CAPSTONE PROJECT

THYROID PREDICTION USING MACHINE LEARNING IN IBM CLOUD

Presented By:
J.POOJA
VIGNAN'S INSTITUTE OF MANAGEMENT AND TECHNOLOGY FOR WOMEN
CSE-AIML

edunet
foundation

OUTLINE

- **Problem Statement** (Should not include solution)
- **Proposed System/Solution**
- **System Development Approach** (Technology Used)
- **Algorithm & Deployment**
- **Result (Output Image)**
- **Conclusion**
- **Future Scope**
- **References**

PROBLEM STATEMENT

The recurrence of thyroid cancer is a significant challenge in patient management, with early detection being crucial for improving treatment outcomes. This problem aims to predict the likelihood of thyroid cancer recurrence by leveraging a combination of demographic, clinical, and diagnostic features. The factors considered for prediction include the patient’s age, gender, smoking history, history of radiation therapy, thyroid function, physical examination findings, and the presence of adenopathy. Additionally, pathological characteristics of the tumor, such as its focality, the TNM classification (tumor size, node involvement, and metastasis), and the cancer stage at diagnosis are key predictors. The response to initial treatment is also a critical factor in assessing recurrence risk. By developing a predictive model based on these variables, the goal is to identify high-risk patients who may experience recurrence, allowing for more personalized care and timely interventions. The model will be evaluated using standard classification metrics such as accuracy, precision, recall, and AUC, aiming to provide a reliable tool that assists clinicians in decision-making and improving patient outcomes.

PROPOSED SOLUTION

1. Data Collection

- **Patient Information**:
  - **Age**: The patient's age, as thyroid conditions can vary with age (e.g., risk of thyroid cancer increases with age).
  - **Gender**: Gender information to assess any gender-based predispositions (e.g., women are at higher risk for thyroid disorders).
- **Lifestyle Factors**:
  - **Smoking**: Whether the patient currently smokes, as smoking can affect thyroid health.
  - **History of Smoking (Hx Smoking)**: Whether the patient has a history of smoking, which may impact thyroid function or the development of thyroid cancer.
  - **History of Radiotherapy (Hx Radiotherapy)**: Whether the patient has been exposed to radiation therapy, which can increase the risk of thyroid cancer.
- **Medical History & Examination**:
  - **Thyroid Function**: Test results for thyroid function, such as TSH, T3, and T4 levels. These can help identify thyroid dysfunction (hypothyroidism, hyperthyroidism).
  - **Physical Examination**: Results from a physical examination of the thyroid (e.g., presence of a goiter or palpable nodules).
  - **Adenopathy**: Lymph node enlargement, which can be indicative of thyroid cancer or other thyroid-related conditions.
  - **Pathology**: Information from any pathological exams (e.g., biopsy results) to determine the nature of any thyroid nodules or lesions.

- **Cancer and Tumor Information**:
  - **Focality**: Whether the thyroid condition involves focal or multi-focal lesions (important in diagnosing cancer).
  - **Risk**: Estimated risk level for thyroid cancer or other thyroid-related conditions.
  - **T**: Tumor size or extent, based on TNM classification (part of staging for thyroid cancer).
  - **N**: Regional lymph node involvement, also part of TNM staging.
  - **M**: Presence of distant metastasis, part of TNM classification.
  - **Stage**: Overall stage of the thyroid cancer, from Stage I to Stage IV.
  - **Response**: Response to treatment (e.g., remission, stable disease, progression), which can help predict long-term outcomes and guide treatment decisions.

2. Data Preprocessing

- **Data Cleaning**:
  - Handle missing values (e.g., missing thyroid hormone levels, history of smoking).
  - Identify and handle outliers (e.g., extremely high/low hormone levels, age extremes).
  - Standardize units for pathology and clinical examination results if necessary (e.g., converting radiotherapy exposure to a consistent format).
- **Feature Engineering**:
  - Convert categorical variables (e.g., Gender, Smoking, Pathology) into numerical values using encoding techniques (one-hot encoding, label encoding).
  - Extract relevant features from clinical history (e.g., whether the patient has a history of radiation or smoking).
  - Normalize continuous features (e.g., TSH levels, age) to ensure consistency and improve model performance.

3. Machine Learning Algorithm

- **Classification or Regression Models**: Depending on the prediction task, we can use different models:
  - **Logistic Regression** or **Random Forest**: For classifying whether a patient has a thyroid disorder or predicting the likelihood of developing thyroid cancer.
  - **Support Vector Machines (SVM)**: For distinguishing between different types of thyroid conditions (e.g., benign vs. malignant thyroid nodules).
  - **Gradient Boosting Machines (GBM, XGBoost)**: Known for their high performance in classification tasks, suitable for predicting thyroid conditions like cancer.
  - **Neural Networks (ANN / LSTM)**: For handling more complex patterns and data, particularly useful if there are longitudinal or time-series data available (e.g., thyroid hormone changes over time).
  - **Survival Analysis Models (Cox Proportional Hazards)**: If predicting patient survival outcomes or response to treatment based on clinical features and treatment regimens.

4. Deployment

- **User Interface**: Develop a system or application where:
  - **Patients**: Can enter their medical history, lifestyle factors, and other relevant details for a prediction or risk assessment of thyroid conditions.
  - **Healthcare Providers**: Can receive insights about thyroid disease risk or the likelihood of cancer, and track patient progress.
  - The application can also provide predictions of thyroid function based on lab results or estimate cancer staging (e.g., if cancer is present, its spread, and prognosis).

5. Evaluation

- **Model Evaluation**:
  - **Accuracy**: Measures the proportion of correct predictions for thyroid condition classifications.
  - **Precision, Recall, F1-Score**: Important when dealing with imbalanced classes (e.g., thyroid cancer detection where most cases may be benign).
  - **ROC Curve and AUC**: Evaluate how well the model distinguishes between different outcomes (e.g., malignant vs. benign thyroid conditions).
  - **Mean Absolute Error (MAE)** or **RMSE**: For regression tasks, particularly if predicting thyroid hormone levels or the likelihood of treatment response.
- **Cross-Validation**: Perform k-fold cross-validation to validate the model's generalizability across different datasets.
- **Continuous Improvement**: Use feedback from clinicians to refine predictions, update the model with new data, and ensure that it remains accurate and relevant over time.

SYSTEM APPROACH

SYSTEM REQUIREMENTS

1. Hardware Requirements:

Processor (CPU):

- Minimum: Intel Core i5 or equivalent (4 cores, 2.5 GHz or higher).

Memory (RAM):

- Minimum: 8 GB.

Storage:

- Minimum: 100 GB of free disk space for data storage, model files, and logs.

2. Software Requirements:

Operating System:

- Windows 10 or later, or macOS, or Linux (Ubuntu or other distributions).

Programming Languages:

- **Python** (preferred for machine learning and data science tasks).
  - Version: 3.6 or later.

Integrated Development Environment (IDE):

- Jupyter Notebook, PyCharm, or Visual Studio Code for developing and testing the model.

LIBRARY REQUIRED TO BUILD THE MODEL

watsonx.ai studio

ALGORITHM & DEPLOYMENT

1. Machine Learning Algorithm

- To predict thyroid conditions (e.g., thyroid cancer, hypothyroidism, or hyperthyroidism) based on the given factors, we need to select the right algorithm that can handle both **classification** (e.g., diagnosis) and **regression** (e.g., predicting thyroid function). Here’s an outline of a suitable approach:

Step 1: Problem Formulation
Step 2: Data Preparation
Step 3: Feature Engineering
Step 4: Model Selection
Step 5: Model Training and Hyperparameter Tuning
Step 6: Model Evaluation Metrics

2. Deployment

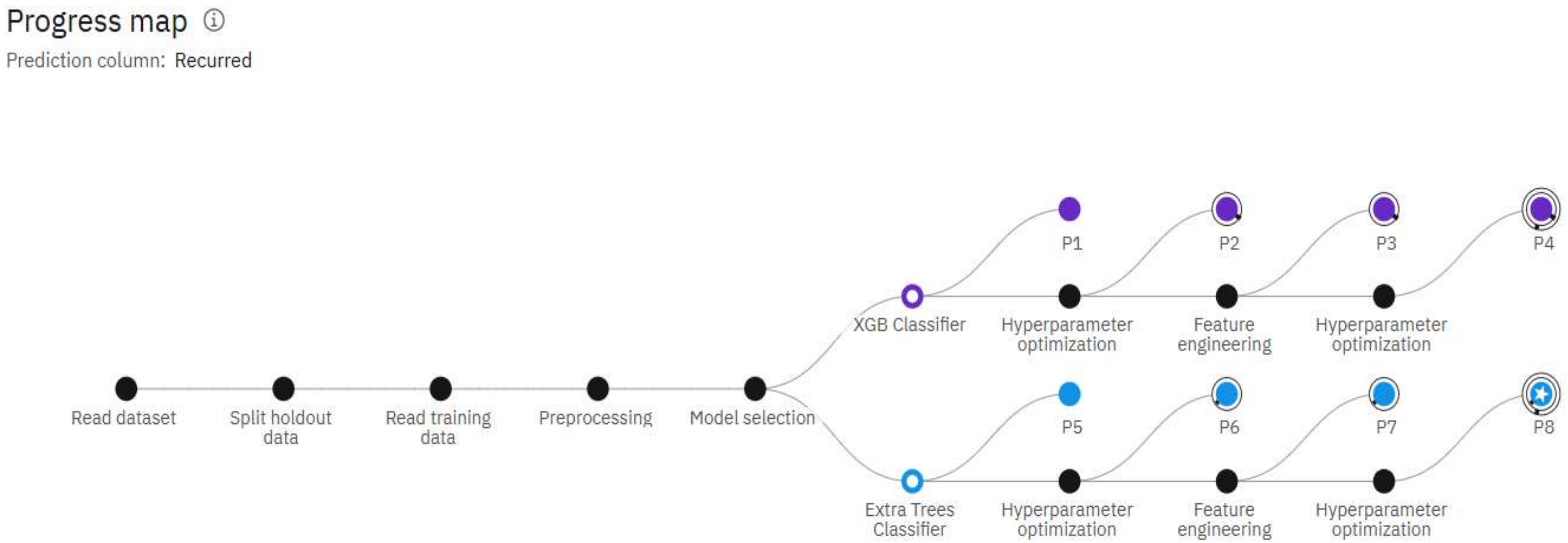
- Deployment of a machine learning model in healthcare requires careful consideration of real-time prediction, scalability, and integration with existing healthcare infrastructure.

- **Deployment Steps:**

  User Interface Development

  Backend System
  Model Serving
  Scalability and Security

RESULT



Progress map
Prediction column: Recurred
P1
P2
P3
P4
XGB Classifier
Hyperparameter optimization
Feature engineering
Hyperparameter optimization
Read dataset
Split holdout data
Read training data
Preprocessing
Model selection
P5
P6
P7
P8
Extra Trees Classifier
Hyperparameter optimization
Feature engineering
Hyperparameter optimization

Prediction results

Display format for prediction results

( ) Table view (•) JSON view

JSON view


```
[
    {
        "fields": [
            "prediction",
            "probability"
        ],
        "values": [
            [
                0,
                [
                    0.9856039877810884,
                    0.014396012218911621
                ]
            ]
        ]
    }
]
```


Prediction results

Display format for prediction results

( ) Table view (•) JSON view

JSON view


```
            "prediction",
            "probability"
        ],
        "values": [
            [
                0,
                [
                    0.9856039877810884,
                    0.014396012218911621
                ]
            ],
            [
                1,
                [
                    0.049352485571632874,
                    0.9506475144283668
                ]
```

CONCLUSION

- In conclusion, this system provides an intelligent, data-driven approach to thyroid health prediction, improving early detection, risk assessment, and personalized care for patients. The combination of powerful machine learning algorithms, a scalable deployment platform, and continuous model refinement positions it as a valuable tool in the modern healthcare ecosystem. This system can contribute significantly to better clinical decision-making, improved patient outcomes, and efficient management of thyroid conditions.

FUTURE SCOPE

- The future scope of the thyroid prediction system is vast, with opportunities for continuous improvement and expansion. As more data sources become available, such as genetic information, lifestyle factors, and longitudinal patient history, the system's predictive capabilities can be significantly enhanced, offering a more comprehensive risk assessment for thyroid disorders. Advanced machine learning models, including deep learning techniques like Convolutional Neural Networks (CNNs) for imaging data and Recurrent Neural Networks (RNNs) for time-series data, can further improve prediction accuracy. Incorporating personalized medicine and real-time monitoring through wearable devices will allow for dynamic, individualized care, while seamless integration with Electronic Health Records (EHR) will streamline data input for healthcare providers, improving workflow efficiency. The system can also be expanded to diverse populations through international collaboration, allowing for more robust, generalized predictions across different regions.

REFERENCES

- https://ieeexplore.ieee.org/document/10183108
- https://pmc.ncbi.nlm.nih.gov/articles/PMC8507551/#:~:text=AI%20represents%20a%20powerful%20tool,methods%20applied%20to%20thyroid%20images.
- https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2022.886953/full
- https://www.sciencedirect.com/science/article/pii/S240584402412587X

edunet foundation

IBM CERTIFICATIONS


In recognition of the commitment to achieve professional excellence

Getting Started with Artificial Intelligence

IBM

Jivilka Pooja

Has successfully satisfied the requirements for:

Getting Started with Artificial Intelligence

Issued on: Jul 15, 2025
Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/9e0bdbdf-5d5f-43a7-bddc-3c624facaa11

IBM

IBM CERTIFICATIONS


In recognition of the commitment to achieve professional excellence

Jivilka Pooja

Has successfully satisfied the requirements for:

Journey to Cloud: Envisioning Your Solution

Issued on: Jul 16, 2025
Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/2ce4ccf0-c070-42c3-b0b3-ed2bae27dd81

IBM CERTIFICATIONS

IBM **SkillsBuild** Completion Certificate

This certificate is presented to

Jivilka Pooja

for the completion of

Lab: Retrieval Augmented Generation with LangChain

(ALM-COURSE_3824998)

According to the Adobe Learning Manager system of record

Completion date: 25 Jul 2025 (GMT) **Learning hours:** 20 mins

THANK YOU