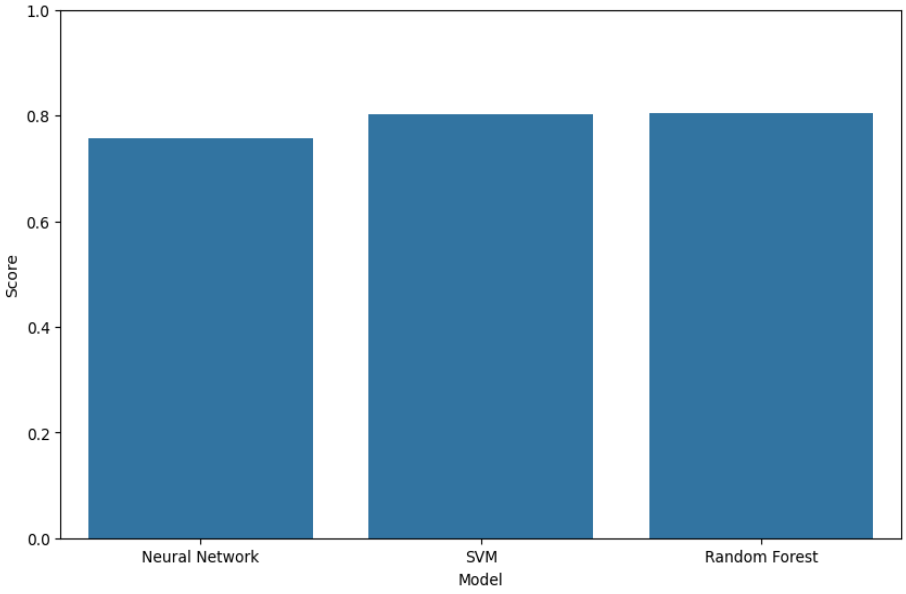
Model	Score
Neural Network	0.7778611376788467
SVM	0.8018761726078798
Random Forest	0.8036272670419013

Rysunek 19: Tabela dokładności modeli bez kolumny wind speed

4.10 Usunięcie kolumny visibility

Podczas przekazywania modelom danych bez kolumny visibility, tak samo jak w poprzednich przypadkach spadek odnotował tylko model sieci neuronowych. Jest

to spadek równy prawie 5 punktom procentowym. Z pewnością brak kolumny visibility wpływa znacząco na dokładność sieci neuronowych.



Rysunek 20: Wykres dokładności modeli bez kolumny visibility

Model	Score
Neural Network	0.7561882189009339
SVM	0.8035021888680426
Random Forest	0.8035647279549719

Rysunek 21: Tabela dokładności modeli bez kolumny visibility

4.11 Wnioski

Jak można zauważyć żadne usunięcie kolumny z danych nie wpłynęło bardzo negatywnie na wyniki dokładności maszyny wektorów nośnych lub lasu drzew losowych, w przypadku sieci neuronowych natomiast procent dokładności spadł

poniżej 80 proc., w niektórych sytuacjach nawet o 5 punktów procentowych. Możemy z tego wyciągnąć wnioski, że modele SVM i lasu drzew losowych są w stanie poprawnie klasyfikować dolegliwość wypadków nawet z brakującymi kolumnami, sieci neuronowe natomiast potrzebują całego zestawu danych, aby robić to z zadowalającą dokładnością.