УДК 519.724

АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ПРИКЛАДНЫХ ЗАДАЧ

Сабитов Б.Р., Кубанычбекова А.К., Антошина А.В., Дюшеналиева А.Т.
Кыргызский технический университет им. И. Раззакова

В статье рассматривается прогнозирования моделей для прикладных на основе машинного обучения. Для сравнительного анализа созданы модели машинного обучения и нейронной сети для прогнозирования различных моделей. В качестве базы данных использовано открытая .csv файл. Модель будет основываться на методах машинного обучения. Определена точность и ошибки модели, который допускает при прогнозировании. Определение матрицы правильное и неправильное прогнозирование меток.

Ключевые слова: машинное обучение, случайный лес, нейронные сети, прогнозирование.

КОЛДОНМО КӨЙГӨЙЛӨРДҮ БОЛЖОЛДОО ҮЧҮН МАШИНА ҮЙРӨНҮҮ АЛГОРИТМДЕРИ

Сабитов Б.Р., Кубанычбекова А.К., Антошина А.В., Дюшеналиева А.Т.
И.Раззаков атындагы Кыргыз техникалык университети

Бул макалада машинаны үйрөнүүгө негизделген прикладдык көйгөйлөрдү болжолдоо моделдери каралат. Салыштырмалуу талдоо үчүн ар кандай моделдерди болжолдоо үчүн машина үйрөнүү жана нейрондук тармак моделдери түзүлгөн. Маалымат базасы катары ачык .csv файлы колдонулган. Модель машинаны үйрөнүү ыкмаларына негизделет. Болжолдоодо жол берилген моделдин тактыгы жана каталары аныкталат. Туура жана туура эмес энбелгилердин алдын ала матрицасы аныкталат.

Баштапкы сөздөр: машина үйрөнүү, кокус токой, нейрон тармактары, болжолдоо.

MACHINE LEARNING ALGORITHMS FOR FORECASTING APPLIED PROBLEMS

Sabitov B.R., Kubanychbekova A.K., Antoshina A.V., Dyushenalieva A.T.
I. Razzakov Kyrgyz Technical University

This article examines forecasting models for applied problems based on machine learning. Machine learning and neural network models for forecasting various models were created for comparative analysis. An open .csv file was used as a database. The model will be based on machine learning methods. The accuracy and errors of the model during forecasting are determined. The matrix of correct and incorrect label predictions is determined.

Keywords: machine learning, random forest, neural networks, forecasting.

Введение.

Для реализации прикладных задач на основе машинного обучения рассмотрим конкретный пример. Основная задача будет построение модели машинного обучения о качестве воздуха AQI. Для этой цели подготовим данные и программное обеспечение для анализа данных. Процесс загрузки пакетов программных систем Python для программной реализации использует следующий набор программ.

Листинг 1. Пакет используемых программ.

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error, r2_score, confusion_matrix,classification_report, roc_curve, auc
from sklearn.ensemble import RandomForestRegressor, RandomForestClassifier
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout
from tensorflow.keras.optimizers import Adam
import matplotlib.pyplot as plt
import seaborn as sns
```

Прогнозирование для моделирования было реализовано Collaboratory и многие пакеты Python программ будем считать уже установленными в Collaboratory.Удобство использования данных пакетов не требуют дополнительной установки. Все пакеты как интеллектуальные

системы представляют собой как система из библиотеки Python и все они готовы к использованию. Следующей важной частью работы является загрузка данных, ее мы загрузим с диска локального компьютера.
Структура базы данных рассматриваемой задачи имеет вид
Листинг 2. Структура базы данных для определения качества воздуха. Целевая прогнозируемая функция CO.

```
data.head()
```

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)
0	10/03/2004	18.00.00	2.6	1360.0	150.0	11.9	1046.0	166.0
1	10/03/2004	19.00.00	2.0	1292.0	112.0	9.4	955.0	103.0
2	10/03/2004	20.00.00	2.2	1402.0	88.0	9.0	939.0	131.0
3	10/03/2004	21.00.00	2.2	1376.0	80.0	9.2	948.0	172.0
4	10/03/2004	22.00.00	1.6	1272.0	51.0	6.5	836.0	131.0

PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH	Unnamed: 15	Unnamed: 16
1056.0	113.0	1692.0	1268.0	13.6	48.9	0.7578	NaN	NaN
1174.0	92.0	1559.0	972.0	13.3	47.7	0.7255	NaN	NaN
1140.0	114.0	1555.0	1074.0	11.9	54.0	0.7502	NaN	NaN
1092.0	122.0	1584.0	1203.0	11.0	60.0	0.7867	NaN	NaN
1205.0	116.0	1490.0	1110.0	11.2	59.6	0.7888	NaN	NaN

Преобразование в задачу классификации (хорошее/плохое качество воздуха). Будем считать, что CO(GT)- угарный газ, или окись углерода (CO) – это газ без запаха, вкуса и цвета, который образуется в результате неполного сгорания. Угарный газ является ядовитым для людей и животных. Угарный газ поступает в организм по дыхательным путям, блокирует приток кислорода к крови, что нарушает деятельность организма- наш целевой показатель в данной базе данных. С этой целью в работе необходимо прогнозировать ее содержание в атмосфере.

Самое важное для построения модели машинного обучения мы должны сделать предварительный анализ данных, которые будут отражать особенности базы данных и записей, в которых мы найдем взаимосвязи между данными. Ниже представлен код, который указывает на предварительную обработку данных и удаление ненужных столбцов и строк с пропущенными значениями. Также мы пишем код для Преобразование в задачу классификации (хорошее/плохое качество

воздуха). Будем считать, что оксид углерода CO(GT) - наш целевой показатель.

Она в задачах по определению качества воздуха играет ключевую роль. В воздухе ее показатель колеблется в зависимости от примесей образующееся в окружающей среде. Выполним следующие операции на разделение базы на признаки и целевую переменную, а также мы произведем разделение на обучающую и тестовую выборки. Обязательным требованием при моделировании является процесс масштабирования элементов базы данных к интервалу 0,1.

Методы и методологии оптимизация нейронных сетей для прогнозирования качества воздуха

Основными инструментами при прогнозировании являются инструменты машинного обучения, которые предназначены для предварительной обработки данных и анализа некоторых алгоритмов. Рассмотрены оптимизаторы глубокого обучения, которые корректируют параметры модели, построенной с помощью глубокого обучения. Для определения наилучшей модели решается задача минимизации функцию потерь. Функция потерь измеряет, насколько хорошо модель может делать прогнозы для данного набора данных, а цель обучения модели — найти набор параметров модели, который дает минимально возможные потери.

Результатами исследований являются следующие прикладные задачи.

1. **Моделирование качества воздуха с применением машинного обучения.**

Построено многомерная регрессионная модель прогноза качества воздуха. Она состоит из нескольких частей. Создание самой модели многомерной регрессии и нейронной сети для результатов которых сравнивается в определенной метрике. Приведена компиляция и обучение модели, прогнозирование на тестовых данных и построение самой модели в виде графиков точности модели и ошибки модели.

Модель проверяется на валидационных данных, т.е. на данных тестирования. Построена нейронной сеть, в котором участвует процесс применения регуляризации модели. Данная технология использовано для переобученной модели. В данном случае мы применяем технологию Dropout, которая может отключать работу некоторых нейронов. Приведена также оценка модели. Было получено оценка регрессионной модели, и графики обучения модели и сравнительный анализ точной и прогнозированное значения модели. Результаты процесса проверки и определения MSE и R^2:

Листинг 3. Определения метрик.

```
/usr/local/lib/python3.11/dist-packages/keras/src/layers/core/dense.py:87: UserWarning:
  super().__init__(activity_regularizer=activity_regularizer, **kwargs)
88/88 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step
Нейронная сеть (регрессия):
MSE: 2529.2577
R2: 0.5657
```

Построенные регрессионные модели графически выглядят следующим образом.

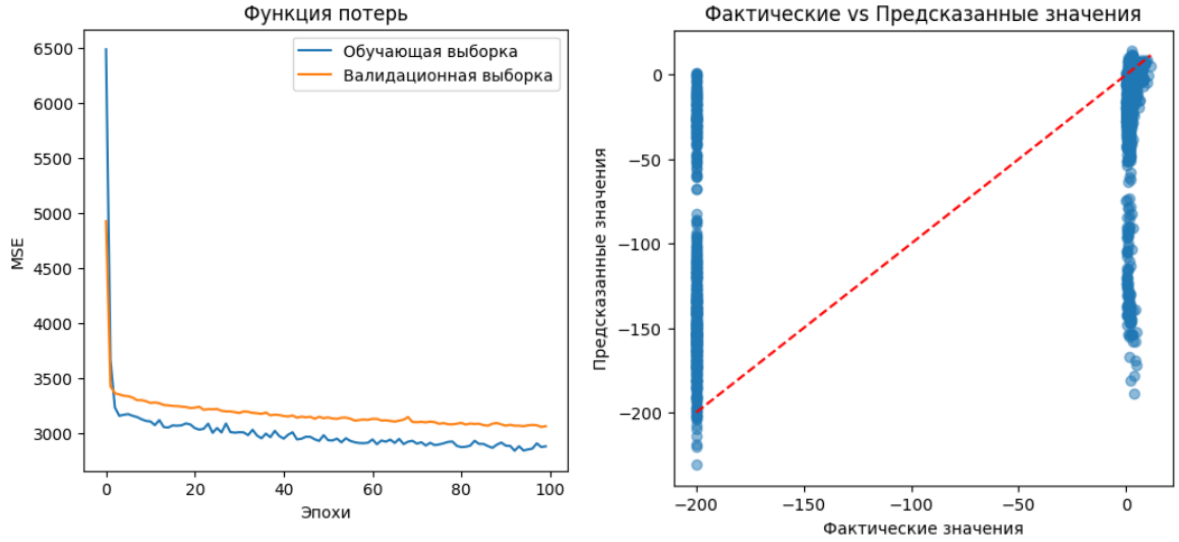


Рис.1. График модели функции потерь и регрессионный модель.

При построении моделей на базе машинного обучения необходимо определить важность признаков исследуемой базы данных. Ниже на листинге приведен результат алгоритма случайный лес для оценки среднеквадратичного отклонения и R-квадрат, называется, как коэффициент детерминации, которая определяет долю дисперсии, используемую в регрессионном анализе для оценки того, насколько хорошо модель соответствует наблюдаемым данным. Значение R-квадрата варьируется от 0 до 1, где 1 указывает на идеальное соответствие модели данным, а 0 означает, что модель не объясняет ни одной из вариаций. Вот результаты.

Листинг 4.Случайный лес-регрессия.

```
Случайный лес (регрессия):
MSE: 2164.9218
R2: 0.6283
```

Рис.2. График важности признаков участвующих в базе данных модели.

Теперь мы приступаем к построению нейронной сети для задачи классификации задачи ее применения к задачам качества воздуха.

2. **Прогнозирование качество воздуха машинного обучения и нейронных технологий.**

Задача была реализовано с соответствующим программным обеспечением для создания модели нейронной сети для классификации качества воздуха. Архитектура нейронной сети модели была обучена на большом количестве эпох. Получена вероятностей качества воздуха и оценка модели. Ниже представлен отчет прогнозировании задачи классификации на основе нейронной сети. Приведен показатель качества модели матрица ошибок, кривые ROC и AUC, которые отражают точность построенной модели. Вот результаты прогнозирования нейронной сетью для задачи классификации, 1-хорошое качества воздуха 0-плохое. Листинг 5.

```
/usr/local/lib/python3.11/dist-packages/keras/src/layers/core/dense.py:87:
  super().__init__(activity_regularizer=activity_regularizer, **kwargs)
88/88 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step

Нейронная сеть (классификация):
              precision    recall  f1-score   support

           0       0.92      0.91      0.92      1443
           1       0.91      0.92      0.91      1365

    accuracy                           0.91      2808
   macro avg       0.91      0.91      0.91      2808
weighted avg       0.91      0.91      0.91      2808
```

Здесь мы написали код построения графиков.

Матрица ошибок выглядит таким образом, которая предсказывает правильно и неправильно предсказанные данные т.е. разницу между истинным и прогнозным классом. Следующий график ROC кривая модели,

максимальная значение которой равно 1 , показывает на качество модели. Как видно из рисунка ниже точность достигает 97%, что является хорошим показателем качества модели.

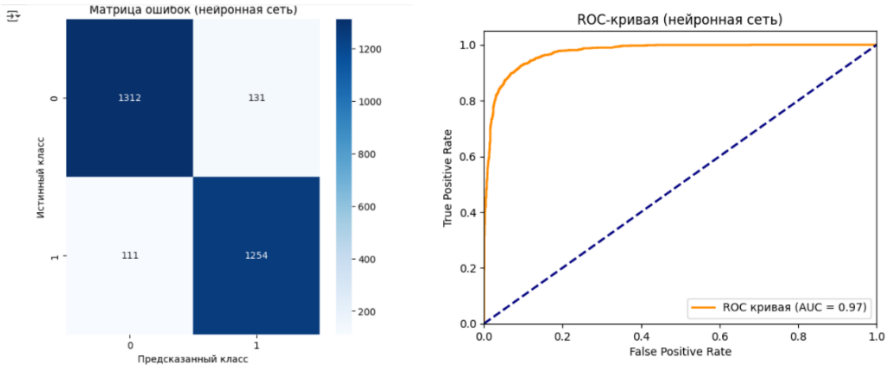


Рис.3. Матрица погрешностей и ROC кривая построенной модели.

Аналогичный результат ниже показывает алгоритм машинного обучения случайный лес, точность в данном случае достигает 98%. Приведем результаты прогноза случайного леса.

Листинг 2. Прогнозирование на основе случайный лес.

```
Случайный лес (классификация):
              precision    recall  f1-score   support

           0       0.94      0.91      0.92      1443
           1       0.90      0.93      0.92      1365

    accuracy                           0.92      2808
   macro avg       0.92      0.92      0.92      2808
weighted avg       0.92      0.92      0.92      2808
```

Матрица погрешностей выглядит так

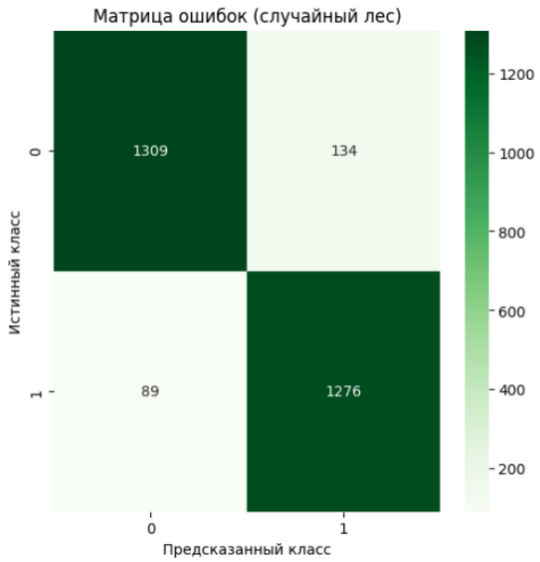


Рис.4. Матрица погрешностей модели с применением случайный лес.

2. Прогнозирования аварийности в энергетических системах.

Рассмотрим задачу прогнозирования аварийности в энергетических системах.

Модель ARIMA

Изучим регрессию скользящего среднего обычно описывается системой ARIMA и широко используется почти во всех краткосрочных и долгосрочных прогнозировании. Для прогнозирования с данной системой используют сезонные и не сезонные прогнозирование. У нее есть 3 составляющие p, d, q, где p — порядок члена AR, q — порядок члена MA, а d — количество разностей, которые преобразуют ряд стационарным. В случае когда требуется сезонность его преобразуют в Сарима. Вот задача влияние на прогнозирования загрязнения атмосферы в зависимости от времени на энергетические системы.

```
df = df.sort_values(by=['Date'])
df.head()
```

	Date	Total_Accident
0	2014-01-01	267
31	2014-01-02	409
59	2014-01-03	392
90	2014-01-04	432
120	2014-01-05	434

Здесь представлены графические это представляется в виде графика.

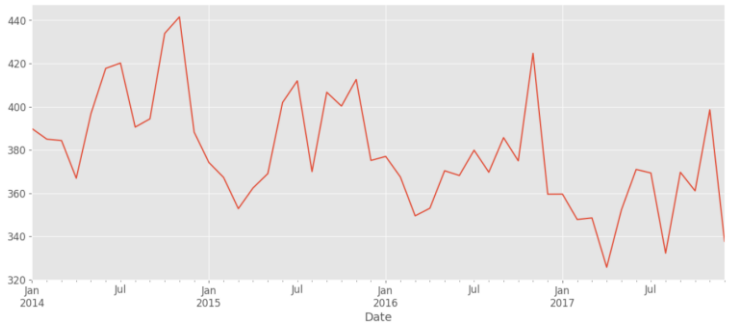


Рис.5. Влияние загрязнение воздуха на системы энергетики.

Установим дату для индекса. Вот результаты.

```
Index(['01/01/2014', '02/01/2014', '03/01/2014', '04/01/2014', '05/01/2014',
       '06/01/2014', '07/01/2014', '08/01/2014', '09/01/2014', '10/01/2014',
       ...
       '22/12/2017', '23/12/2017', '24/12/2017', '25/12/2017', '26/12/2017',
       '27/12/2017', '28/12/2017', '29/12/2017', '30/12/2017', '31/12/2017'],
      dtype='object', name='Date', length=1461)
```

Вот данные, качества показателя потери в энергетических системах в определенный период времени и ее значение.

Date	Total_Accident
2014-01-01	389.870968
2014-02-01	385.000000
2014-03-01	384.354839
2014-04-01	366.933333
2014-05-01	396.870968

dtype: float64

Важная задача обнаружение места и время аварийности энергетических систем. Здесь ниже указана общая аварийность энергетических систем по месяцам

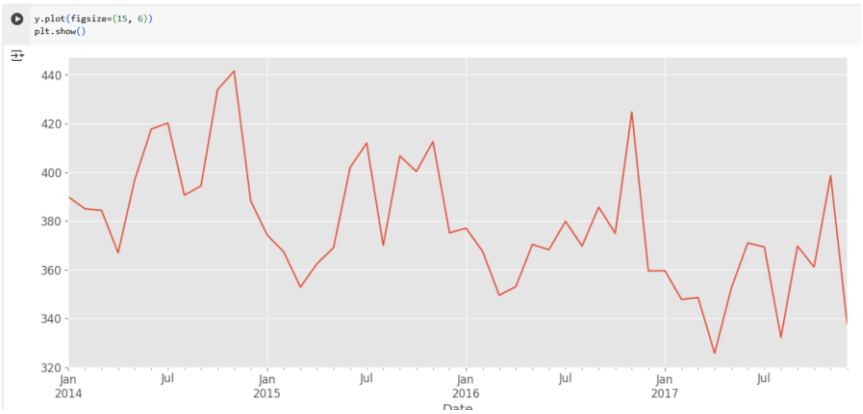


Рис.6. Ситуация аварийности в энергетических системах.

Давайте визуализируем данные, используя декомпозицию временного ряда, которая позволяет нам разложить наш временной ряд на три отдельных компонента: тренд, сезонность и шум. Один из важных показателей во временных рядах тренд, сезонность и сложность аварии.

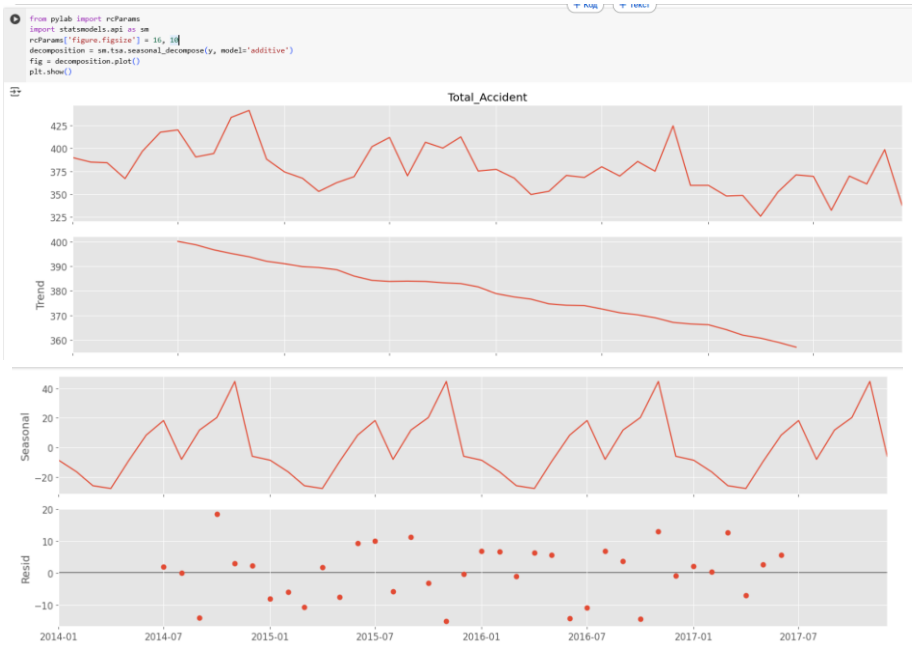


Рис.7. Это тренды аварийности, сезонности и сложность аварий в энергетических системах.

Ниже указано регрессионная модель. Которая построена на основе машинного обучения. Здесь проведена анализ поведения кривой аварийности. диагностику модели, чтобы исследовать любое необычное поведение энергетической сети. Вот построенный регрессионный модель и гистограмма данных аварийности системы.

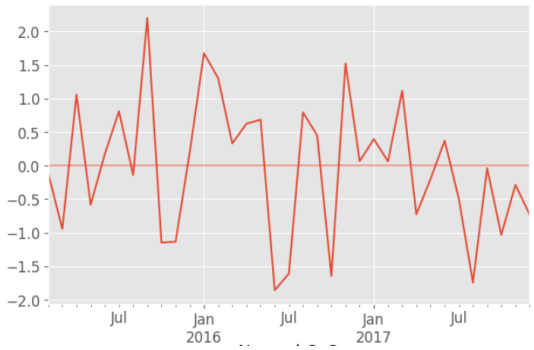


Рис.8. Это тренды аварийности, сезонности и сложность аварий в энергетических системах.

Во многих случаях требуется общий анализ данных требует на ее подчинённость критериям Пуассона, это указывает на дискретность распределения аварийности в системах и ее вероятностное распределения. Ее дискретность указывает на вероятностное распределение событий произошедшее за фиксированный промежуток времени. И она является критерием сходимости временного ряда. Ниже представлена полная картина история аварийности системы. Частота появлений аварийности сужено до отрезка 0,1.

Во всех прогнозах с временными рядами очень важно тестировать модель на данных, которые не участвовали в процессе обучения. Ниже приставлен процесс прогнозирования на данных, которые вообще не участвовали и валидации и при обучении модели для системы. Конкретные даты, взятые из открытых систем указано в прогнозе ниже построенной с помощью системы Арима.

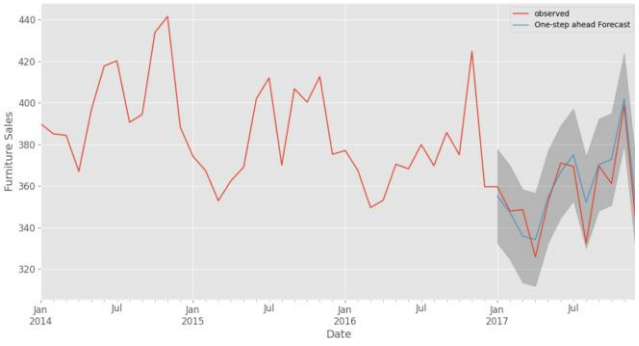
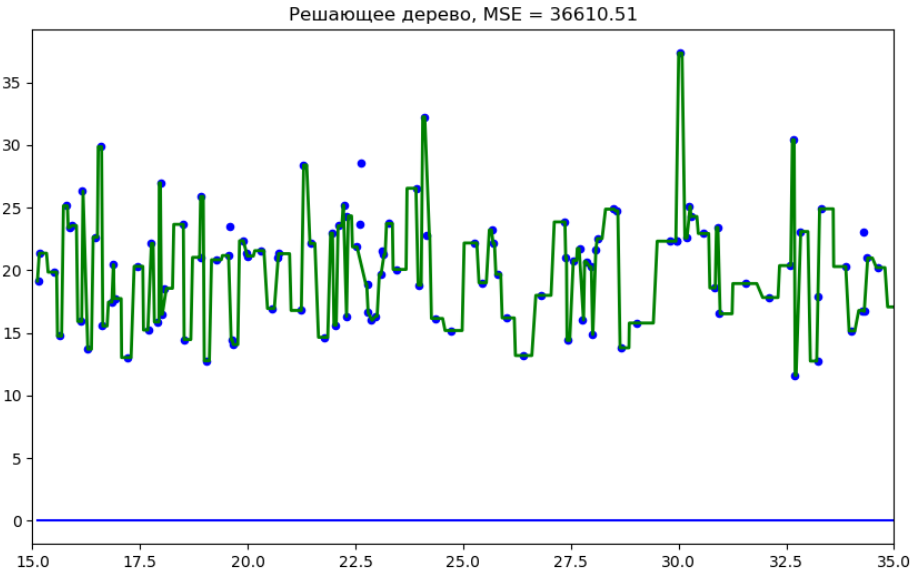


Рис.9. Прогнозирование с применением системы Арима аварийности энергетической системы.

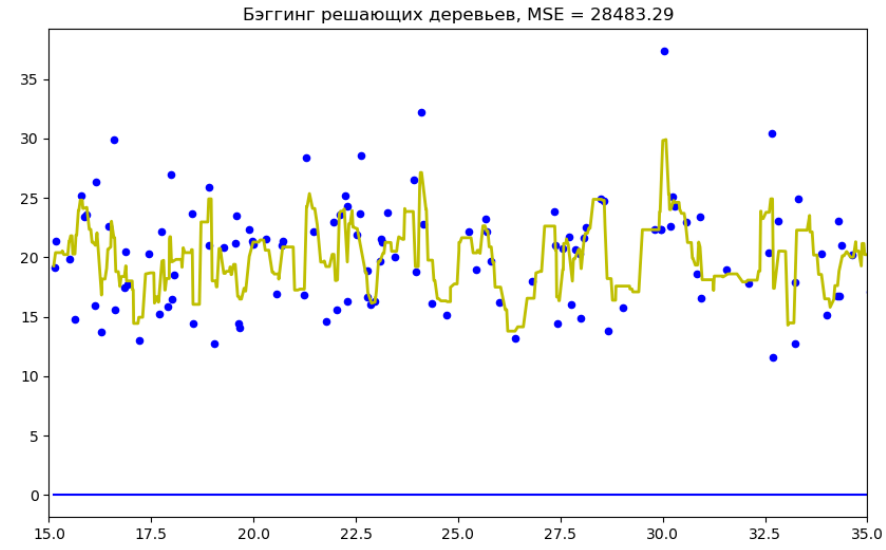
3. Реализация машинного обучение для задач растениеводства.

Используются алгоритмы машинного обучения для задач растениеводства. Рассмотрим применение алгоритма случайный лес. Данный алгоритм случайный лес — это метод применения бэггинга над решающими деревьями. Здесь чрезвычайно важно, что при обучении модели, множество признаков выбираются как подмножество общего признакового пространства.

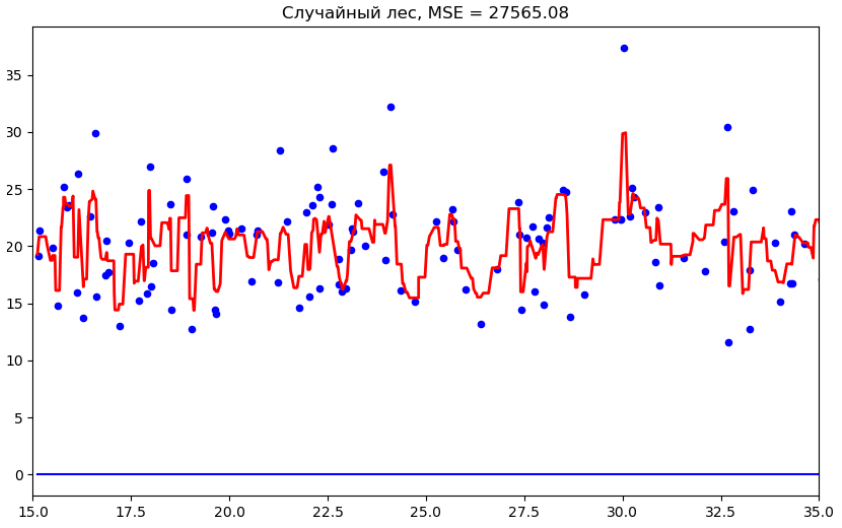
Ниже в рисунках для прикладной задачи урожайности в сельском хозяйстве построены классификаторы и проведен сравнительный анализ алгоритмов решающих деревьев, беггинга и случайный лес.



a)



b)



c)

Рис.10. Реализация задач регрессии дерево решений (a), беггинга(b) и случайный лес(c).

Ниже представлен применение данного метода к задаче регрессии сельского хозяйства. Например, результат прогноза урожайности с полным набором данных, где у означает урожайность в тоннах картофеля в зависимости от обработанной площади x в га с применением данной технологии, результат реализации Лассо представлено на следующем графике 11.

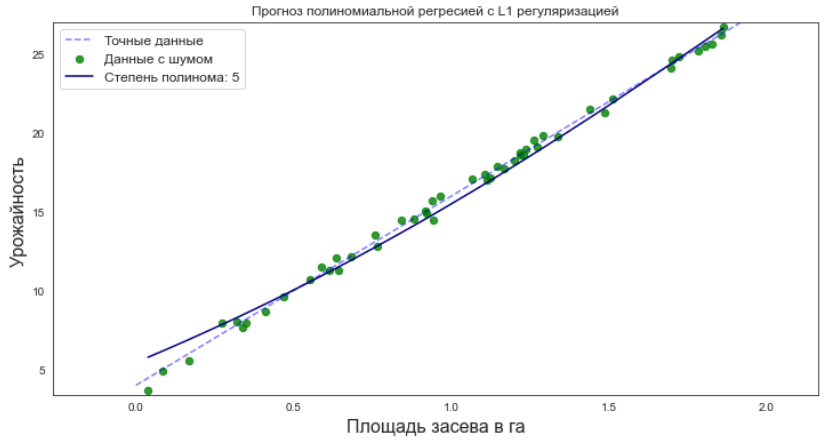


Рис.11. Результат применения технологии регуляризации L_1 для прогнозирования урожайности картофеля

Рассмотрено прогнозирование влияния количества вносимых пестицидов с применением полиномиальной регрессии на почвенные характеристики с применением регуляризации Тихонова-Риджа, выглядит следующим образом

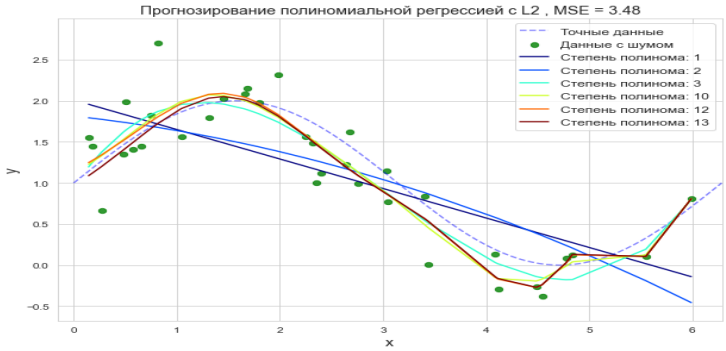


Рис.12. Результат применения технологии регуляризации для прогнозирования влияние пестицидов на урожайность картофеля

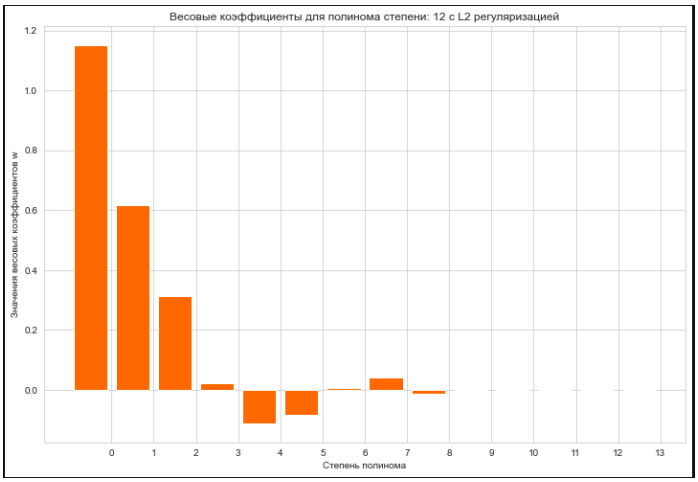


Рис.13. Значения весовых коэффициентов полиномов высокой степени после применения L_2 регуляризации

Ниже приведен метод регуляризации в L_1.

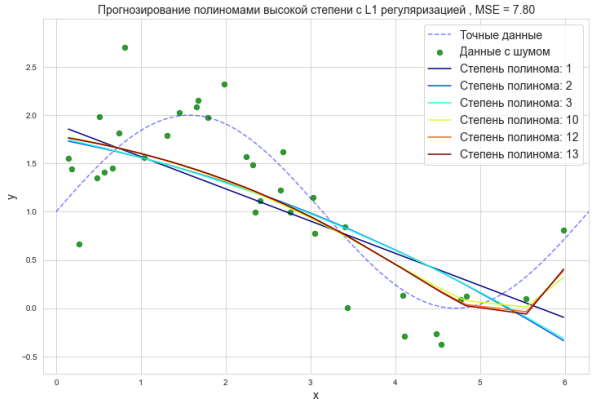


Рис.14. Значения весовых коэффициентов полиномов высокой степени после применения L_1 регуляризации

Рассмотрены различные прикладные задачи. Построены различные модели с применением различных алгоритмов машинного обучения. Основная цель работы была направлена на реализацию надежной модели машинного и глубокого обучения, которые могут эффективно

прогнозировать прикладные задачи, в частности качества воздуха, аварийность в энергетических системах, моделирования в растениеводстве и основы модели нейронных сетей. Получены следующие результаты.

• Сфокусировано на раннем обнаружении качества воздуха, что имеет решающее значение для охраны общественного здоровья и окружающей среды, с потенциалом спасти миллионы жизней во всем мире.

• Разработано передовая система прогнозирования прикладных задач , которая объединяет различные методы машинного обучения, включая регрессионной модели, случайный лес, опорную векторную машину, дерево решений, достигая замечательной высокой точности классификации как с моделями случайного леса, так и дерева решений.

• Построены нейронная сеть для прогнозирования конкретных задач

• Проведен сравнительный анализ нейронного и машинного моделирования прикладных задач

• Для переобученных моделей рассмотрен метод регуляризации.

• Исследован класс задач применительно к растениеводству.

ЛИТЕРАТУРА

1. Сабитов, Б.Р. Идентификация болезней томатов на основе многоклассовой классификации. [Электронный ресурс] / Б.Р.Сабитов, Н.С.Сейтказиева, А.Дж. Картанова // Проблемы автоматики и управления 2022. - № 3(45). – С.11.-Режим доступа: https://www.elibrary.ru/item.asp?id=50020292 . – Загл.с экрана.
2. Сабитов, Б.Р. Моделирование и прогнозирование задач сельского хозяйства на основе машинного обучения [Электронный ресурс] / Б.Р.Сабитов // Тр.Межд. научно-практ.конф. Научно-технологическое развитие АПК для целей устойчивого развития. – Режим доступа: https://www.conferences.org/articles/e3sconf/abs/2023/17/contents/contents.html. – Загл.с экрана.
3. Sabitov, B.R. Deep learning Methods for Recognition of Orchard Crops [Электронный ресурс] / B.R. Sabitov, S.Biibsunova, A.Kashkaroeva et al. // IJCSNS. – 2022. – Vol.22. – No.10. – Режим доступа: http://paper.ijcsns.org/07_book/202210/20221033.pdf . – Загл.с экрана.
4. Pu, Y. Variational autoencoder for deep learning of images, labels and captions [EB/OL] [Text] / Y.Pu, Z.Gan, R.Henano et al. // 2016. – 09.28. – arxiv. – No.1609.08976.