|  |  |
| --- | --- |
|  | **POLITECHNIKA ŁÓDZKA** |
|  | Wydział Elektrotechniki, Elektroniki, Informatyki  i Automatyki |
|  | Instytut Mechatroniki i Systemów Informatycznych |

Praca dyplomowa  
 magisterska

na temat:

**System prognozowania warunków meteorologicznych z wykorzystaniem algorytmów eksploracji danych**

**(System for forecasting meteorological conditions using data mining algorithms)**

|  |  |
| --- | --- |
| Imię i Nazwisko: | **Kinga Sochacka** |
| Nr albumu: | **234005** |
| Specjalność: | **Inteligentne systemy baz danych** |
| Kierunek: | **Informatyka** |

Opiekun pracy:

prof. dr hab. inż. **Adam Pelikant**

Łódź, wrzesień 2021

**SPIS TREŚCI**

[1. Wstęp 2](#_Toc71717611)

[2. Cel i zakres pracy. 2](#_Toc71717612)

[3. Podstawy teoretyczne 2](#_Toc71717613)

[3.1. Eksploracja danych 2](#_Toc71717614)

[3.2. Szeregi czasowe 2](#_Toc71717615)

[3.3. Sztuczne sieci neuronowe 2](#_Toc71717616)

[3.4. Wykorzystane technologie 2](#_Toc71717617)

[3.2.1 Python 2](#_Toc71717618)

[3.4.1. Keras 2](#_Toc71717619)

[3.4.2. PyCharm 2](#_Toc71717620)

[3.4.3. Git 2](#_Toc71717621)

[3.4.4. Java 2](#_Toc71717622)

[3.4.5. InteliJ 2](#_Toc71717623)

[3.4.6. HTML 2](#_Toc71717624)

[3.4.7. Thymeleaf 2](#_Toc71717625)

[4. Praktyczna realizacja systemu 2](#_Toc71717626)

[4.1. Wykorzystany zbiór danych 2](#_Toc71717627)

[4.2. Przetwarzanie danych 2](#_Toc71717628)

[4.3. Wykonany model predykcyjny 2](#_Toc71717629)

[4.4. Implementacja aplikacji webowej 2](#_Toc71717630)

[Literatura 2](#_Toc71717631)

# Wstęp

Informatyka to obecnie jedna z najszybciej rozwijających się dziedzin nauki. Dostęp do ogromnych ilości danych wymusza poszukiwanie sposobów na poprawne i efektywne ich przechowywanie oraz przetwarzanie, a właśnie tym zajmuje się informatyka. Powstają coraz to nowsze i lepsze technologie, które sprostają tym zadaniom. Ilość generowanych w dzisiejszym świecie informacji doprowadziła do wyodrębnienia takich działów informatyki jak hurtownie danych czy Big Data.

Przechowywanie, przetwarzanie i analiza dużych, zmiennych, różnorodnych zbiorów danych nie jest prostym zadaniem, jednak bardzo wartościowym. Może prowadzić do zdobycia nowych, cennych informacji. Z pomocą w tym zadaniu przychodzi eksploracja danych, która wykorzystuje szybkość komputera do odszukiwania prawidłowości w gromadzonych danych.

Eksploracja danych znajduje zastosowanie w wielu dziedzinach np. biznesie czy przemyśle. W niniejszej pracy zostanie ona wykorzystana w meteorologii. Nie jest to łatwa nauka, gdyż warunki pogodowe są bardzo zmienne i zależne od wielu czynników, a prognozowanie pogody jest procesem złożonym i trudnym.

Szybki rozwój technologii umożliwił gromadzenie ogromnych ilości danych. Jednak wydobycie z nich przydanych informacji jest dużo trudniejszym wyzwaniem. Tradycyjne metody analizy danych przestały być wystarczające. To doprowadziło do powstania nowych technologii, które za pomocą specjalistycznych algorytmów przetwarzają duże ilości danych.

# Cel i zakres pracy.

Celem pracy jest wykonanie analizy historycznych danych meteorologicznych w celu prognozowania warunków pogodowych z wykorzystaniem technik eksploracji danych.

Proces realizacji pracy podzielono na kilka etapów:

* wybór zbioru danych do analizy,
* przygotowanie historycznych danych meteorologicznych,
* wykonanie modelu predykcyjnego,
* implementacja aplikacji wykorzystującej wykonany model dla aktualnych danych.

# Podstawy teoretyczne

## Eksploracja danych

Eksploracja danych (ang. data mining, inaczej zgłębianie, ekstrakcja danych) to proces pozyskiwania nowej wiedzy z posiadanych danych. Wykorzystuje szybkość komputera oraz wyspecjalizowane algorytmy w celu odkrywania wzorców i prawidłowości w zbiorach danych. Daje także możliwość prognozowania wyników dla przyszłych obserwacji.

Zgłębianie danych jest integralną częścią odkrywania wiedzy z baz danych (ang. Knowledge discovery in databases – KDD), czyli procesu przekształcania surowych danych w przydatne informacje. Na rysunku 3.1 przedstawiono etapy KDD, które prowadzą do otrzymania wartościowych informacji z danych wejściowych. Pierwszy etap - wstępne przetwarzanie danych obejmuje zebranie danych z różnych źródeł, usunięcie duplikatów i niepełnych danych, wybór cech, redukcję wymiarowości, normalizację, wyodrębienie podzbiorów. Często jest to proces żmudny i bardzo czasochłonny. Celem wstępnego przetworzenia danych jest przekształcenie surowych danych w dane odpowiednio sformatowane pod kątem analizy. Drugi etap to eksploracja danych, w ramach której do konretnych danych dobiera się i stosuje odpowiednie algorytmy. Na ostatni krok – końcowe przetwarzanie składają się filtrowanie wzorców, wizualizacja oraz interpretacja wyników. Stosowane są także statystyczne miary oceniające wydajność wykonanych modeli. Dzięki temu etapowi cały proces znajduje zastosowanie w praktyce – może zostać wykorzystany w aplikacjach biznesowych, ponieważ zapewnia, że tylko poprawne i przydatne informacje zostaną zastosowane w systemie.



Rys. 0.1 Proces odkrywania wiedzy z baz danych (KDD) [1].

Eksploracja danych jest wykorzystywana do dwóch typów zadań. Pierwszy typ do zadania predykcyjne, których celem jest przewidywanie wartości określonego atrybutu na podstawie wartości innych atrybutów. Drugi typ to zadania opisywania danych, czyli szukanie wzorców i prawidłowości w danych, określanie korelacji między danymi, trendów i zauważanie anomalii.

Zgłębianie danych musi sprostać wielu wymaganiom.

## Szeregi czasowe

Tradycyjne zbiory danych używane w uczeniu maszynowym to kolekcja informacji, w której każda próbka jest traktowana tak samo. Inaczej jest w przypadku szeregów czasowych, które są ciągiem obserwacji uporządkowanych w czasie ze stałym krokiem. Wprowadza to dodatkową informację, która może być istotna w przypadku wielu problemów prognostycznych. Z drugiej strony przetwarzanie takich danych jest trudniejsze i wymaga zastosowania specjalnych metod.

Na potrzeby rozwiązywania problemów szeregów czasowych wprowadzono standardowe terminy używane do opisywania tego typu danych:

* t: aktualny czas, który stanowi punkt odniesienia,
* t – n: czas opóźniony, przeszłość w stosunku do czasu aktualnego,
* t + n: czas przyszły w stosunku do czasu aktualnego.

Aby móc zastosować szeregi czasowe w uczeniu maszynowym, należy odpowiednio przygotować posiadany zbiór danych. Na rysunku przedstawiono przykładowy zbiór danych, w którym dane są posortowane względem czasu.

|  |  |
| --- | --- |
| czas | zmienna |
| 10-05-2021 | 100 |
| 11-05-2021 | 50 |
| 12-05-2021 | 120 |
| 13-05-2021 | 90 |

Restrukturyzacja tego zbioru polega na tym, że wartość poprzedniego kroku jest używana do prognozowania wartości zmiennej w kroku następnym. Na rysunku widać, że poprzedni krok czasowy jest wejściem (X), a następujący po nim krok czasowy jest wyjściem (y). Kolejność obserwacji jest zachowana i musi taka pozostać w trakcie trenowania modelu predykcyjnego. Dla pierwszego wiersza brakuje wartości wejściowej, a dla ostateniego wartości wyjściowej, dlatego należy je usunąć.

|  |  |
| --- | --- |
| X | y |
| NaN | 100 |
| 100 | 50 |
| 50 | 120 |
| 120 | 90 |
| 90 | NaN |

Wykorzystanie wcześniejszych kroków czasowych do przewidywania następnego nazywane jest metodą przesuwnego okna (ang. sliding window method).

Ze względu na ilość obserwacji wyróżniamy dwa rodzaje szeregów czasowych:

* jednowymiarowe (ang. Univariate Time Series) – obserwowana jest tylko jedna zmienna w każdym kroku czasowym,
* wielowymiarowe (ang. Multivariate Time Series) – obserwowane są dwie lub więcej zmiennych w każdym kroku czasowym.

Prognozowanie w przypadku wielowymiarowych zbiorów jest dużo trudniejsze i bardziej złożone niż jednowymiarowych. Na rysunku przedstawiono przykład wielowymiarowego szeregu czasowego.

|  |  |  |  |
| --- | --- | --- | --- |
| X1 | X2 | X3 | y |
| NaN | NaN | 100 | 200 |
| 100 | 200 | 50 | 150 |
| 50 | 150 | 120 | 60 |
| 120 | 60 | 90 | 30 |
| 90 | 30 | NaN | NaN |

Ze względu na liczbę przewidywanych kroków czasowych wyróżniamy dwa rodzaje prognozowania:

* jednoetapowy (ang. One-step Forecast) – przewidywany jest jeden następny krok (przykład z rysunków ),
* wieloetapowe (ang. Multi-step Forecast) – prognozowane są dwa lub więcej kroki czasowe (rys. ).

|  |  |  |  |
| --- | --- | --- | --- |
| X1 | X2 | y1 | y |
| NaN | NaN | 100 | 200 |
| 100 | 200 | 50 | 150 |
| 50 | 150 | 120 | 60 |
| 120 | 60 | 90 | 30 |
| 90 | 30 | NaN | NaN |

## Sztuczne sieci neuronowe

## Wykorzystane technologie

### Python

Python jest interpretowanym językiem programowania wysokiego poziomu. Nie wymaga kompilacji przed uruchomieniem. Pisanie kodu w tym języku jest szybkie, ale uruchomienie zazwyczaj wolniejsze niż w przypadku języków kompilowanych. Jest dostosowany do programowania zorientowanego obiektowo, umożliwia definiowanie klas wraz z dziedziczeniem. Wyróżnia się dynamicznym deklarowaniem typów zmiennych. Twórcy Pythona położyli nacisk na czytelność kodu, co zmniejsza koszty jego utrzymania.

Python znajduje szerokie zastosowanie w takich dziedzinach informatyki jak aplikacje webowe, nauka o danych, sztuczna inteligencja.

### Keras

### PyCharm

### Git

### Java

### InteliJ

### HTML

### Thymeleaf

# Praktyczna realizacja systemu

## Wykorzystany zbiór danych

## Przetwarzanie danych

## Wykonany model predykcyjny

## Implementacja aplikacji webowej

# Podsumowanie

# Literatura

[1] Tan, Steinbach, Kumar, Introduction to Data Mining, Pearson, 2014

[2] Jason Brownlee, Introduction to Time Series Forecasting with Python, Machine Learning Mastery, 2020

[3] Jason Brownlee, Long Short–Term Memory Networks With Python, Machine Learning Mastery, 2017

[4] <https://www.edureka.co/blog/interview-questions/python-interview-questions/> 21 stycznia 2021

[5] <https://www.python.org/doc/essays/blurb/> ostatni dostęp 6 maja 2021

[6] <https://www.badania-statystyczne.pl/data-mining> ostatni dostęp 12 maja 2021