## Министерство образования Республики Беларусь

Учреждение образования

## БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Инженерно-экономический факультет

### Кафедра ЭИ

Отчет

по лабораторной работе № 6

«Деревья решений»

по курсу

«Машинное обучение»

Вариант 7

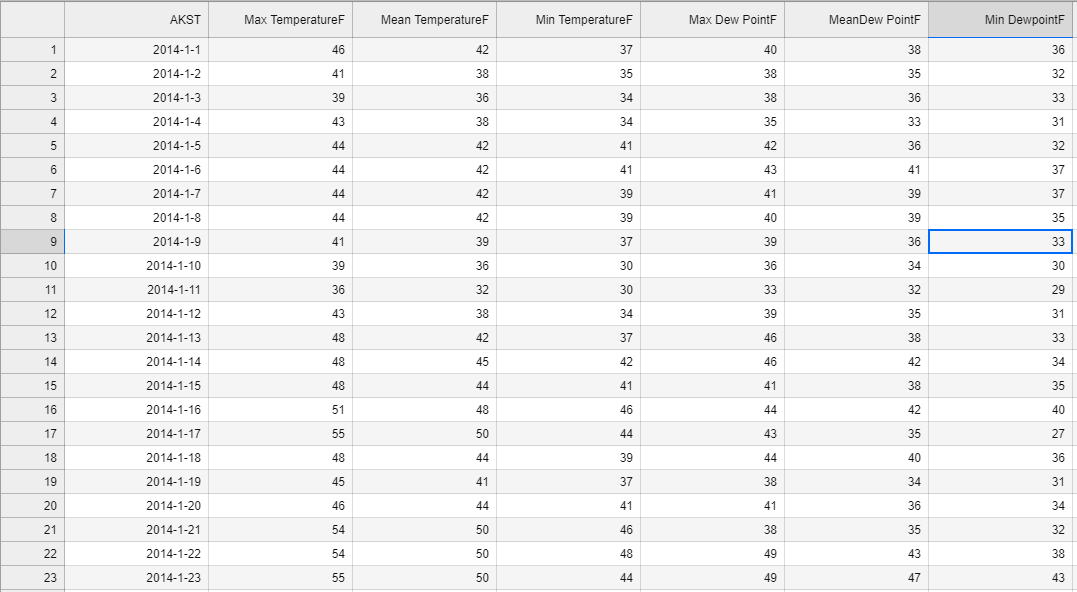
|  |  |
| --- | --- |
| Выполнил:  Ермаков К. Ю. | Проверил:  Милентьев В.А. |

Минск 2024

**Цель работы:** получение навыков прогнозирования с помощью метода дерева решений и оценки результатов в Python.

Для проведения анализа был выбран датасет изменения температуры в 2014 году.

Вид датасета:



*import pandas as pd*

*from sklearn.model\_selection import train\_test\_split*

*from sklearn.preprocessing import LabelEncoder*

*from sklearn.tree import DecisionTreeClassifier*

*from sklearn.metrics import confusion\_matrix, accuracy\_score*

*import seaborn as sns*

*import matplotlib.pyplot as plt*

*# Загрузка данных из файла*

*df = pd.read\_csv("sitka\_weather\_2014.csv")*

*# Удаление столбца 'Date', так как он не несет информации для обучения модели*

*df.drop('AKST', axis=1, inplace=True)*

*df.columns = df.columns.str.lower().str.replace(' ', '')*

*# Замена пустых значений на 0*

*df.fillna(0, inplace=True)*

*# Преобразование целевой переменной в числовой формат*

*df['events'] = df['events'].astype(str)*

*# Преобразование целевой переменной в числовой формат*

*label\_encoder = LabelEncoder()*

*df['events'] = label\_encoder.fit\_transform(df['events'])*

*# Разделение данных на признаки (X) и целевую переменную (y)*

*X = df.drop('events', axis=1)*

*y = df['events']*

*# Разделение данных на обучающий и тестовый наборы*

*X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)*

*# Создание и обучение модели дерева решений*

*model = DecisionTreeClassifier(random\_state=42)*

*model.fit(X\_train, y\_train)*

*# Оценка точности модели на тестовом наборе*

*y\_pred = model.predict(X\_test)*

*accuracy = accuracy\_score(y\_test, y\_pred)*

*print("Accuracy:", accuracy)*

*# Построение матрицы сходства*

*conf\_matrix = confusion\_matrix(y\_test, y\_pred)*

*print("Confusion Matrix:")*

*print(conf\_matrix)*

*# Важность признаков*

*feature\_importances = pd.DataFrame(model.feature\_importances\_, index=X.columns, columns=['Importance'])*

*print("Feature Importances:")*

*print(feature\_importances)*

*corr\_matrix = df.corr()*

*plt.figure(figsize=(12, 8))*

*sns.heatmap(corr\_matrix, annot=True, cmap='coolwarm', fmt=".2f", annot\_kws={"size": 10})*

*plt.title("Correlation Matrix")*

*plt.show()*

*# Построение матрицы сходства*

*plt.figure(figsize=(8, 6))*

*sns.heatmap(conf\_matrix, annot=True, cmap='Blues', fmt="d", annot\_kws={"size": 10})*

*plt.title("Confusion Matrix")*

*plt.xlabel("Predicted")*

*plt.ylabel("Actual")*

*plt.show()*

Из анализа были получены данные:

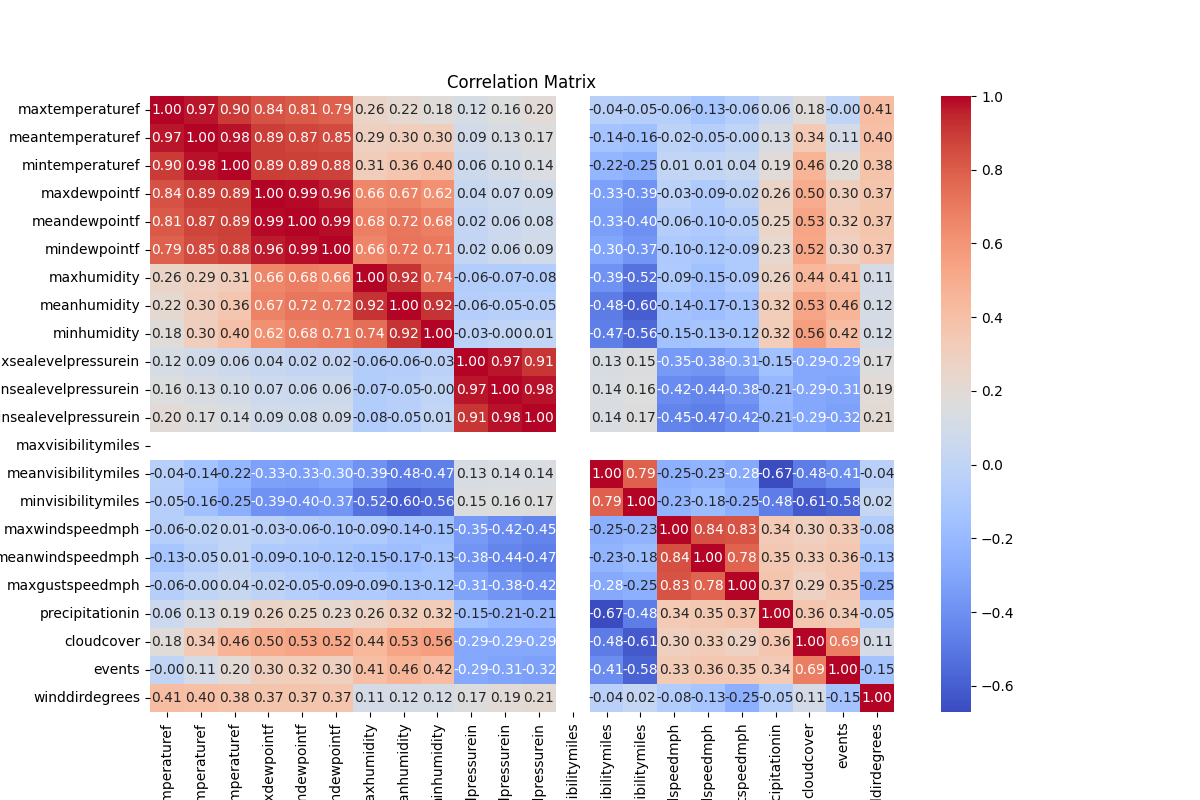


Рисунок 1 – Матрица корреляции

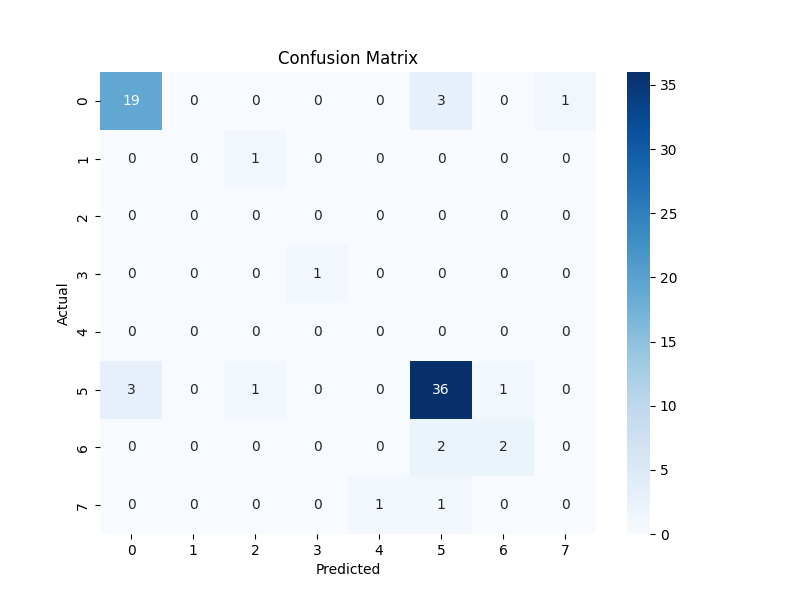
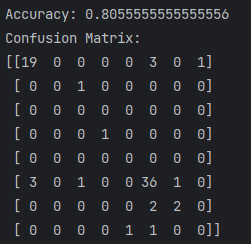
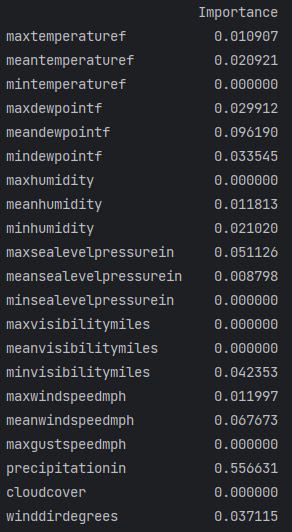


Рисунок 2- Матрица сходства





В ходе выполнения лабораторной работы был проведен анализ данных о погоде за январь 2014 года с использованием модели дерева решений.

Точность модели: Модель дерева решений показала достаточно высокую точность прогнозирования, составив примерно 80.56%. Это означает, что модель в большинстве случаев правильно классифицирует тип погодных событий на основе предоставленных признаков.

Матрица сходства (Confusion Matrix): Анализ матрицы сходства позволяет оценить, какие классы модель чаще всего путает между собой. Например, класс "0" (вероятно, связанный с дождем) имеет несколько случаев, когда он был неправильно классифицирован как класс "5" (вероятно, связанный с осадками).

Важность признаков: Оценка важности признаков позволяет выявить, какие факторы наиболее сильно влияют на прогнозирование погодных событий. Наибольшее влияние оказывают такие признаки, как количество осадков, средняя точка росы и средняя скорость ветра.

Исходя из проведенного анализа, можно сделать вывод о том, что модель дерева решений демонстрирует хорошую производительность при прогнозировании погодных событий на основе предоставленных данных.