Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 46 (2025), No. 4-A, pp. 1045–1057
DOI : 10.47974/JIOS-1890

Semantic web empowered prediction of cervical cancer through machine learning on cloud platform using SemanticAutoMLCloud

Sheetal Agarwal *
Rupal Gupta †
College of Computing Sciences and Information Technology
Teerthanker Mahaveer University
Moradabad 244001
Uttar Pradesh
India

Abstract

Semantic web technologies and machine learning have shown great results in predictive analytics in many areas. Applying machine learning to semantic web data can solve use cases in various domains including the healthcare domain. With rapid growth of data on the web, there is a need for graph database solutions that can be used to store and process semantic web data. This paper proposes an algorithm SemanticAutoMLCloud that further demonstrates the ability with the proposed architecture to integrate semantic web technologies with machine learning using the Amazon SageMaker Autopilot on the AWS cloud platform, which uses Amazon Neptune as a graph database solution to store a large volume of semantic web data. Furthermore, Amazon SageMaker Autopilot training modes, including the Ensembling mode and Hyperparameter Optimization (HPO) mode, and different algorithms of these training modes have been applied to cervical cancer data extracted from the RDF dataset using SPARQL query processing. The XGBoost algorithm of HPO mode outperformed as compared to other algorithms with 99.5% accuracy in the classification of cervical cancer. The purpose of this research is to integrate the machine learning implementation over semantic web data on a cloud platform to solve diverse use cases. This paper proposes a novel architecture that can work on heavy datasets having billions of triples and can be used to process large amounts of data. This paper also illustrates the implementation details of the proposed architecture, which focuses on machine learning.

Subject Classification: *Primary 68P20, Secondary 68T99.*

Keywords: *Cloud computing, Amazon neptune, Machine learning (ML), Semantic web, RDF, SPARQL.*

* *E-mail:* **sheetalagarwal544@gmail.com (Corresponding Author)**
† *E-mail:* **rupal.gupta07@gmail.com**

© Taru Publications

1. Introduction

A vision introduced by Sir Tim Berners Lee for the semantic web aims to expand the existing World Wide Web by enabling machines to understand meaning from information over the Internet [1]. The semantic web is a global database containing a universal network of linked data that allows users to represent data as a rich network [2]. RDF and SPARQL are the fundamental parts of the semantic web stack, which are used to represent and interrogate data to answer complex business questions respectively. SPARQL is a query language which is used for processing RDF data [3]. A knowledge graph (KG) is a network of facts connected via different relationships which is similar to a mind map [4]. SPARQL can be used to process information from knowledge graphs, which further allows machine learning algorithms to retrieve filtered information from complex relationships and patterns stored as knowledge graphs [5]. The implementation of semantic web technologies on cloud platforms in the dynamic landscape of information technology has emerged as a revolutionary change, opening new horizons. As the volume and complexity of data continue to escalate, storing large volumes of semantic web data requires either an extensive data storage infrastructure or the adoption of a scalable cloud platform [6]. The integration of semantic web technologies and machine learning on cloud platforms highlights the challenges and opportunities in the field of semantic web.

This research contributes to a novel strategy that is based upon the proposed algorithm SemanticAutoMLCloud which further proposed an architecture to apply machine learning to semantic web data on a cloud platform. For the same, the information on cervical cancer has been extracted using SPARQL query processing from RDF Data stored on the Amazon Neptune [10] database. It has further been illustrated for predicting cervical cancer by leveraging integration between semantic web technologies and machine learning. The exploration of this integrated approach showcases its potential to enhance predictive accuracy and contributes to the broader understanding of the integration of semantic web technologies, cloud computing, and machine learning in the context of the healthcare domain.

2. Methodology and Proposed Solution

In this paper, a high level novel architecture is represented using a block diagram shown in Figure 1 below, which gives an implemented

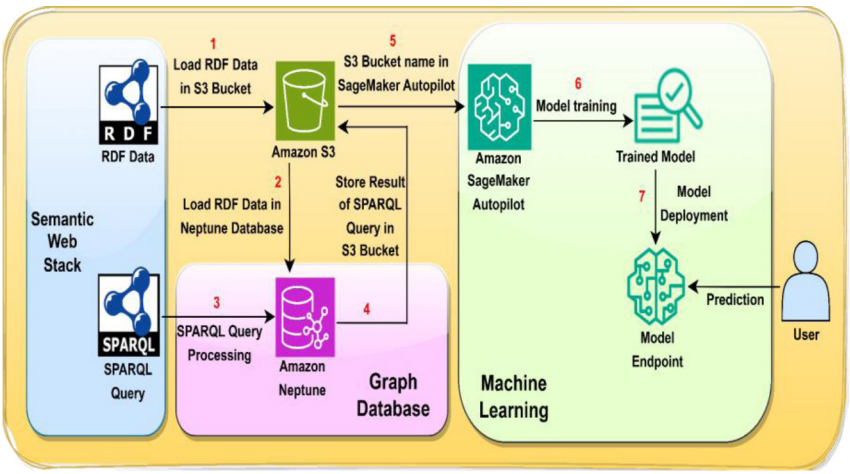


Figure 1

Proposed Architecture of Semantic Web technologies integration with Machine Learning on cloud platform

aspect while utilizing the power of cloud computing and automation using AWS cloud platform. In the proposed architecture, the RDF data (N-Triples format) is loaded on Amazon S3, which is subsequently loaded to Amazon Neptune with the help of Neptune bulk loader and Neptune workbench magic commands in the AWS-Neptune-Notebook environment. With the help of SPARQL query preprocessing, important and useful information is extracted from the RDF dataset. Utilizing the AutoML feature of Amazon SageMaker (SageMaker Autopilot), machine learning algorithms are applied to the results of the SPARQL query. Once the model is trained, the best model provided by Amazon SageMaker Autopilot is deployed resulting in the model endpoint being used for prediction.

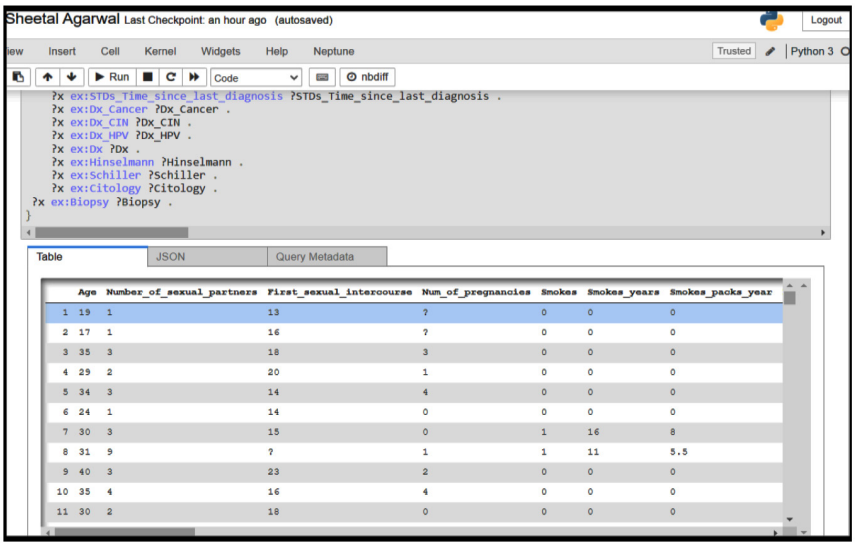
2.1 Semantic Web Stack

The extension of the World Wide Web, which is a web of data, aims to enable machines to interpret information found on the Web. Semantic web connects data points with meaning which allows machines to perform reasoning on the data. It is the foundation for building large-scale knowledge graphs that represent information in a structured and interconnected way and support the integration of diverse RDF datasets [7]. RDF is a framework that is used to represent data as graphs (networks of linked data). SPARQL allows users to create queries that target particular patterns inside RDF graphs, and then extract information from linked



Figure 2(a)

SPARQL query executed on RDF data, stored in the Amazon Neptune database, to extract cervical cancer data



	Age	Number_of_sexual_partners	First_sexual_intercourse	Num_of_pregnancies	Smokes	Smokes_years	Smokes_packs_year
1	19	1	13	?	0	0	0
2	17	1	16	?	0	0	0
3	35	3	18	3	0	0	0
4	29	2	20	1	0	0	0
5	34	3	14	4	0	0	0
6	24	1	14	0	0	0	0
7	30	3	15	0	1	16	8
8	31	9	?	1	1	11	5.5
9	40	3	23	2	0	0	0
10	35	4	16	4	0	0	0
11	30	2	18	0	0	0	0

Figure 2(b)

Results of SPARQL query executed on RDF data, stored in the Amazon Neptune database, to extract cervical cancer data

data sources on the web [8]. Figure 2(a) and Figure 2(b) present the SPARQL query on the RDF dataset having cervical cancer data.

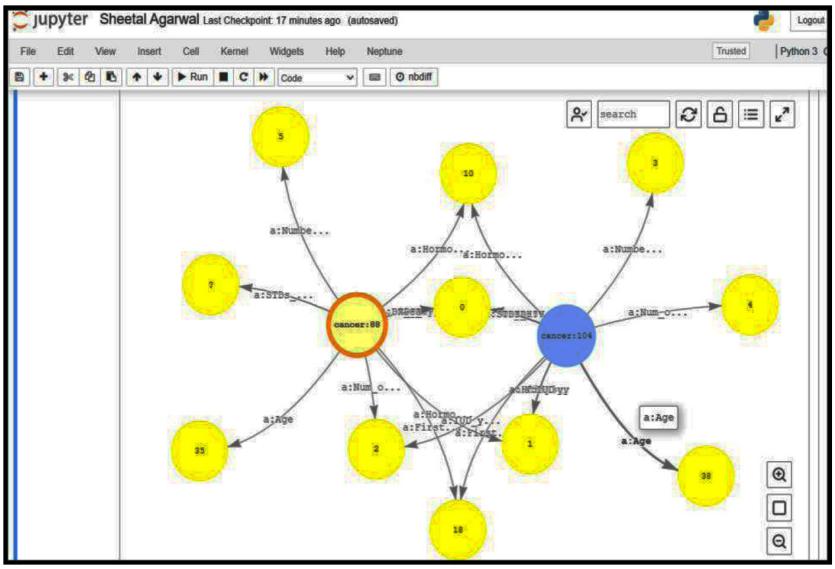


Figure 3

Knowledge Graph of cervical cancer data from RDF data stored in Amazon Neptune database

3. Proposed Algorithm

In this section, an algorithm SemanticAutoMLCloud has been proposed that illustrates the semantic web integration with machine learning over the AWS cloud platform with Amazon SageMaker Autopilot. The algorithm takes the input and produces the best trained model of machine learning using the AWS utilities. The proposed algorithm has three inputs called RDF dataset ($R_{RDF_dataset}$), Neptune database ($N_{Neptune_database}$), and S3 Bucket (S_{S3_Bucket}) respectively. The output produced by the algorithm is the best model (M_{best_model}) that can be used for prediction for any unseen instance of data from the knowledge graph. Figure 3, visualizes the knowledge graph from cervical cancer RDF data stored in the Amazon Neptune database.This algorithm will assist while integrating machine learning over the semantic web using the AWS cloud platform.

Algorithm 1: Semantic web technologies integration with machine learning on AWS cloud platform

Algorithm 1: SemanticAutoMLCloud

Input: $R_{RDF_dataset}$: RDF dataset (cervical cancer in N-Triples format), $N_{Neptune_database}$: Amazon Neptune Database, S_{S3_Bucket}: Amazon S3 Bucket
Output: $M_{best_model} \in$ Model (Trained Model)

begin

S_{S3_Bucket} <- $R_{RDF_dataset}$
$N_{Neptune_Notebook}$ <- access($N_{Neptune_database}$)
$N_{Neptune_database}$ <- S_{S3_Bucket}($R_{RDF_dataset}$)
$E_{Extracted_data}$ <- SPARQL_Query($R_{RDF_dataset}$)
S_{S3_Bucket} <- $E_{Extracted_data}$
SageMaker_Autopilot_Experiment <- S3($E_{Extracted_data}$)
    if training_mode = HPO ***then***
        algorithm <- select_algorithm()
    else if training_mode = ensembling ***then***
        algorithm <- select_algorithm()
    else
        algorithm <- auto_select_algorithm()
experiment <- create_AutoML_experiment($E_{Extracted_data}$, algorithm)
M <- best_model(experiment)
return M_{best_model}
prediction <- perform_prediction(M_{best_model})

end

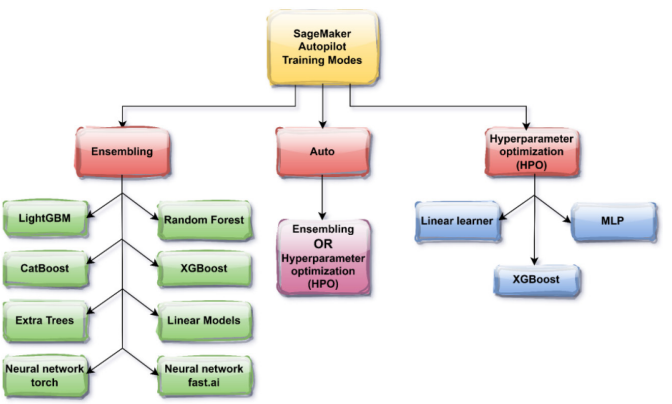


Figure 4

Amazon SageMaker Autopilot training modes and algorithms

Amazon Sagemaker Autopilot

Automation is the key to everything in cloud computing [9] and Amazon SageMaker Autopilot (AutoML) is a completely automated machine learning service managed by Amazon Web Services [10]. This research explores the efficiency and power of SageMaker Autopilot in semantic web technologies, emphasizing its role in automating ML. Amazon SageMaker Autopilot provides different training modes including Ensembling, Hyperparameter Optimization (HPO), and Auto mode. In Figure 4 different training modes and algorithms supported by those training modes are presented.

3.1 *Execution Workflow for the Proposed Algorithm*

The execution workflow for the proposed algorithm has been demonstrated which is designed to facilitate the storage for an extensive amount of semantic web data (RDF form), in a graph database (Amazon Neptune) on a cloud platform (Amazon Web Services) [10]. The workflow empowers users to effortlessly extract meaningful insights from extensive semantic web data through SPARQL query processing. The proposed workflow seamlessly integrates with the AutoML Service (SageMaker Autopilot) provided by Amazon Web Services, allowing users to save time and effort and harness the power of machine learning efficiently. This research not only contributes to the efficient management of semantic web data but also opens the way for users to unlock the transformative potential of automated machine learning in addressing complex challenges. For the same, Figure 5 represents the workflow of the methodology proposed in this research.

This workflow and the processing steps will ease the understanding of the researchers to apply machine learning on RDF data towards the integration of semantic web and machine learning technologies. For the same, the above steps of the workflow have been performed on a sample cervical cancer RDF dataset.

Amazon SageMaker Autopilot simplifies the machine learning process with complete visibility and transparency. Once the model is ready, sagemaker autopilot provides the performance report of the model. Users can check the hyperparameters for the model training, maximum runtime (seconds), and all other details. The performance report includes confusion matrix, precision recall curve, model quality report, AUC, ROC curve, and so on. Once the best model is created, it creates a candidate definition notebook and data exploration report as well and saves all the

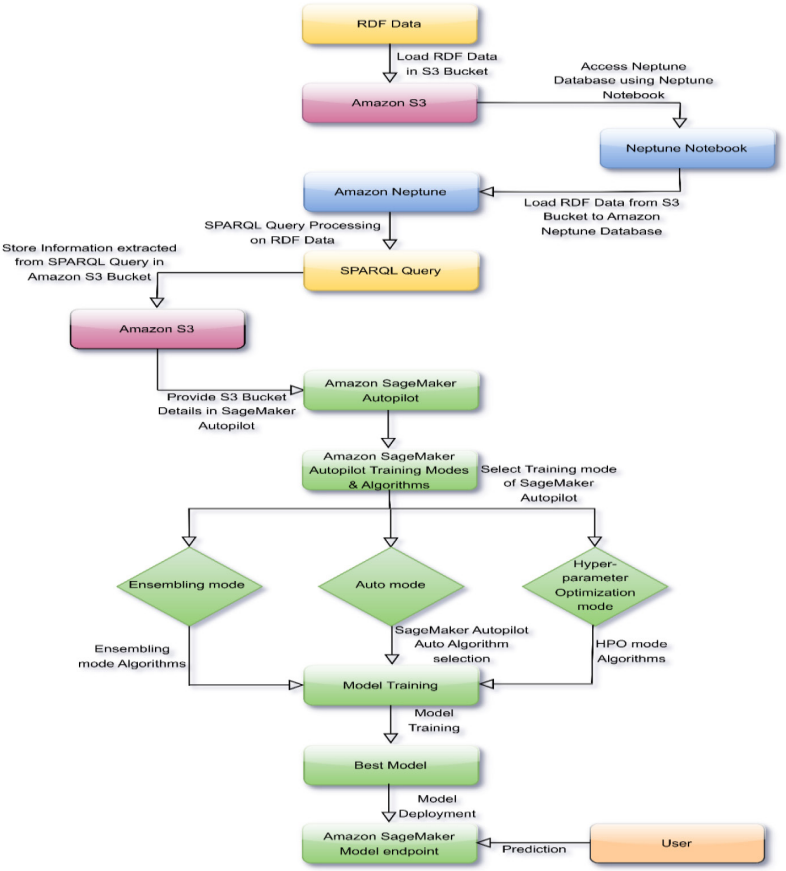


Figure 5

Workflow of machine learning applied to semantic web data on the AWS

splitted data and reports in the S3 bucket provided in the output data destination path.

Figure 6 represents the cross column correlation of cervical cancer information between the top 10 most predictive features generated by Amazon SageMaker Autopilot data exploration notebook for the XGBoost algorithm of Hyperparameter Optimization mode. Amazon SageMaker Autopilot generates the cross column correlation, histogram of target variable for analysis, features summary along with the plot diagram of prediction power of top 10 features and the detailed summary of data and features in the data exploration notebook.

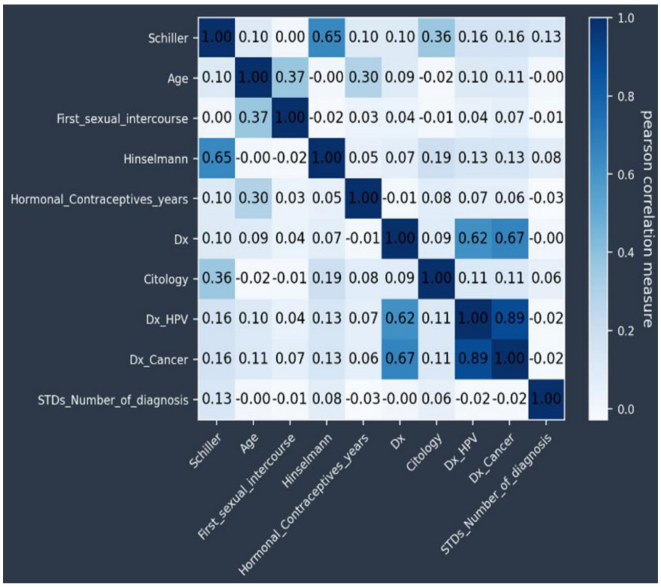


Figure 6

Snapshot of the cross-column correlation of cervical cancer information generated by Amazon SageMaker Autopilot data exploration notebook for XGBoost Algorithm of HPO mode

4. Results And Discussion

This section describes the implementation and results of the applicability of the machine learning algorithms on the cervical cancer dataset. A total of 11 algorithms from the Ensembling mode and HPO mode have been applied to the dataset, and the accuracy measures have been discussed. A high performance graph database service provided by Amazon Web Services called Amazon Neptune has been used in the proposed solution to store an extensive amount of semantic web data,

Table 1

RDF Data Loading Time, SPARQL Query Execution Time, and Response Content Size (bytes) in Amazon Neptune database

RDF Data	Loading time in Amazon Neptune	SPARQL Query Execution Time (ms)	Response Content Size (bytes)
Cervical Cancer	6 seconds	324.08	2717500

presented in Figure 1. Furthermore, Table 1 shows that the RDF data in N-Triples serialization format having cervical cancer patient's information has taken 6 seconds of loading time from Amazon S3 to the Amazon Neptune database accessed through Neptune Notebook.

4.1 *Experimental Setup*

To store RDF data and perform SPARQL query processing, "Neptune Engine Version 1.3.0.0'' with Instance class "db.t3.medium with resource configuration on AWS cloud (2 vCPU, 4GiB RAM, 1500 Mgps EBS)" has been used. The Neptune Notebook with the "ml.t3.medium" instance type has been utilized. Furthermore, to create the Amazon SageMaker domain, Single user setup has been created. All the implementation has been performed in ap-south-1, Asia Pacific (Mumbai) Region in the AWS cloud platform.

4.2 *Dataset description*

The cervical cancer data has been generated based on the UCI repository [11].

4.3 *Evaluation Metrics*

The analysis of the performance for algorithms of Ensembling mode and Hyperparameter Optimization (HPO) mode has been performed with the metrics named accuracy, F1, precision, and recall.

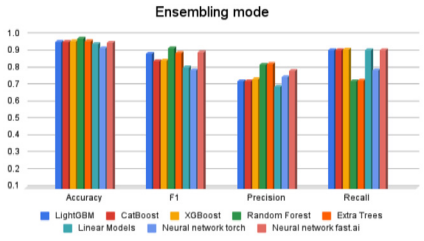


Figure 7

Accuracy measures for Ensembling mode algorithms

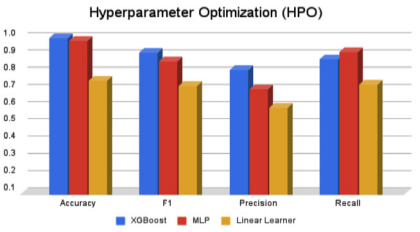


Figure 8

Accuracy measures for HPO mode algorithms

4.4 *Performance analysis*

Performance analysis for Ensembling mode and Hyperparameter Optimization mode algorithms have been analyzed and compared with

the metrics called accuracy, F1, precision, and recall in Figure 7 and Figure 8 respectively.

4.4.1 Analysis with Ensembling Mode Algorithms

The accuracy, F1, precision, and recall for LightGBM, CatBoost, XGBoost, Random Forest, Extra Trees, Linear Models, Neural network torch, and Neural network fast.ai algorithms of SageMaker Ensembling mode are presented in Table 2 respectively.

Table 2

Comparing Amazon SageMaker Autopilot's Ensembling mode algorithms for prediction of cervical cancer RDF data

Ensembling Mode Algorithms				
Algorithms	**Accuracy**	**F1**	**Precision**	**Recall**
LightGBM	0.959	0.891	0.725	0.909
CatBoost	0.961	0.843	0.727	0.912
XGBoost	0.964	0.849	0.741	0.916
Random Forest	0.981	0.921	0.825	0.725
Extra Trees	0.963	0.895	0.829	0.731
Linear Models	0.948	0.81	0.696	0.909
Neural network torch	0.924	0.792	0.751	0.791
Neural network fast.ai	0.953	0.897	0.788	0.909
LightGBM	0.959	0.891	0.725	0.909
CatBoost	0.961	0.843	0.727	0.912

Figure 7 analyzes the performance of LightGBM, CatBoost, XGBoost, Random Forest, Extra Trees, Linear Models, Neural network torch, and Neural network fast.ai algorithms of SageMaker Ensembling mode on cervical cancer patient's information extracted from RDF data.

4.4.2 Analysis with Hyperparameter Optimization (HPO) Mode

The accuracy, F1, precision, and recall for XGBoost, MLP, and Linear Learner algorithm for Amazon SageMaker Hyperparameter Optimization mode are presented in Table 3 respectively.

Table 3

Comparison of Hyperparameter Optimization mode algorithms of Amazon SageMaker Autopilot used on cervical cancer RDF data

Hyperparameter Optimization (HPO) Mode				
Algorithms	**Accuracy**	**F1**	**Precision**	**Recall**
XGBoost	0.995	0.911	0.812	0.874
MLP	0.981	0.859	0.698	0.916
Linear Learner	0.75	0.718	0.591	0.727

Figure 8 analyzes the performance of XGBoost, MLP, and Linear Learner algorithm for Amazon SageMaker Hyperparameter Optimization mode on cervical cancer patient's information extracted from RDF data. The analysis summarizes that the XGBoost algorithm for Amazon SageMaker Hyperparameter Optimization mode outperforms with 99.5% accuracy.

The findings presented in Table 2 and Table 3 from the Ensembling mode and Hyperparameter Optimization mode algorithms applied to cervical cancer information show that the XGBoost algorithm from Amazon SageMaker Hyperparameter Optimization mode provides better accuracy. This research endeavors that the Hyperparameter Optimization mode gives great results as compared to Ensembling modes and the XGBoost algorithm of Hyperparameter Optimization mode outperforms in cervical cancer prediction with 99.5% accuracy.

5. Conclusion

The research focuses on the integration of semantic web and machine learning on a cloud platform. For the same, an algorithm SemanticAutoMLCloud has been proposed with a novel architecture that focuses on the semantic web technologies on the cloud platform for predictions. Different algorithms from Ensembling mode and HPO mode have been applied. Figure 7 and Figure 8 represent the algorithms' performance on the data related to cervical cancer on the basis of accuracy, F1, precision, and recall respectively. The research concluded with the observation that the XGBoost algorithm of HPO mode excels and provides higher accuracy as compared to other algorithms presented in Table 2 and Table 3. This research directs the researcher to work with multi cloud as a solution can be explored in which Amazon Neptune as a graph database solution and GCP Vertex AI OR Azure machine learning as an AutoML solution can be utilized.

References

[1] A. Patel and S. Jain, "Present and future of semantic web technologies: a research statement," *International Journal of Computers and Applications*, vol. 43, no. 5, pp. 413–422 (2021).

[2] M. Wylot, M. Hauswirth, P. Cudré-Mauroux, and S. Sakr, "RDF data storage and query processing schemes: A survey," *ACM Computing Surveys* (CSUR), vol. 51, no. 4, pp. 1–36 (2018).

[3] J. Pérez, M. Arenas, and C. Gutierrez, "Semantics and complexity of SPARQL," *ACM Transactions on Database Systems* (TODS), vol. 34, no. 3, pp. 1–45 (2009).

[4] V. V. Nigam, S. Paul, A. P. Agrawal, and R. Bansal, "A review paper on the application of knowledge graph on various service providing platforms," in Proc. 2020 10th Int. Conf. Cloud Comput., Data Sci. & Eng. (Confluence), Noida, India, pp. 716–720, Jan. (2020).

[5] R. Sapna, H. G. Monikarani, and S. Mishra, "Linked data through the lens of machine learning: an Enterprise view," in Proc. 2019 IEEE Int. Conf. Electr., Comput. and Commun. Technol. (ICECCT), Coimbatore, India, pp. 1–6, Feb. (2019).

[6] H. Brabra, A. Mtibaa, L. Sliman, W. Gaaloul, and F. Gargouri, "Semantic web technologies in cloud computing: a systematic literature review," in Proc. 2016 IEEE Int. Conf. Services Comput. (SCC), San Francisco, CA, USA, pp. 744–751, Jun. (2016).

[7] M. Kejriwal, "Knowledge graphs: A practical review of the research landscape," *Information*, vol. 13, no. 4, p. 161 (2022).

[8] C. Peng, F. Xia, M. Naseriparsa, and F. Osborne, "Knowledge graphs: Opportunities and challenges," *Artificial Intelligence Review,* vol. 56, no. 11, pp. 13071–13102 (2023).

[9] A. V. Joshi and A. V. Joshi, "Amazon's machine learning toolkit: Sagemaker," in *Machine Learning and Artificial Intelligence*, vol. 1, pp. 233–243 (2020).

[10] Amazon Web Services, "Getting Started with AWS." [Online]. Available: https://docs.aws.amazon.com/ [Accessed: Apr. 26, 2025].

[11] UCI Machine Learning Repository, "Cervical Cancer (Risk Factors) Data Set." [Online]. Available: https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors [Accessed: Apr. 26, 2025].

Received October, 2024