Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Comparative Analysis of ChatGPT-Generated Code and Kaggle Champion Performance in Water Potability Prediction: A Few-Shot Learning prompts Approach

Malusi Sibiya

Computer Science Department, University of South Africa, Florida, South Africa
Email: sibiym@unisa.ac.za

This research compares the predictive performance of machine learning models generated by ChatGPT, an AI language model, with those developed by Kaggle champions in predicting water potability. Utilizing few-shot learning prompts, ChatGPT constructs predictive models with minimal dataset information, including total columns, missing values, and total entries. Incorporating state-of-the-art regularization techniques, ChatGPT-generated models aim to enhance predictive accuracy. Performance metrics such as accuracy, precision, recall, F1 score, and area under the ROC curve are evaluated to assess model effectiveness. The findings indicate the potential of AI language models like ChatGPT in rapidly constructing predictive models with limited information, providing insights into their comparative performance against human-developed models in real-world predictive tasks.

Keywords: ChatGPT, Few-shot learning, Machine Learning, Natural Language Processing, Water potability.

1. Introduction

The term "potability" refers to whether water is safe and clean enough for people to drink without causing harm or illness. It involves assessing various scientific and regulatory factors to ensure that water meets specific standards and doesn't pose health risks to consumers. Recent research by Yaroshenko et al. has observed advancements in methods for detecting water quality. Their study offers a comprehensive review of the suitability of various technologies for real-time water quality monitoring, particularly those tested in practical settings. The performance of sensors based on molecularly imprinted polymers is extensively evaluated, shedding light on their operational principles, stability in real-world applications, and potential for mass production [1]. Their study not only carefully evaluates the performance of sensors based on molecularly imprinted polymers but also highlights the practical implications of these findings. The research underscores their potential significance in addressing pressing issues related to water quality monitoring. ML solves many problems in these days. An online platform called Kaggle(https://www.kaggle.com/) always endeavors to initiate competitions whenever datasets of high importance become available. Kaggle is an online platform renowned for hosting data science competitions, offering datasets, collaborative coding environments, and forums, thereby fostering a global community of data scientists and machine learning practitioners engaged in collaborative problem-solving and skill development. In the Kaggle leaderboard for the competition of predicting water potability, the highest ROC AUC score at the time of writing this paper was 99.5%, achieved by the authors who identified themselves as EDA&GBC (https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability/code). The code with solutions is also available on this platform. This research aims to compare the predictive performance of machine learning models generated by ChatGPT [2], an AI language model, against those developed by Kaggle champions EDA&GBC in predicting water potability. In the methodology section, we briefly explain how the few-shot learning prompts, carefully designed from the water potability dataset, are used to generate complete code, the ROC AUC score of which will be compared with that of the authors leading in the Kaggle leaderboard in the prediction of water potability competition. The results suggest that AI language models such as ChatGPT have the capability to swiftly construct predictive models using restricted information, thereby shedding light on their relative efficacy compared to models developed by humans in practical predictive endeavors. In the next section, a review of research on water potability and the contemporary applications of ChatGPT is conducted.

2. Literature Review

A significant amount of research has been dedicated to water quality assessment and monitoring. Evidence in the literature suggests that traditional scientific methods of water quality assessment have been supplanted by artificial methods, involving the use of sensors and ML algorithms in the monitoring processes. However, it is important to note the distinction between water quality and water potability. To distinguish between water quality and water potability, this section reviews the approaches employed by researchers in these two disciplines. The efficacy of machine learning (ML) techniques for water quality prediction was investigated in a recent study by Yaroshenko et al. A machine learning classifier model was constructed using real-world data, with all measured characteristics utilized as significant features. The dataset was partitioned into training and testing subsets, and various ML algorithms were employed, with support vector machine and k-nearest neighbor demonstrating superior performance in terms of F1-score and ROC AUC values. Conversely, the LASSO-LARS and stochastic gradient descent methods exhibited higher recall values [1]. The Root Zone Water Quality Model (RZWQM), developed by USDA–ARS scientists, integrated physical, chemical, and biological processes to simulate water and agrochemical movement in agricultural fields. Ahuja et al. evaluated the model's performance using field data, demonstrating reasonable simulation of soil water movement and pesticide persistence [3]. Shrestha and Kazama applied multivariate statistical techniques to assess temporal and spatial variations in water quality within the Fuji River basin. Cluster analysis categorized sampling sites into pollution gradient clusters, while factor analysis revealed key factors driving water quality variations across different pollution levels. Discriminant analysis effectively reduced data dimensionality and identified indicator parameters for water quality assessment, pollution source identification, and river water quality management [4]. The utilization of multivariate statistical techniques for the evaluation and interpretation of water quality datasets was highlighted in the study by researchers focusing on the Gomti river in India [5]. Through cluster analysis, distinct groups within the river's catchment regions were identified, while factor analysis/principal component analysis revealed key factors responsible for variations in water quality across different catchment areas. Discriminant analysis facilitated data reduction and pattern recognition, aiding in the identification of indicator parameters for effective water quality management. Additionally, receptor modeling techniques provided insight into pollution sources/factors contributing to river contamination. The Athens Water Supply and Sewerage Company (EYDAP SA) played a crucial role in supplying potable water to millions of inhabitants in Attica, Greece [6]. Stringent quality control measures were enforced, with thorough analysis conducted to ensure compliance with established guidelines. Statistical tools were employed to enhance quality control processes, with particular attention given to parameters such as turbidity, residual chlorine, and aluminum levels. Statistical process control techniques were utilized to evaluate control limits and improve process quality. In addressing the imperative need for accurate detection and identification of contaminants in drinking water, a real-time event adaptive detection, identification, and warning (READiw) methodology was explored [7]. Through pilot-scale pipe flow experiments, various chemical and biological contaminants were examined, with adaptive transformation techniques enhancing sensor detection capabilities. Kinetic and chemical differences among contaminants allowed for their distinguishability, providing a reliable method for contamination event identification. The optimization and artificial intelligence (AI) techniques applied in the simulation and operation of the Barra Bonita reservoir in Brazil were elucidated in the methodology proposed by Chaves et al. [8]. A fuzzy stochastic dynamic programming model was developed to calculate optimal operation procedures, considering multiple fuzzy objectives. Markov chain technique handled the stochastic nature of river flow, while water quality analysis employed artificial neural network models to predict organic matter and nutrient loads based on river discharge. The proposed methodology demonstrated efficacy in reservoir operation, providing a valuable tool for water resource management. The necessity for comprehensive information on water quality, especially concerning sediment, phosphorus, and nitrogen exports from catchments, is emphasized by catchment managers and stakeholder groups [9]. Due to the limited availability of intensive spatial and temporal data on nutrient concentrations or loads, there's a demand for nutrient export models capable of providing valuable insights with sparse data inputs. This paper evaluates four such models and various direct estimation methods for their efficacy in predicting loads in Australian catchment scenarios. The discussion underscores the significance of coordinated data collection over extended periods and fine temporal scales to improve load prediction accuracy. Artificial neural network procedures were employed to define and predict diatom assemblage structures in Luxembourg streams using environmental data [10]. Self-organizing maps (SOM) classified samples based on their diatom composition, while a multilayer perceptron with a backpropagation learning algorithm (BPN) predicted these assemblages. Classical methods were then utilized to identify relationships between diatom assemblages and SOM cell numbers. The study demonstrated high predictability of diatom assemblages using physical and chemical parameters within a limited geographical area. In planning sampling regimes, minimizing estimation error or sampling effort for a desired accuracy is essential [11]. This paper compares classical and geostatistical approaches for matching sampling effort to accuracy using airborne thematic mapper images of British lakes. It illustrates that the systematic scheme outperforms the random scheme, especially with increasing sample size and spatial dependence. The study underscores the necessity of calibrating sampling regimes to the spatial dynamics of the lake and suggests remote sensing as an ideal means for determining such dynamics. A pilot study was conducted to assess the hormonal activity of freshwaters in Victoria using recombinant receptor-reporter gene bioassays [12]. Water samples from the Yarra River were analyzed for toxicity, genotoxicity, and receptor assay activity. Results indicated weak to moderate toxicity with no significant location-based trends along the river. Estrogenic, thyroid, and retinoic acid receptor activity was negligible, while AhR activity increased downstream, possibly influenced by bush fires. Approximately 24% of total AhR activity was associated with suspended solids. The preceding reviews focused on evaluating various aspects of water quality. In the following reviews, the focus shifts towards assessing water potability. These reviews examine the suitability of water for human consumption, considering factors such as chemical composition, microbial contamination, and adherence to regulatory standards. Nyende-Byakika et al. provide insights into the raw water quality of Bospoort dam in South Africa [13]. Through a comprehensive time-series analysis, various parameters were monitored, revealing that while most parameters remained within recommended threshold levels for the majority of the study period, conductivity, hardness, and high coliform counts exceeded acceptable limits. The water exhibited excessive hardness and high conductivity, surpassing alarm levels for a considerable portion of the study duration despite dissolved solids being below their alarm thresholds. Notably, total coliform and E. coli counts were found to be significantly elevated, indicating potential microbial contamination concerns. Pehlivan and Emre investigate the environmental and hydrological processes in the Sarma Stream basin, located southwest of Akcakoca in the Duzce Province of Turkey [14]. Samples from various sources, including rocks, soil, stream water, rain, snowmelt, and bed and suspended sediment, were collected and analyzed. The study reveals that sandstone and soil samples contribute to the stream's muddy flow during the rainy season, with chlorite-type minerals prevalent in the bed and suspended sediments. The water chemistry indicates a calcium bicarbonate-rich composition, influenced by acid rain and containing elevated levels of certain heavy metals and elements, necessitating treatment of water in the Sariyayla Reservoir. Comparatively, a study by an undisclosed author assesses potable water filtration methods commonly used in rural Ghanaian communities [15]. Physico-chemical and microbiological analyses were conducted on water samples from ponds, dams, and rivers, revealing elevated levels of total suspended solids, turbidity, total coliforms, and bacterial counts. However, filtration methods, including ceramic filters and household sand filters, effectively reduced these parameters to acceptable standards. The study suggests that a combination of filtration methods, including the use of alum and activated carbon, could further improve water quality, recommending follow-up research in this area. Elizabeth and Rajpramikh monitored the microbiological and physico-chemical parameters of drinking water samples from two villages in the Vizianagaram District. Borewell water from Somi

Naiduvalasa village exhibited elevated coliform levels and exceeded permissible limits for various physic-chemical parameters, rendering it non-potable [16]. In contrast, the historical shift in perceptions of water quality assessment was discussed, highlighting the transition from sensory-based evaluations to reliance on standardized analytical methods. This shift, driven by institutional and regulatory practices, marginalized consumer sensory knowledge as merely aesthetic, focusing instead on objective analytical data. However, the exclusion of sensory information from water quality assessment overlooks the subjective experiences of consumers, calling for new practices that engage consumers as valuable participants in ensuring water quality [17]. Furthermore, an evaluation of the water potability of various regions in Ludhiana, Punjab, revealed suboptimal potability levels despite acceptable hardness and pH values. Physicochemical and bacteriological analysis conducted across six areas of Ludhiana city showed low levels of potability, highlighting the need for interventions by local water authorities to ensure the supply of safe drinking water to the population [18]. The integration of sensor materials into new-generation transducers and the use of household electronic devices for signal registration offer potential for the development of economical, portable detectors operating in real-time mode [19]. In another study, physico-chemical and microbial parameters of water quality in hand-dug wells in Bolgatanga, Ghana, were assessed. The study revealed elevated coliform levels in dry seasons and increased concentrations of various parameters during the rainy season, suggesting infiltration from stormwater. Proximity to pollution sources also influenced coliform counts, indicating the need for disinfection of well water before use [20]. Similarly, research on the potability of packaged sachet water within the Federal University of Agriculture, Abeokuta campus, Nigeria, found that while physicochemical parameters met WHO and Nigerian standards, bacteriological analysis revealed total bacteria count in all samples and contamination with total coliforms in two brands. The study underscores the importance of routine water quality examination and regulatory oversight to ensure safe drinking water supply [21]. Assessment of water quality in a village involved analyzing various physicochemical parameters and calculating a water quality index. While most parameters met Indian standards, coliform levels exceeded permissible limits, indicating contamination and leading to waterborne diseases. Although some water sources were classified as excellent, disinfection of coliform before use was recommended [22]. Furthermore, groundwater samples from different areas in Ariyalur District, Tamil Nadu, were analyzed for various physicochemical parameters. The majority of samples were found unsuitable for drinking purposes, highlighting the need for comprehensive water quality management [23]. Thus far, researchers have found ChatGPT to be beneficial in several areas, leading to its application in various research endeavors [24-29].

3. Methodology

The methodology section of this study outlines a systematic approach to address the research objectives regarding the predictive modeling of water potability. Specifically, it details the methodology adopted to assess and compare models generated by ChatGPT, an AI language model, with those developed by human experts, notably the Kaggle champions EDA&GBC, in the prediction of water potability. Through a thorough examination of the methodologies used in constructing predictive models and generating associated code, this study aims to uncover the subtle differences in performance between AI-generated models and those crafted by human experts. Central to this methodology is the use of few-shot learning prompts derived from the water potability dataset, which facilitate the rapid construction and evaluation of predictive models. By adhering to established scientific principles and employing robust analytical techniques, this methodology aims to provide comprehensive insights into the comparative predictive performance of ChatGPT-generated models and those produced by human experts in practical applications.

3.1. An experimental approach by EDA&GBC Kaggle champions

The Python code by the authors is available on Kaggle platform (https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability/code) outlines an approach aimed at tackling a classification problem concerning water potability. Initially, the dataset was loaded and explored using libraries like Pandas for data manipulation and Seaborn for visualization. Basic data checks were conducted to identify any missing values, and an overview of the data types was obtained. Visualization techniques such as count plots, pair plots, and histograms were employed to gain insights into the distribution of features and the target variable, facilitating a better understanding of the dataset's characteristics. Additionally, observations suggested a Gaussian-like distribution in some parameters, prompting the consideration of standardization using StandardScaler during preprocessing to ensure consistency in feature scaling. Subsequently, outlier detection and removal were performed utilizing KMeans clustering. By assigning data points to clusters, outliers were identified and excluded from further analysis. Following outlier removal, the dataset was prepared for model training by partitioning it into training and testing sets. To enhance model performance, feature standardization using StandardScaler was applied, aiming to mitigate potential issues arising from varying feature scales. Various classification algorithms, including Logistic Regression, Support Vector Machine (SVM), Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, AdaBoost Classifier, and K-Nearest Neighbors Classifier, were imported from the Scikit-learn library for model training. Functions were defined to evaluate model performance metrics such as accuracy, precision, recall, and F1-score for each algorithm. These metrics provided insights into the models' predictive capabilities and aided in selecting the most suitable model for the classification task. Upon training and evaluation, the Gradient Boosting Classifier emerged as a primary candidate, demonstrating promising performance. Visualization techniques, including the plotting of confusion matrices and Receiver Operating Characteristic (ROC) curves, were employed to assess the model's performance visually. Finally, the Area Under the Receiver Operating Characteristic Curve (ROC AUC) was calculated to quantitatively evaluate the Gradient Boosting Classifier's ability to predict water potability probabilities. This metric offered a comprehensive measure of the model's discriminative power, further validating its efficacy in addressing the classification problem at hand.

3.2. An experimental approach by ChatGPT through a few-shot learning prompts

Initially, we imported the water potability dataset utilizing the Pandas library, accessing it from its designated location within the computer and transferring it to the working environment of the Jupyter notebook. Subsequently, we proceeded to exhibit the information encapsulated within the water potability dataset. The insights gleaned from this presentation of dataset information served as a foundational basis for formulating instructional prompts tailored to the framework of few-shot learning, guiding ChatGPT to harness a comprehensive array of regularization methods, address outliers effectively, and ascertain the Receiver Operating Characteristic Area Under the Curve (ROC AUC) metric. Once ChatGPT generated the code as guided by the few-shot learning prompts, we incorporated the water potability dataset and ran it to get the results. Few-shot learning prompts involve presenting ChatGPT with tasks or examples along with corresponding labels when making prompts. Figure 1 shows the displayed dataset information that was used for designing a few-shot learning prompts based on the dataset.

```
water_port.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          2495 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64
dtypes: float64(9), int64(1)
memory usage: 256.1 KB
```

Figure 1. Dataset info used to develop few-shot learning prompts.

The highlighted data entries in Figure 1 were used to design prompts that informed ChatGPT in terms of how many entries were missing in those columns. The main prompt depicted in Figure 2 was utilized to generate the optimal ChatGPT code, which randomly selected the Random Forest Classifier.

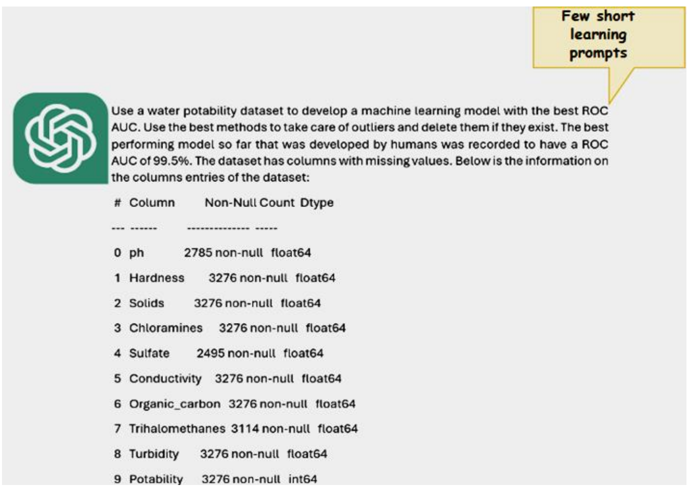

Figure 2. Depiction of a few-short learning main prompt

In the subsequent prompt, ChatGPT was provided with the ROC AUC results from the previous experiment and tasked with enhancing the ROC AUC based on the reported outcomes. The subsequent prompts were derived from inputting ChatGPT with the previous ROC AUC results of the generated code. Figure 3 depicts this procedure.

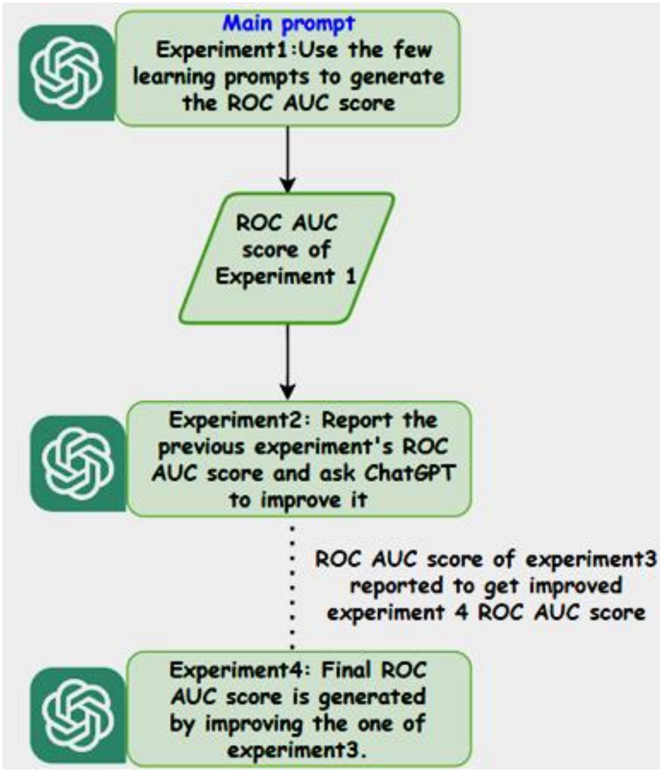

Figure 3. A complete prompt design procedure.

Below are the actual experimental prompts that were used on ChatGPT to get the ROC AUC scores.

Main prompt(experiment1)-Prompt: ‘Use a water potability dataset to develop a machine learning model with the best ROC AUC. The best performing model so far that has been by humans was recorded to have a ROC AUC of 99.5%. The dataset has columns with missing values. Below is the information on the column’s entries of the dataset:

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	ph	2785 non-null	float64
1	Hardness	3276 non-null	float64
2	Solids	3276 non-null	float64
3	Chloramines	3276 non-null	float64

4	Sulfate	2495 non-null	float64
5	Conductivity	3276 non-null	float64
6	Organic_carbon	3276 non-null	float64
7	Trihalomethanes	3114 non-null	float64
8	Turbidity	3276 non-null	float64
9	Potability	3276 non-null	int64’

Experiment 2-Prompt: ‘Your ROC AUC Score was 0.6798852061117301. Can you improve it?’

Experiment 3-Prompt: ‘Your Improved ROC AUC score was 0.6993474454878243, improve it again.’

Experiment 4-Prompt: ‘Your Further Improved ROC AUC Score was 0.6998249243991723. Get a final further improved score that can be around 0.9.’

The performance scores of the models, as ChatGPT was provided with the ROC AUC results, were recorded, and the findings are presented and discussed in Section 4.

4. Results and Discussion

Following the procedure outlined in the Kaggle platform, the EDA&GBC Kaggle champions attained a ROC AUC score of 99.5%. The python code by the authors is available at Kaggle platform and be accessed anytime. (https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability/code). At the time of writing this paper, this achievement represented the highest score on the Kaggle leaderboard. In this section, we present and analyze the results of the ChatGPT-generated models in comparison with those of the EDA&GBC Kaggle champions. Initially, we elaborate on the procedures and the classifiers selected by ChatGPT for each experiment.

4.1 Experiment 1 – ChatGPT’s code generation approach and results

The generated code aimed to construct a machine learning model using the Random Forest Classifier to predict water potability. It followed a systematic approach: initially, the dataset was loaded and split into features and target variables, with subsequent division into training and testing sets for model evaluation. Missing values were addressed by replacing them with feature means, followed by feature scaling for uniformity. The Random Forest Classifier, comprising 100 decision trees, was then trained on the data. Predictions were made on the testing set to assess model performance, evaluated using the ROC AUC score, indicating its ability to discern between positive and negative classes. This comprehensive process aimed to create a robust model for water potability prediction. The achieved ROC AUC score achieved with this approach was 68%.

4.2 Experiment2 – ChatGPT’s code generation approach and results

The procedure employed by ChatGPT to generate the code involved several steps. Initially, the dataset was loaded using Pandas, and missing values were handled through mean imputation. Subsequently, the features were separated from the target variable, and the dataset was split into training and testing sets using the train_test_split function from scikit-learn. Feature scaling was performed using StandardScaler to ensure uniformity in feature scales. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied for oversampling. A Random Forest Classifier with 100 estimators was then trained on the resampled training data. Finally, the classifier's performance was evaluated using the ROC AUC metric and the score of 70% resulted.

4.3 Experiment3 – ChatGPT’s code generation approach and results

ChatGPT’s procedure to generate the code involved several steps. Initially, the dataset was loaded using the Pandas library, and missing values were handled through mean imputation. Subsequently, the features were separated from the target variable, and the dataset was split into training and testing sets using the train_test_split function from scikit-learn. Feature scaling was performed using StandardScaler to ensure uniformity in feature scales. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied for oversampling. A Random Forest Classifier was then defined, and hyperparameters for grid search were specified. Grid search was conducted with cross-validation to find the best combination of hyperparameters. The best model from the grid search was identified, and its performance was evaluated by predicting probabilities on the test set and calculating the ROC AUC score. Finally, ROC AUC score of 70% resulted showing no change when compared to the results of the previous experiment (Experiment 3).

4.4 Experiment4 – ChatGPT’s code generation approach and results

The procedure undertaken by ChatGPT to develop the generated code involved several sequential steps. Initially, the dataset was imported using the Pandas library, with missing values being handled through mean imputation using the SimpleImputer from scikit-learn. Following this, the dataset was split into features and the target variable, with the features being separated from the target variable. Subsequently, the data was divided into training and testing sets using the train_test_split function, and feature scaling was performed using StandardScaler to maintain uniformity in feature scales. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was implemented for oversampling. The Gradient Boosting Classifier was defined, and hyperparameters for grid search were specified to fine-tune the model. Grid search was then conducted with cross-validation to identify the optimal combination of hyperparameters. The best model resulting from the grid search was selected, and its performance was evaluated by predicting probabilities on the test set and calculating the ROC AUC score. However, the ROC AUC dropped to 66.5%. Figure 4 illustrates the results of ROC AUC scores by the ChatGPT classifier models in all four experiments.

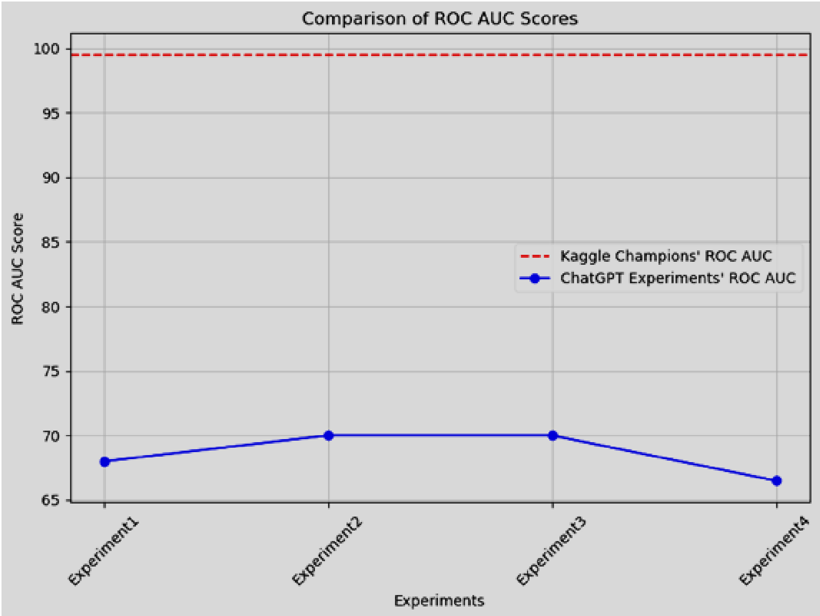

Figure 4. Comparison of ROC AUC scores.

The experiments compared the predictive performance of machine learning models developed by Kaggle champions with those generated by ChatGPT, an AI language model. ChatGPT utilized few-shot learning prompts to construct predictive models with minimal dataset information and incorporated state-of-the-art regularization techniques to enhance predictive accuracy. The experiments yielded varying ROC AUC scores for ChatGPT, ranging from 66.5% to 70%, while the ROC AUC score achieved by Kaggle champions stood at 99.5%. Despite ChatGPT's ability to rapidly construct predictive models with limited information, the results suggest that the models developed by Kaggle champions outperformed those generated by ChatGPT in terms of ROC AUC score. However, it's noteworthy to appreciate the existence of ChatGPT and the efforts it made to approach the ROC AUC performance of the Kaggle champions, indicating the potential of AI language models in real-world predictive tasks. Further analysis may be required to explore additional performance metrics and potential areas of improvement for both approaches.

References

1. Yaroshenko, I., Kirsanov, D., Marjanovic, M., Lieberzeit, P. A., Korostynska, O., Mason, A., ... & Legin, A. (2020). Real-time water quality monitoring with chemical sensors. Sensors, 20(12), 3432.
2. OpenAI. (2024). ChatGPT [3.5]. Retrieved from [https://chat.openai.com/c/53c0468f-e40d-439c-a90b-e224d64afdc8]
3. Ahuja, L. R., Ma, Q. L., Rojas, K. W., Boesten, J. J., & Farahani, H. J. (1996). A field test of root zone water quality model—pesticide and bromide behavior. Pesticide Science, 48(2), 101-108.
4. Shrestha, S., & Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji River basin, Japan. Environmental Modelling &

Software, 22(4), 464-475.
5. Singh, K. P., Malik, A., & Sinha, S. (2005). Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques—a case study. Analytica Chimica Acta, 538(1-2), 355-374.
6. Smeti, E. M., Thanasoulias, N. C., Kousouris, L. P., & Tzoumerkas, P. C. (2007). An approach for the application of statistical process control techniques for quality improvement of treated water. Desalination, 213(1-3), 273-281.
7. Yang, Y. J., Haught, R. C., & Goodrich, J. A. (2009). Real-time contaminant detection and classification in a drinking water pipe using conventional water quality sensors: Techniques and experimental results. Journal of environmental management, 90(8), 2494-2506.
8. Chaves, P., Tsukatani, T., & Kojiri, T. (2004). Operation of storage reservoir for water quality by using optimization and artificial intelligence techniques. Mathematics and Computers in Simulation, 67(4-5), 419-432.
9. Gevrey, M., Rimet, F., Park, Y. S., Giraudel, J. L., Ector, L., & Lek, S. (2004). Water quality assessment using diatom assemblages and advanced modelling techniques. Freshwater biology, 49(2), 208-220.
10. Letcher, R. A., Jakeman, A. J., Calfas, M., Linforth, S., Baginska, B., & Lawrence, I. (2002). A comparison of catchment water quality models and direct estimation techniques. Environmental Modelling & Software, 17(1), 77-85.
11. Hedger, R. D., Atkinson, P. M., & Malthus, T. J. (2001). Optimizing sampling strategies for estimating mean water quality in lakes using geostatistical techniques with remote sensing. Lakes & Reservoirs: Research & Management, 6(4), 279-288.
12. Allinson, M., Shiraishi, F., Kamata, R., Kageyama, S., Nakajima, D., Goto, S., & Allinson, G. (2011). A pilot study of the water quality of the Yarra River, Victoria, Australia, using in vitro techniques. Bulletin of environmental contamination and toxicology, 87, 591-596.
13. Nyende-Byakika, S., et al. "Potability analysis of raw water from Bospoort dam, South Africa." Water Practice and Technology 11.3 (2016): 634-643.
14. Pehlivan, R., & Emre, H. (2017). Potability and hydrogeochemisty of the Sarma Stream water, Duzce, Turkey. Water Resources, 44, 315-330.
15. Achio, S., Kutsanedzie, F., & Ameko, E. (2016). Comparative analysis on the effectiveness of various filtration methods on the potability of water. Water Quality Research Journal of Canada, 51(1), 42-46.
16. Elizabeth, K. M., & Rajpramikh, K. E. (2000). Potability of Water among the Tribals of Vizianagaram Sub-plan Area, Andhra Pradesh: Microbiological and Physico-Chemical Analysis. The Anthropologist, 2(3), 181-184.
17. Spackman, C., & Burlingame, G. A. (2018). Sensory politics: The tug-of-war between potability and palatability in municipal water production. Social studies of science, 48(3), 350-371.
18. Mahajan, M., & Bhardwaj, K. (2017). Potability analysis of drinking water in various regions of Ludhiana District, Punjab, India. International Research Journal of Pharmacy, 8(6), 87-90.
19. Lvova, L., Di Natale, C., & Paolesse, R. (2019). Chemical sensors for water potability assessment. Bottled and Packaged Water, 177-208.
20. Abanyie, S. K., Boateng, A., & Ampofo, S. (2016). Investigating the potability of water from dug wells: A case study of the Bolgatanga Township, Ghana. African Journal of Environmental Science and Technology, 10(10), 307-315.
21. Opafola, O. T., Oladepo, K. T., Ajibade, F. O., & David, A. O. (2020). Potability assessment of packaged sachet water sold within a tertiary institution in southwestern Nigeria. Journal of King Saud University-Science, 32(3), 1999-2004.
22. Chauhan, J. S., Badwal, T., & Badola, N. (2020). Assessment of potability of spring water and its health implication in a hilly village of Uttarakhand, India. Applied Water Science, 10, 1-10.
23. Arulnangai, R., Sihabudeen, M. M., Vivekanand, P. A., & Kamaraj, P. (2021). Influence of

physico chemical parameters on potability of ground water in ariyalur area of Tamil Nadu, India. Materials Today: Proceedings, 36, 923-928.

24. An, H., Li, X., Huang, Y., Wang, W., Wu, Y., Liu, L., ... & Jiang, G. (2024). A new ChatGPT-empowered, easy-to-use machine learning paradigm for environmental science. Eco-Environment & Health.
25. BARBERIO, A. (2022). Large language models in data preparation: opportunities and challenges.
26. Hassani, H., & Silva, E. S. (2023). The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field. Big data and cognitive computing, 7(2), 62.
27. Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and Open-AI Models: A Preliminary Review. Future Internet, 15 (6), 192.
28. Mujahid, M., Rustam, F., Shafique, R., Chunduri, V., Villar, M. G., Ballester, J. B., ... & Ashraf, I. (2023). Analyzing sentiments regarding ChatGPT using novel BERT: A machine learning approach. Information, 14(9), 474.
29. Lubiana, T. (2023). Ten Quick Tips for Harnessing the Power of ChatGPT. GPT-4 in Computational Biology. arXiv [q-bio. OT].