RV COLLEGE OF ENGINEERING®

Bengaluru – 560059

(Autonomous Institution Affiliated to Visvesvaraya Technological University, Belagavi)

Summer Internship Report

21EEI57

on

"HEALTH INSURANCE PRICE PREDICTOR MODEL"

Submitted by

Maaz Ahmed

USN – 1RV21EE035

Carried out at

CoE - AI Research and Business Solutions

Under the Guidance of

Internal Guide
Dr. K. M. Ajay
Asst. Professor
EEE Department
RVCE

CoE Mentor
Prof. Rajesh RM
Asst. Professor
Dept. of AI/ML
RVCE

Submitted in partial fulfillment for the award of degree of
Bachelor of Engineering (B.E.)
in

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

2023-24

RV COLLEGE OF ENGINEERING®

Bengaluru - 560059

(Autonomous Institution Affiliated to Visvesvaraya Technological University, Belagavi)

CERTIFICATE

Certified that the Internship titled "**Heath Insurance Price Prediction Model**" is carried out by **Maaz Ahmed (1RV21EE035)**, a bonafide student of R.V Collegeof Engineering, Bengaluru, submitted in partial fulfillment for the fifth semester examination of Bachelor of Engineering in Electrical and Electronics Engineering affiliated to Visvesvaraya Technological University, Belagavi, during the year 2023-24. It is certified that all corrections/suggestions indicated for the internal assessment have been incorporated in the report and deposited in the departmental library. The Internship report has been approved as it satisfies the academic requirement in respect of internship work prescribed by the institution for the said degree.

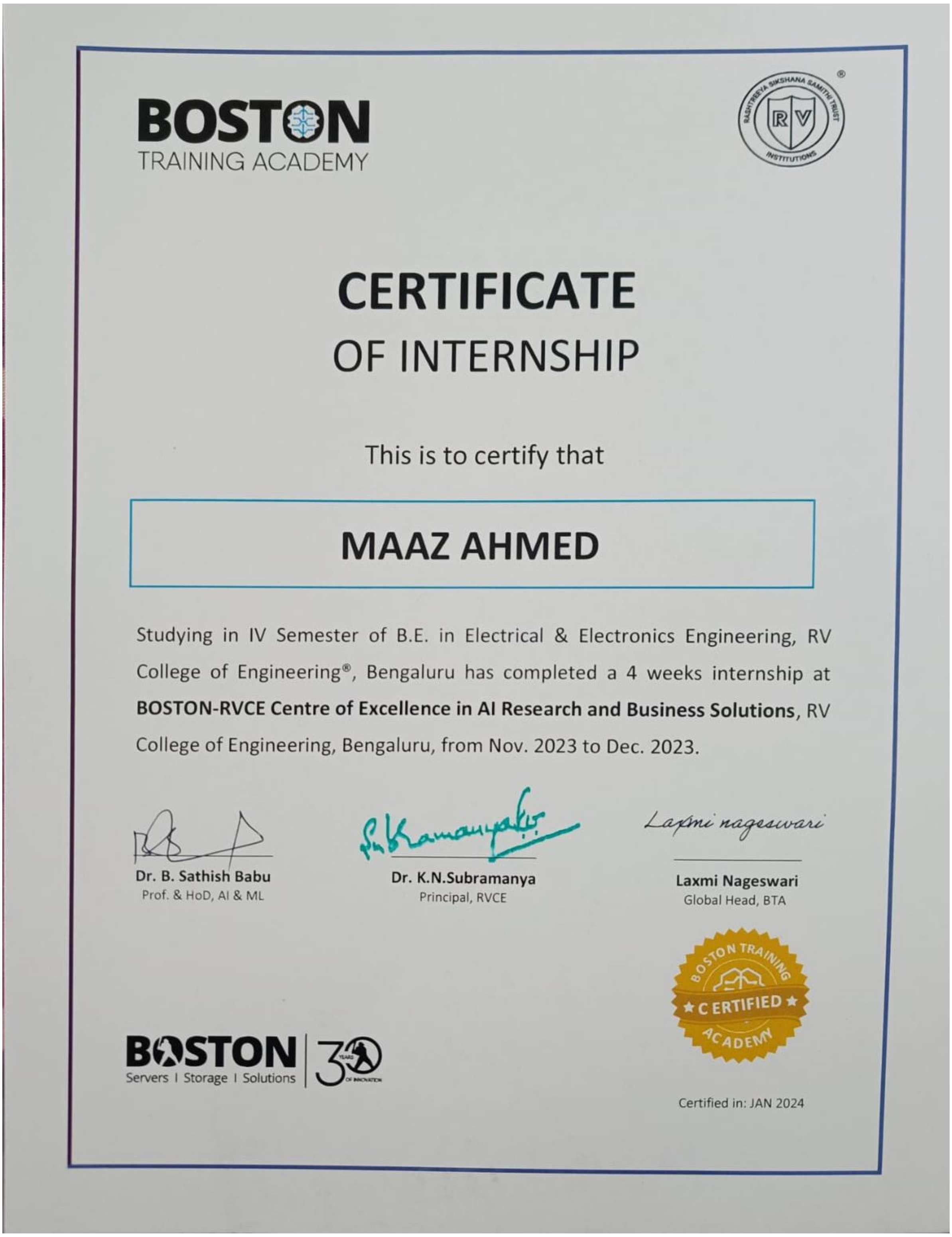
Internal Guide
Dr. K. M. Ajay
Asst. Professor
EEE Department
RVCE

Dr. S. G. Srivani
Head of Department
Department of EEE
RVCE, Bengaluru–59

Dr. K. N. Subramanya
Principal
RVCE
Bengaluru–59

Name of the Examiners	**Signature with Date**
1.___________________	___________________
2.___________________	___________________

CERTIFICATE FROM COE



BOSTON
TRAINING ACADEMY

RASHTREEYA SIKSHANA SAMITHI TRUST
RV
INSTITUTIONS

CERTIFICATE
OF INTERNSHIP

This is to certify that

MAAZ AHMED

Studying in IV Semester of B.E. in Electrical & Electronics Engineering, RV College of Engineering®, Bengaluru has completed a 4 weeks internship at **BOSTON-RVCE Centre of Excellence in AI Research and Business Solutions**, RV College of Engineering, Bengaluru, from Nov. 2023 to Dec. 2023.

Dr. B. Sathish Babu
Prof. & HoD, AI & ML

Dr. K.N.Subramanya
Principal, RVCE

Laxmi nageswari

Laxmi Nageswari
Global Head, BTA

BOSTON TRAINING ACADEMY
★ CERTIFIED ★

BOSTON
Servers | Storage | Solutions
30 YEARS OF INNOVATION

Certified in: JAN 2024

ACKNOWLEDGEMENT

I am indebted to my guide**, Dr. S. G. Srivani, Head of Department**, Department of Electrical and Electronics Engineering and **Prof. Rajesh R M, Assistant Professor, CoE of AI RESEARCH AND BUSINESS SOLUTIONS,** for their wholehearted support, suggestions and invaluable advice throughout the Internship and also helped in the preparation of this report.

I sincerely express my gratitude to my examiner **Dr. K.M. Ajay**, **Assistant Professor,** Department of Electrical and Electronics Engineering, for their valuable comments and suggestions.

I also express our gratitude to our coordinator **Dr. Hemalatha J N., Associate Professor**, Department of Electrical and Electronics Engineering, for her constant support throughout.

We express sincere gratitude to our beloved **Principal, Dr. K. N. Subramanya** for his appreciation towards this internship work.

We thank all the **teaching staff and technical staff** of the Electrical and Electronics Engineering department, RVCE for their help.

Lastly, I take this opportunity to thank our **family** members and **friends** who provided all the backup support throughout the internship work.

MAAZ AHMED
1RV21EE035
Department of Electrical and Electronics Engineering
RV College of Engineering,
Bengaluru-59

EXECUTIVE SUMMARY

The CoE of AI Research and Business Solutions, a joint initiative between RV College of Engineering and Boston Ltd. UK, was inaugurated on 25th August 2022. It aims to provide advanced solutions using AI, machine learning, and deep learning, as well as infrastructure for startups and technology-enabled training. The CoE operates in five verticals: AI Infrastructure, Training and Education, Incubation and Bootstrapping, Industrial Consultancy, and Research and Development. The CoE of AI Research and Business Solutions at RV College provides a range of products and services in artificial intelligence, machine learning, and deep learning. It offers access to a Graph core IPU M2000 machine, curated training programs, and supports incubation and bootstrapping of ideas. The CoE also offers industrial consultancy in various industries and conducts research in areas like natural language processing, computer vision, speech processing, and reinforcement learning. It also conducts workshops, certification courses, internships, and projects in collaboration with industry partners and academic institutions.

The Machine Learning Domain, a key component of the Center of Excellence - AI RESEARCH AND BUSINESS SOLUTIONS, focuses on advancing the field of machine learning through cutting-edge research and development initiatives. The mission is to develop innovative solutions, foster collaboration with industry partners, and contribute to the academic community through publications and knowledge dissemination. The domain provides a structured environment for individuals to apply theoretical knowledge to practical scenarios, develop essential skills, and gain valuable insights into machine learning applications in various industries. Internships in the machine learning domain contribute to professional growth and preparation for careers in data science, artificial intelligence, and related fields. Interns engage in project work, data analysis, algorithm development, coding and programming, collaboration, learning and skill development, problem-solving, and documentation and reporting. The need for Data Science and AI/ML solutions in precision agriculture is driven by the desire to make agriculture more efficient, sustainable, and resilient in the face of environmental and economic challenges. These technologies empower farmers with actionable insights, promoting precision in decision-making and fostering a more productive and environmentally friendly approach to agriculture.

The learning outcomes include a deep understanding of machine learning and deep learning, proficiency in Python programming, and problem-solving skills, applying data analysis techniques, proactively identifying and resolving challenges, and demonstrating effective communication. Participation in project planning, execution, and time management, and demonstrating adaptability and commitment to continuous learning is achieved.

Table of Contents

LIST OF FIGURES

CHAPTER 1

PROFILE OF THE ORGANISATION

CHAPTER 1

PROFILE OF THE ORGANISATION

1.1 Organizational Structure

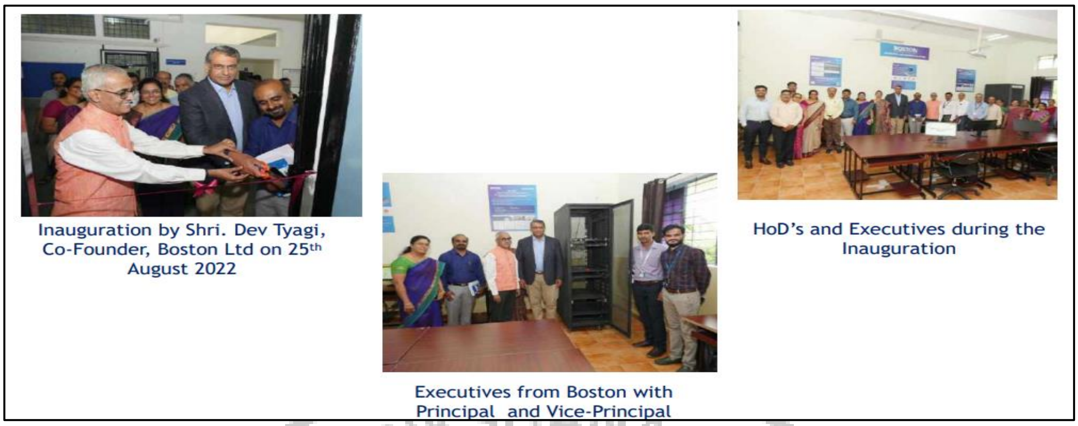
The CoE of AI Research and Business Solutions is a joint initiative by RV College of Engineering and Boston Ltd. UK. The CoE was inaugurated on 25th August 2022, by Sri. Dev Tyagi, Co-founder, Boston Ltd., UK, Mrs. Laxmi Nageswari, Global Head of AI Education, Boston Ltd. Bengaluru, Dr. K.N. Subramanya Principal, RVCE, Dr K.S Geetha, Vice Principal, RVCE, Dr B Satish Babu, HoD, AIML seen in fig 1.1.1.

The CoE aims to provide cutting-edge solutions for various domains using artificial intelligence, machine learning, and deep learning. The CoE also provides the necessary infrastructure for startups & technology enabled training to encourage and support start-up ecosystems.

The CoE is headed by Dr. B. Sathish Babu, Professor and HoD, Dept. of AIML, RVCE. He is assisted by a team of faculty members and students from the Department of Artificial Intelligence and Machine Learning, RVCE. The CoE also collaborates with experts and stakeholders from Boston Ltd. UK, and other industry and academic partners.

It operates in five verticals:

- AI Infrastructure,
- Training and Education,
- Incubation and Bootstrapping,
- Industrial Consultancy, and
- Research and Development.



Inauguration by Shri. Dev Tyagi, Co-Founder, Boston Ltd on 25th August 2022

Executives from Boston with Principal and Vice-Principal

HoD's and Executives during the Inauguration

Fig 1.1.1: Inauguration of CoE

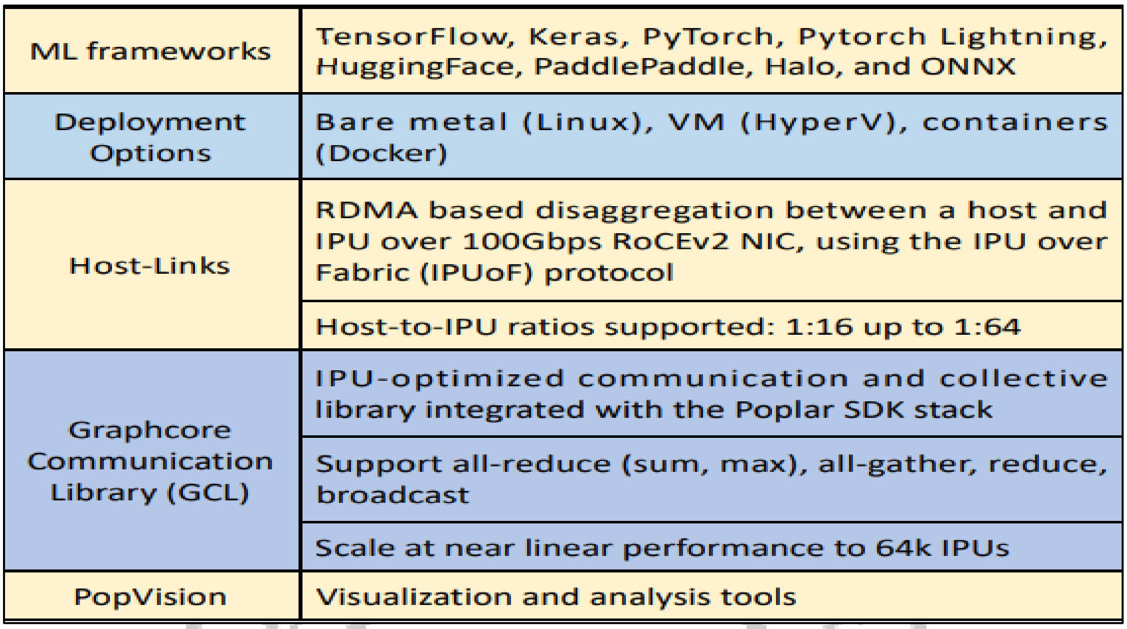
1.2 Products and Services

The CoE of AI Research and Business Solutions at RV College offers various products and services in the field of artificial intelligence, machine learning, and deep learning.

Some of them are shown in below figure i.e., fig 1.2.1

Fig 1.2.1: Five Verticals of Operation

AI Infrastructure: The CoE provides access to a Graph core IPU M2000 machine with POD4 capacity, which can handle high-end AI workloads without dependency on the cloud infrastructure. It also provides various deployment options, communication libraries, and visualization and analysis tools for AI applications shown in below fig.1.2.2.



ML frameworks	TensorFlow, Keras, PyTorch, Pytorch Lightning, HuggingFace, PaddlePaddle, Halo, and ONNX
Deployment Options	Bare metal (Linux), VM (HyperV), containers (Docker)
Host-Links	RDMA based disaggregation between a host and IPU over 100Gbps RoCEv2 NIC, using the IPU over Fabric (IPUoF) protocol
	Host-to-IPU ratios supported: 1:16 up to 1:64
Graphcore Communication Library (GCL)	IPU-optimized communication and collective library integrated with the Poplar SDK stack
	Support all-reduce (sum, max), all-gather, reduce, broadcast
	Scale at near linear performance to 64k IPUs
PopVision	Visualization and analysis tools

Fig 1.2.2.: AI in Infrastructure

Training and Education: The CoE offers curated training programs on the latest and cutting-edge technologies, such as Intel One API, for industry participants and students. It also conducts workshops, certification courses, internships, and projects in collaboration with industry partners and academic institutions.

Incubation and Bootstrapping: The CoE supports incubating the ideas under industrial mentorship and gives bootstrapping services to launch the ideas as workable products and business services. It also works on inter-disciplinary/multi-disciplinary problems in various domains, such as FinTech, AgriTech, Healthcare, Genomics, Smart cities, and others.

Industrial Consultancy: The CoE invites industrial consultancy inquiries in verticals such as commerce, science, healthcare, smart cities, agriculture, and others, where data science and AI technologies are needed. It also provides solutions for various AI-related challenges and opportunities faced by the industry.

Research and Development: The CoE invites research scholars to explore working of their research algorithms and solutions on high-end infrastructure with minimum cost.

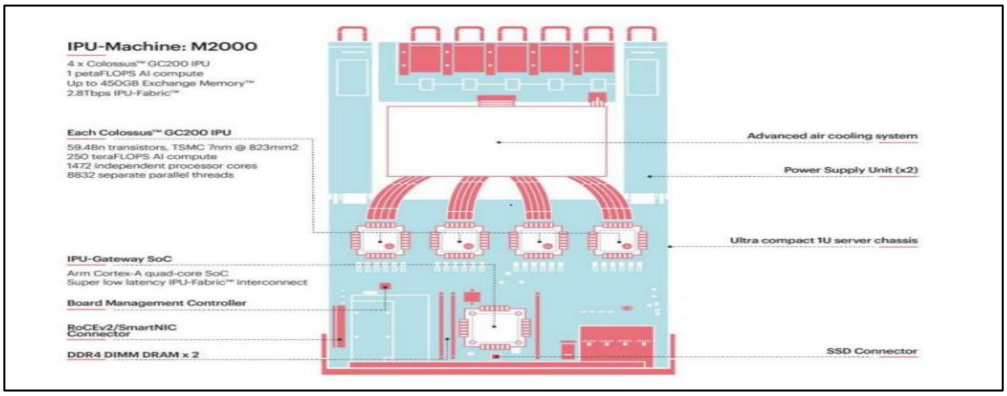

IPU-Machine: M2000
4 x Colossus™ GC200 IPU
1 petaFLOPS AI compute
Up to 450GB Exchange Memory™
2.8Tbps IPU-Fabric™
Each Colossus™ GC200 IPU
59.4Bn transistors, TSMC 7nm @ 823mm2
250 teraFLOPS AI compute
1472 independent processor cores
8832 separate parallel threads
IPU-Gateway SoC
Arm Cortex-A quad-core SoC
Super low latency IPU-Fabric™ interconnect
Board Management Controller
RoCEv2/SmartNIC Connector
DDR4 DIMM DRAM x 2
Advanced air cooling system
Power Supply Unit (x2)
Ultra compact 1U server chassis
SSD Connector

Fig 1.2.3: IPU Machine M2000

1.3 Industrial Interaction

The Collaboration with Companies shown in fig 1.3.1, fig 1.3.2 and fig 1.3.3

1. Company – **BullWork Mobility**

- Duration of the Project - 20 weeks
- Title of the work - DL model for Plant Health Monitoring
- Funding: Rs.3,00,000/-

Faculty (AIML Dept.):

- Dr. B. Sathish Babu
- Dr. Vijayalakshmi M N
- Prof. Narasimha Swamy
- Students (AIML Dept.): SubhashGupta and Om Mangalg

Fig 1.3.1: Collaboration with BullWork

2. Company – **Toyota Industries Engine India Pvt. Ltd.**

- Duration of the Project - 12 Months (Completed)
- Title of the work - Implementation of Machine Learning Solution for GD Engines drilling operation
- Funding: Rs.1,56,000/-

Faculty: Dr. B. Sathish Babu (PI-AIML)

- Dr. Sridhar (Mech.)
- Dr. Gangadhar (Mech.)

Students:

- Ajay Brightson (III Sem. AIML) and Hrithik (V Sem. CSE)

Fig 1.3.2: Collaboration with Toyota Industries

3. Company – **Saint-Gobain Gyproc**

- Duration of the Project - 02 Months
- Title of the work - Implementation of Machine Learning Solution for predicting moisture correction factor
- Funding: Joint research project

Faculty: Dr. B. Sathish Babu

Students:

- Arun Kumar (Saint-Gobain Gyproc)
- Ajay Brightson (III Sem. AIML)
- Swarna A N (III Sem).

Fig 1.3.3: Collaboration with Gyproc

1.3.1 Activity and Research Collaboration

- Successfully launched the first batch of the certification course in Data Science on 20th August,2022 shown in fig 1.3.3.1

  Number of Participants: 10 (05 Industry + 05 Academics).

- “Train-the-Trainer” workshop on Intel Unnati Gaudi DL Lab

Fig 1.3.1.1: Activities under CoE

1.4 Business Partners

The business partners of the COE of AI Research and Business Solutions at RV College are:

RV College of Engineering: RVCE is a private engineering college in Bangalore, India, that established the CoE in 2022 in collaboration with Boston Ltd. UK. RVCE provides the faculty, students, and infrastructure for the CoE's operations.

Boston Ltd. UK: Boston Ltd. is a global provider of high-performance, mission-critical server, storage, and cloud solutions. Boston Ltd. Sponsored the Graphcore IPU M2000 machine with POD4 capacity for the CoE, and also provides industrial mentorship and training.

Other industry and academic partners: The CoE collaborates with various industry and academic partners for conducting workshops, certification courses, internships, projects, consultancy, and research. Some of the partners are Intel, BullWork Mobility, M R Ambedkar Dental College, RR Institute of Technology, etc.

Research grants: The CoE conducts research in various areas of AI, such as natural language processing, computer vision, speech processing, reinforcement learning, and others. The CoE may receive research grants from various funding agencies, such as the Government of India, DST, DRDO, etc.

1.5 Societal Concerns

The CoE faces the challenge of attracting and retaining talented and motivated manpower, as well as providing them with continuous learning and development opportunities.

Ethical, fair, transparent, and accountable AI: The CoE aims to create AI solutions that are ethical, fair, transparent, and accountable, meaning that they respect the values, rights, and norms of the society, and that they can be explained, verified, and controlled by the users and the stakeholders. The CoE faces the challenge of ensuring that the AI solutions do not cause harm, bias, discrimination, or injustice to any individual or group, and that they comply with the relevant laws and regulations.

Privacy and security of data and users: The CoE aims to protect the privacy and security of the data and the users that are involved in the AI solutions, meaning that they respect the confidentiality, integrity, and availability of the data and the users, and that they prevent unauthorized access, use, or disclosure of the data and the users. The CoE faces the challenge of ensuring that the data and the users are not exposed to any risks, threats, or attacks, and that they have the right to consent, access, correct, or delete their data and their participation in the AI solutions.

Awareness and education of the public about AI: The CoE aims to raise awareness and educate the public about the benefits and risks of AI, meaning that they inform, engage, and empower the public to understand, use, and participate in the AI solutions, and that they address the myths, fears, and misconceptions about AI. The CoE faces the challenge of ensuring that the public has the knowledge, skills, and attitudes to make informed and responsible decisions about AI, and that they have the opportunity to voice their opinions, feedback, and concerns about AI.

Trust and collaboration among the stakeholders of AI: The CoE aims to foster trust and collaboration among the stakeholders of AI, meaning that they build and maintain positive and productive relationships among the users, developers, providers, regulators, and researchers of AI, and that they share the goals, values, and responsibilities of the AI solutions. The CoE faces the challenge of ensuring that the stakeholders have the mutual respect, understanding, and communication to work together effectively and efficiently, and that they have the mechanisms to resolve any conflicts, disputes, or issues that may arise from the AI solutions.

1.6 Manpower

Fig 1.6.1: Faculty at the CoE

- Dr. B Sathish Babu, Professor and HoD, Department of Artificial Intelligence and Machine Learning, RVCE
- Dr. Vijayalakshmi M.N, Associate Professor, Department of Artificial Intelligence and Machine Learning, RVCE
- Dr. S. Anupama Kumar, Associate Professor, Department of Artificial Intelligence and Machine Learning, RVCE
- Prof. Narasimha Swamy S, Assistant Professor, Department of Artificial Intelligence and Machine Learning, RVCE
- Prof. Somesh Nandi, Assistant Professor, Department of Artificial Intelligence and Machine Learning, RVCE
- Prof. Rajesh R M, Assistant Professor, Department of Artificial Intelligence and Machine Learning, RVCE
- Prof. Priya TV, Assistant Professor, Department of Artificial Intelligence and Machine Learning, RVCE
- Prof. Sonika CT, Assistant Professor, Department of Artificial Intelligence and Machine Learning, RVCE

CHAPTER 2

ACTIVITIES OF DEPARTMENT

CHAPTER 2
ACTIVITIES OF THE DEPARTMENT

2.1 About the domain

It was a privilege to work within the domain of Machine Learning under the Center of Excellence - AI RESEARCH AND BUSINESS SOLUTIONS. This domain is a key component of the CoE, focusing on advancing the field of machine learning through cutting-edge research and development initiatives.

Mission and Objectives: The primary mission of the Machine Learning domain is to conduct impactful research in the field of artificial intelligence and machine learning. This includes developing innovative solutions, fostering collaboration with industry partners, and contributing to the academic community through publications and knowledge dissemination.

Machine learning as a domain provides a structured environment for individuals to apply theoretical knowledge to practical scenarios, develop essential skills, and gain valuable insights into the applications of machine learning in various industries. Internships in this domain contribute to the professional growth and preparation of individuals for careers in data science, artificial intelligence, and related fields.

Interns in the machine learning domain typically engage in activities such as:

1. Project Work: Undertaking specific projects that involve the application of machine learning techniques to solve practical problems. This could include tasks like developing predictive models, image recognition systems, or natural language processing applications.
2. Data Analysis: Interns may be involved in collecting, cleaning, and analyzing data, as data is a crucial component of machine learning. This involves understanding the data, identifying patterns, and preparing it for use in machine learning algorithms.
3. Algorithm Development: Actively participating in the design, development, and optimization of machine learning algorithms.

This includes experimenting with different models, fine-tuning parameters, and evaluating their performance.

4. Coding and Programming: Writing code to implement machine learning algorithms and models. Proficiency in programming languages such as Python, along with familiarity with relevant libraries and frameworks (e.g., TensorFlow, PyTorch), is often a key aspect of machine learning internships.

5. Collaboration: Working collaboratively within a team, which may include data scientists, engineers, and domain experts. Collaboration skills are crucial in understanding project requirements, sharing insights, and collectively working towards project goals.

6. Learning and Skill Development: Actively learning about new developments in the field of machine learning, staying updated on the latest tools and technologies, and continuously improving technical skills.

7. Problem Solving: Addressing real-world challenges by applying machine learning techniques. This involves critical thinking, problem-solving skills, and the ability to adapt machine learning solutions to specific business or research problems.

8. Documentation and Reporting: Documenting the processes, methodologies, and outcomes of machine learning projects. This may involve preparing reports, presenting findings, and effectively communicating technical concepts to non-technical stakeholders.

2.2 About the subdomain-Medical Domain

The health sector plays a pivotal role in societal well-being, encompassing various components such as healthcare providers, facilities, pharmaceuticals, and insurance. Health insurance, a key aspect of the health sector, offers financial protection against the escalating costs of medical care. It functions as a contractual agreement between individuals and insurance companies, covering expenses for services like doctor visits, hospital stays, and medications. The unpredictability of healthcare costs, influenced by regulatory changes and external events, makes predicting health insurance prices challenging. Government policies, market trends, and advancements in medical technology further contribute to pricing dynamics.

The availability and accessibility of quality healthcare services are essential for a robust health sector. Health insurance serves as a critical tool in ensuring that individuals have affordable access to necessary medical treatments. Public health initiatives, preventive care, and technology-driven innovations also shape the evolving landscape of the health sector.

2.2.1 Challenges and Concerns

1. Data Complexity: Health insurance price prediction involves analyzing complex and diverse datasets, including medical histories, demographic information, and healthcare trends. Integrating and interpreting such varied data sources can be challenging.

2. Dynamic Healthcare Landscape: The healthcare landscape is dynamic, with constantly evolving medical treatments, technologies, and healthcare practices. Predicting insurance prices must account for these changes, adding a layer of complexity to the modeling process.

3. Regulatory Changes: Frequent changes in healthcare regulations can impact insurance pricing models. Staying compliant with evolving regulations while maintaining pricing accuracy is a persistent challenge.

4. Uncertain Future Events: Unforeseen events, such as pandemics or major health crises, can significantly impact healthcare costs. Incorporating such unpredictable factors into price prediction models poses a challenge for insurers.

5. Individual Health Variability: The health status of individuals can vary widely. Predicting how different demographics and health conditions will affect insurance costs requires sophisticated modelling techniques to account for this variability.

6. Ethical and Privacy Concerns: Handling sensitive health data raises ethical and privacy concerns. Striking a balance between utilizing data for accurate predictions and safeguarding individuals' privacy is a challenge in health insurance price prediction.

2.2.2 Need for AIML Solutions

The Need for Data Science and AI/ML Solutions in Predicting Health Insurance Prices:

- Data Complexity: The healthcare sector generates vast and diverse datasets, including patient histories, treatment outcomes, and demographic information. Data science techniques enable the extraction of valuable insights from these complex datasets, aiding in the accurate prediction of health insurance prices.

- Predictive Modelling: AI and machine learning algorithms excel at predictive modelling, allowing insurers to analyze historical data to identify patterns and trends. This predictive capability is crucial for estimating future healthcare costs and determining appropriate insurance premiums.

- Risk Assessment: Data science and AI/ML solutions play a pivotal role in assessing and managing risks associated with insuring diverse populations. These technologies can analyze individual health profiles, lifestyle factors, and historical claims data to better predict and quantify potential financial risks for insurance providers.

- Adaptability to Changes: The healthcare landscape is dynamic, with continuous advancements in medical treatments and technologies. AI/ML solutions can adapt to these changes, ensuring that health insurance price predictions remain accurate and reflective of the evolving nature of healthcare practices.

- Enhanced Accuracy: Machine learning models can enhance the accuracy of health insurance pricing by identifying non-linear relationships and intricate patterns within datasets. This allows insurers to make more precise predictions, reducing the likelihood of underestimating or overestimating insurance prices.

- Enhanced Accuracy: Machine learning models can enhance the accuracy of health insurance pricing by identifying non-linear relationships and intricate patterns within datasets. This allows insurers to make more precise predictions, reducing the likelihood of underestimating or overestimating insurance prices.

- Personalized Pricing: AI/ML solutions enable insurers to move towards personalized pricing based on individual risk factors, providing a more tailored approach to health insurance. This not only enhances fairness but also attracts and retains policyholders by offering competitive and customized premium rates.

- Efficiency and Automation: Data science and AI/ML streamline the pricing process, automating repetitive tasks and allowing insurers to handle large and complex datasets efficiently. This not only saves time but also reduces the likelihood of errors in the pricing models.

In summary, the integration of data science and AI/ML solutions in predicting health insurance prices addresses the complexities of the healthcare landscape, enhances accuracy, and enables insurers to make data-driven decisions. These technologies contribute to more effective risk management, personalized pricing, and regulatory compliance, ultimately benefiting both insurance providers and policyholders.

CHAPTER 3

TASKS PERFORMED

CHAPTER 3

TASKS PERFORMED

Machine learning serves as a trans-formative tool, profoundly shaping the decision-making landscape for crop and fertilizer recommendations in agriculture. By harnessing the power of advanced algorithms, this technology empowers farmers to navigate the complexities of agricultural choices with precision and data-driven insights. Through the automated analysis of extensive datasets encompassing soil attributes, climate conditions, and historical crop yields, machine learning models seamlessly classify optimal crops for specific environmental contexts. Moreover, these models extend their functionality to provide nuanced recommendations for fertilizer types and quantities, custom-tailored to the unique nutrient needs of the predicted crops. The integration of machine learning not only streamlines the decision-making process but also cultivates a sustainable and optimized approach to farming, enabling farmers to navigate dynamic agricultural landscape with confidence and efficiency. [7]

3.1 Project Workflow

This project's workflow is shown in the Fig 3.1.1. For any machine learning project , the problem definition and understanding the problem. Then comes the Data-preprocessing stage that helps us understand the importance of features and its affects on the project. Also, data is transformed for better use for the machine learning during this stage of the project. According to the understanding of the project, a machine learning algorithm is selected or it can be selected by trial and error. The pre-processed data is used to train and test the accuracy of machine learning model and if the accuracy of the model is good, the projects is complete.

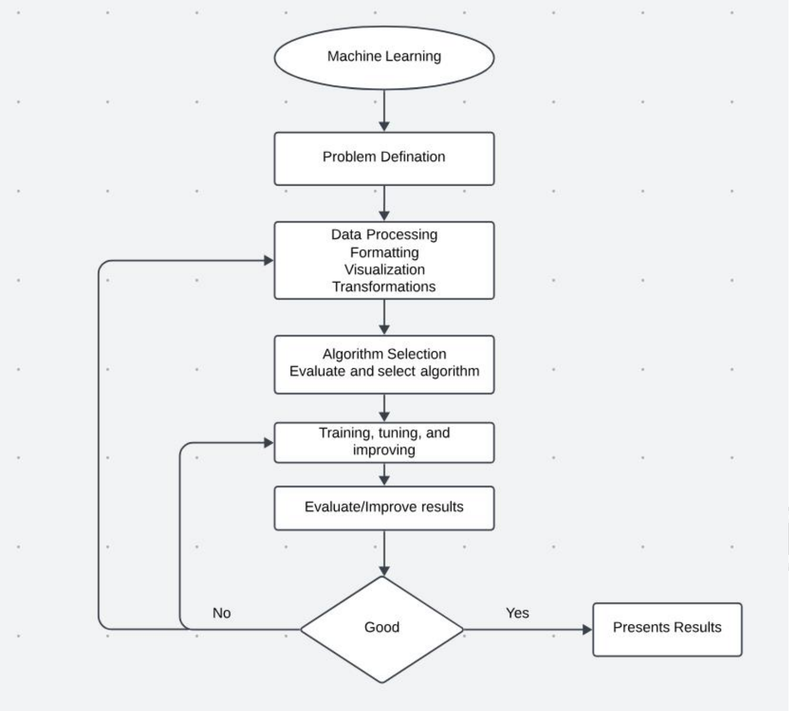


Fig. 3.1.1 Six Stages of Machine Learning

1. Data Collection:
   - Gather a diverse dataset containing age, gender, no of children, region,smoker etc .
2. Data Preprocessing:
   - Clean and preprocess the dataset, handling missing values, outliers, and standardizing numerical features.
3. Exploratory Data Analysis (EDA):
   - Conduct EDA to understand feature distributions, correlations, and visualize data patterns.
4. Algorithm Selection:
   - Based on the nature of the problem (classification and regression) and the dataset characteristics, choose appropriate algorithms.
5. Model Development – Price Prediction:
   - Split the dataset into training and testing sets.
   - Train and evaluate the selected insurance price algorithm.
6. User Interface Development:
   - Create an interactive user interface for user input and displaying

recommendations.

7. Model Evaluation:

- Evaluate the performance of both models using metrics such as accuracy, precision, and recall.

8. Monitoring:

- Monitor system performance, user interactions, and model predictions.

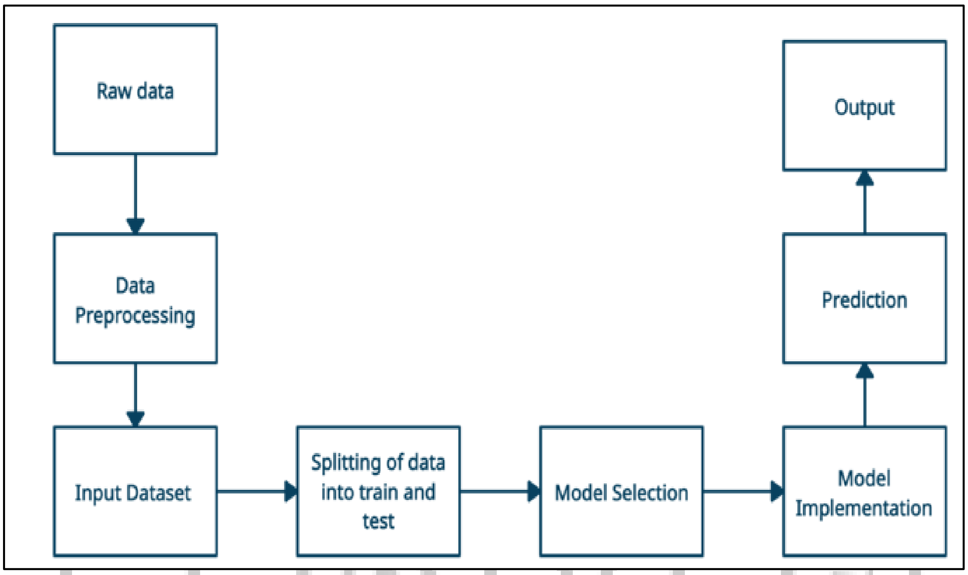


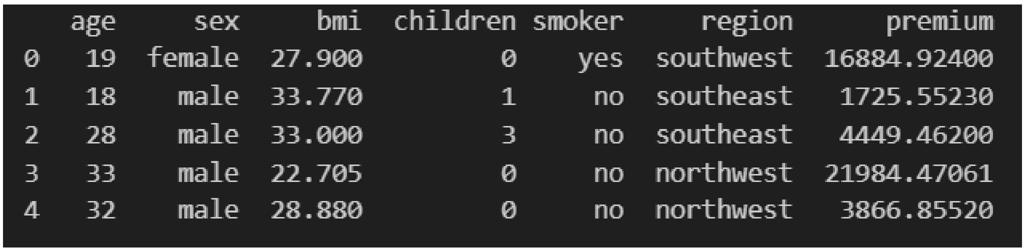
Fig. 3.1.2 Project Workflow

3.2 Dataset Collection

For the purpose of developing this project, the dataset is taken from Kaggle. Kaggle is a well-known online platform that serves as a hub for data science and machine learning competitions, collaboration, and learning. Established in 2010, Kaggle has become a go-to destination for data scientists, machine learning enthusiasts, and professionals looking to showcase their skills and solve real-world problems using data.

Health Insurance Dataset:

The dataset shown in Fig. 4.2.1 shows the first five entries in the dataset. The dataset contains 7 features and 1338 unique entries. It contains 6 inputs and 1 target. Among the 7 features, it contains 4 numerical features : age, BMI, no. of children, and expenses and 3 nominal features : sex, smoker and region. Among these 7 features the "premium" is our target, which is the price paid by the customer per year for the health insurance and the other 6 are the input.



```
   age     sex     bmi  children smoker     region      premium
0   19  female  27.900         0    yes  southwest  16884.92400
1   18    male  33.770         1     no  southeast   1725.55230
2   28    male  33.000         3     no  southeast   4449.46200
3   33    male  22.705         0     no  northwest  21984.47061
4   32    male  28.880         0     no  northwest   3866.85520
```

Fig. 3.2.1 Health Insurance Dataset

About the features of this dataset:

- **Age**: This one of the critical factors that affect the premium amount. Higher the age, higher the premium. Older people are at a higher risk of suffering from illness than youngsters. So, if it is advisable to buy a health insurance premium when you are young; you will give comprehensive coverage and better benefits at an affordable premium.
- **Gender**: The impact of sex on health insurance pricing is often considered by insurance providers due to differences in healthcare utilization between males and females. Factors such as maternity care needs for women have traditionally influenced premium rates.
- **BMI (Body Mass Index)**: Generally, people with higher BMI are charged higher premium than those who have a normal BMI. This is because people with high BMI are at a high risk of suffering from various diseases like diabetes, and heart-related problems, and therefore need regular medical care.
- **Number of Children**: The number of children in a family can impact health insurance prices. Insurance providers consider larger families as they often entail higher healthcare expenses, including pediatric care and vaccinations.
- **Smoking Habits**: Smoking increases health risks, and the insurance companies view smokers as high-risk insurance buyers, and therefore charge high premium. People who smoke pay high premiums as compared to non-smokers.
- **Region**: The location where you stay determines your policy premium cost. For certain geographic locations, the premium rates are high due to lack of healthy food options, climate and health issues.

The distribution of features in the dataset is shown in the figures below. The Fig. 4.2.2 shows the gender division in the dataset. Fig 4.2.3 shows distribution of the customer smokes cigarettes in their day-to-day life. Fig 4.2.4 shows the distribution of customers according to the region.

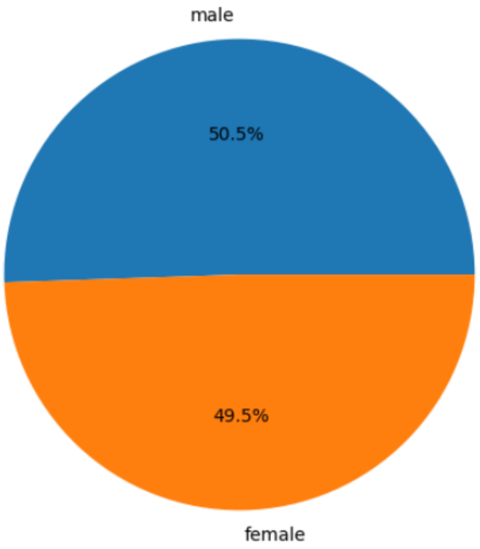


Fig. 3.2.2 Shows the distribution of customers according to gender

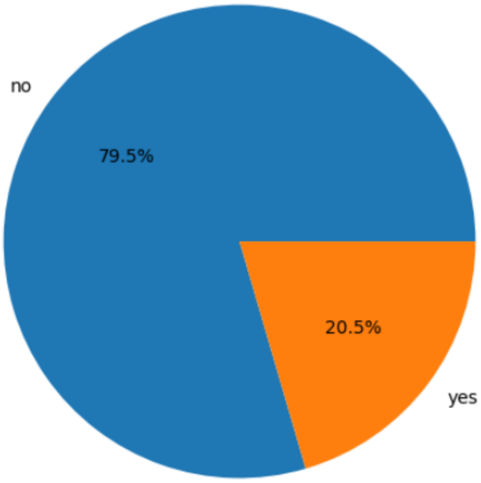


Fig 3.2.3 shows the distribution of customers that smoke.

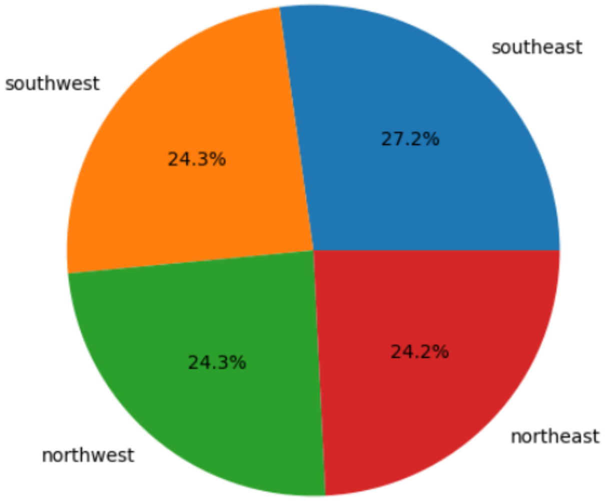


Fig. 3.2.4 shows the distribution of customers according to region

3.3 Algorithm Selection

For predicting health insurance prices, Multivariable Linear Regression models are used. Many different and popular linear regression models are used in this project depending upon their use case and accuracy of the model. This helps in building a better predictive model as it better fits the data and gives the customer more accurate premium estimates. This project uses Multivariable Linear Regression concepts and implements it using four different models altogether. The models used are:

1. **Support Vector Machines (SVM)**: Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It is particularly effective in high-dimensional spaces and is widely used for tasks such as image classification, text classification, and hand-written character recognition. SVMs are known for their ability to handle both linear and non-linear relationships in data. The SVM finds the hyperplane (two-dimensional space, a hyperplane is a line. In three-dimensional space, it becomes a plane, and so on) that maximizes the margin between classes. The margin is the distance between the decision boundary and the nearest data point from each class. SVM aims to maximize this margin. While SVM is primarily used for classification, it can also be adapted for regression tasks. In regression, the objective is to fit as many instances as possible within a specified margin.

   Support Vector Machines have several advantages:

   - Effectiveness in high-dimensional spaces.
   - Versatility in handling both linear and non-linear relationships.
   - Robustness against overfitting.

   Support Vector Machines also have some disadvantages as the choice of the kernel and parameters like C requires careful consideration, and SVMs may be sensitive to the choice of these parameters.

2. **Random Forest**: It is an ensemble learning algorithm in machine learning that is widely used for both classification and regression tasks. It belongs to the family of tree-based models and is known for its robustness and high predictive accuracy. The term "ensemble" refers to the idea of combining the predictions of multiple weak learners to create a stronger overall model. Decision trees are recursive structures that make decisions by recursively splitting the data based on features, creating a tree-like structure of nodes, branches, and leaves. Random Forest creates an

3. ensemble of decision trees. It trains multiple decision trees independently, and the final prediction is a combination (average for regression or voting for classification) of the predictions of individual trees. Random Forest uses randomization to better train its trees. Each tree in the forest is trained on a random subset of the training data. This process is known as bootstrapping or bagging. At each node of a decision tree, a random subset of features is considered for splitting. This helps in decorrelating the trees and increasing the diversity of the forest. Random Forest can provide insights into the importance of different features in making predictions. This is calculated based on the reduction in impurity (for classification) or mean squared error (for regression) brought about by each feature. Random Forest is known for its ability to achieve high predictive accuracy and generalization on various types of datasets. It often performs well without extensive hyperparameter tuning.
4. **Gradient Boost:** It is an ensemble machine learning technique that builds a predictive model in the form of an ensemble of weak learners, typically decision trees. The primary objective of gradient boosting is to combine the predictions of these weak learners, each trained on the errors of the preceding ones, to create a strong predictive model. Gradient boosting is particularly powerful and is widely used for both classification and regression tasks. Gradient boosting uses weak learners, often shallow decision trees, as building blocks. These are also referred to as "base learners" or "weak models." The term "gradient" in gradient boosting comes from the optimization process. The algorithm minimizes a loss function by iteratively adding weak learners to the model. In each iteration, the weak learner is trained on the residuals (errors) of the combined model from the previous iteration. The choice of loss function depends on the type of problem being addressed. For regression tasks, mean squared error is commonly used, while for classification tasks, log loss or exponential loss may be employed. The algorithm proceeds through multiple boosting iterations, with each iteration introducing a new weak learner. The overall model is the sum of the weak learners' predictions. A learning rate parameter controls the contribution of each weak learner to the overall model. A lower learning rate makes the learning process more robust but requires more iterations.
5. **XGBoost:** XGBoost, or Extreme Gradient Boosting, is a popular and highly efficient machine learning algorithm that belongs to the family of gradient boosting methods. Developed by Tianqi Chen, XGBoost is known for its speed, scalability, and high predictive performance. It is widely used for both classification and regression tasks and has been successful in various machine learning competitions. XGBoost is an implementation of the gradient boosting framework, which builds an ensemble of

6. weak learners (typically decision trees) in a sequential manner, where each learner corrects the errors made by the previous ones. XGBoost includes a regularization term in the objective function, allowing it to control the complexity of the weak learners and prevent overfitting. This regularization term includes parameters like alpha (L1 regularization) and lambda (L2 regularization). XGBoost has a built-in capability to handle missing values in the dataset. It can automatically learn how to handle missing data during the training process.

3.4 Tools Used

During the development of this project, there are many python libraries used named NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, Streamlit and Feature-engine.

1. NumPy : NumPy, short for Numerical Python, is a powerful and widely used open-source library in the Python programming language for numerical and mathematical operations. It provides support for large, multi-dimensional arrays and matrices, along with a collection of high-level mathematical functions to operate on these arrays. NumPy is a fundamental library for data science, scientific computing, and machine learning in Python.
2. Pandas: Pandas is an open-source data manipulation and analysis library for the Python programming language. It provides data structures for efficiently storing and manipulating large datasets and tools for working with structured data, such as tabular data, time-series, and more. Pandas is a crucial library in the Python data science ecosystem and is widely used for data cleaning, exploration, and analysis.
3. Matplotlib: Matplotlib is a comprehensive and widely-used data visualization library in the Python programming language. It enables the creation of a variety of static, animated, and interactive plots and charts, making it a fundamental tool for data scientists, researchers, and analysts. Matplotlib provides a flexible and customizable interface for creating high-quality visualizations for both exploratory data analysis and presentation.
4. Seaborn: Seaborn is a statistical data visualization library built on top of Matplotlib. It provides an interface for creating informative and attractive statistical graphics. Seaborn simplifies the process of creating complex visualizations and enhances the default aesthetics of Matplotlib plots. The library is particularly useful for exploring relationships in complex datasets and is widely used in the data science and statistical

analysis communities.

5. Scikit-learn: Scikit-learn, often abbreviated as sklearn, is a comprehensive and widely used machine learning library for the Python programming language. It provides simple and efficient tools for data analysis and modeling, including various machine learning algorithms, preprocessing techniques, model evaluation, and utilities for tasks such as feature selection and dimensionality reduction.
6. Feature-engine: Feature-engine is a Python library with multiple transformers to engineer and select features to use in machine learning models. Feature-engine preserves Scikit-learn functionality with methods fit() and transform() to learn parameters from and then transform the data.
7. Streamlit: Streamlit is an open-source Python library that allows you to create web applications for data science and machine learning with minimal effort. It is designed to simplify the process of turning data scripts into shareable web applications. Streamlit is particularly popular for its ease of use, interactive features, and rapid development capabilities.

3.5 Data Preprocessing and Analysis

Data preprocessing and analysis are critical steps in machine learning workflows. These steps involve preparing and cleaning the raw data to make it suitable for training machine learning models. For this step, firstly the dataset is loaded to a pandas dataframe.

Loading the dataset:

```
health = pd.read_csv("medical-charges.csv")
```

To use this dataset, the data points in the dataset shouldn’t be empty. To check this method is used:

```
health.isnull().sum()
```

If there is any empty data point present, that data point is dropped using:

```
health1=health.dropna(subset=['premium'])
```

The last step used, is encoding the categorical features since machine learning algorithms train best on numerical values.

Checking the data for any outlying data points that will interfere with the model and make it overfit. Boxplot is used to check the outlying data visually. During this process only BMI feature had some outlying data and is shown in the Fig 5.2.1.

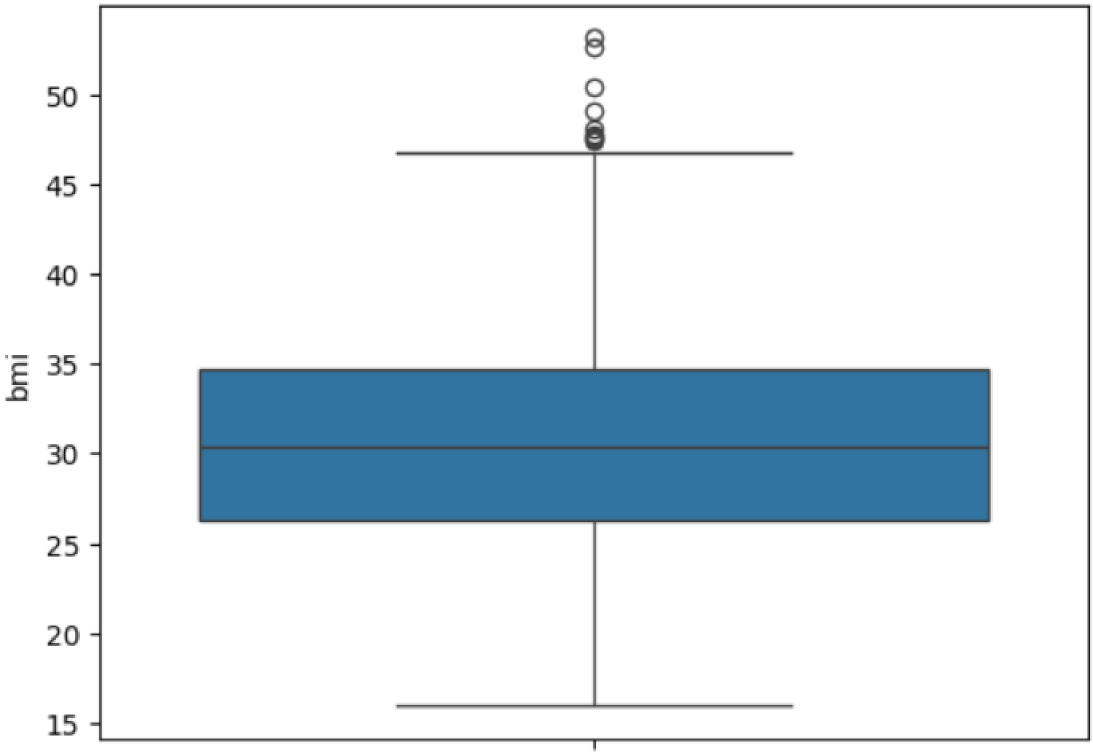


Fig. 3.5.1 Boxplot of BMI

To clean these outlying data we use:

```
      q1 = health1['bmi'].quantile(0.25)
q2 = health1['bmi'].quantile(0.5)
q3 = health1['bmi'].quantile(0.75)
iqr = q3 - q1
lowlim = q1- 1.5*iqr
upplim = q3 + 1.5*iqr
print(lowlim)
print(upplim)
arb=ArbitraryOutlierCapper(min_capping_dict={'bmi':lowlim},max_capping_dict=
{'bmi':upplim})
health1[['bmi']]=arb.fit_transform(health1[['bmi']])
sns.boxplot(health1['bmi'])
plt.show()
```

Thus resulting in outliers free data shown in Fig 5.2.2.

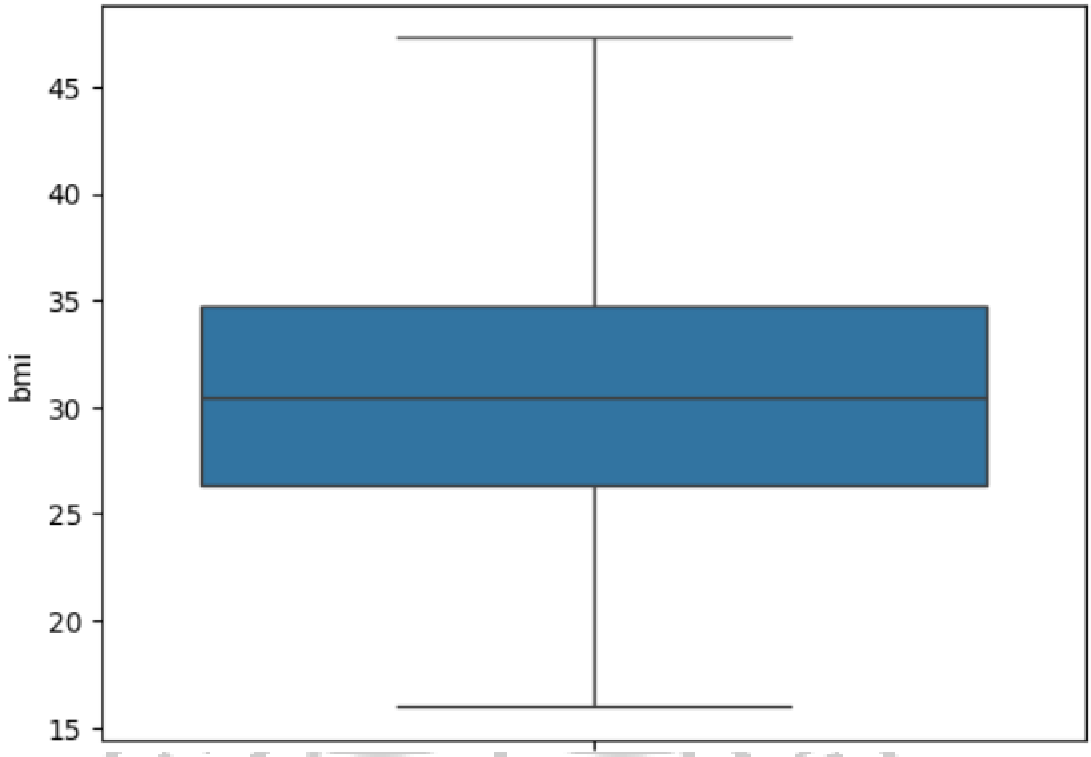


Fig. 3.5.2 Outliers Free BMI Data

Analysis Details:

Analyzing the dataset, to better understand about the data given. Calculating the statistical parameters of the data to know the distribution of the data for different features.

```
print(health1.describe().T)
```

Visualizing the dataset:

It is the representation of data in graphical or visual formats to help people understand the patterns, trends, and insights within the data. Visualization is a powerful tool in the data analysis process, as it allows for the exploration and communication of complex information in a more intuitive and accessible manner. For this step, Microsoft's Power BI which is an industry standard tool for visualization. Data visualization is shown in Fig. 5.2.3 and Fig. 5.2.4.

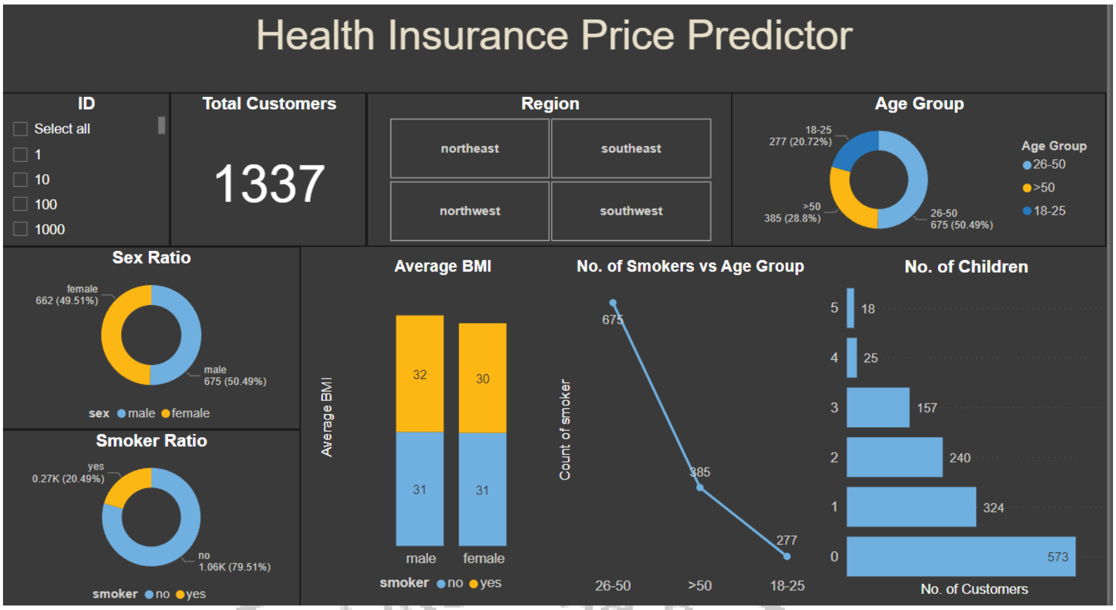


Fig. 3.5.3 Data Visualization using Power BI

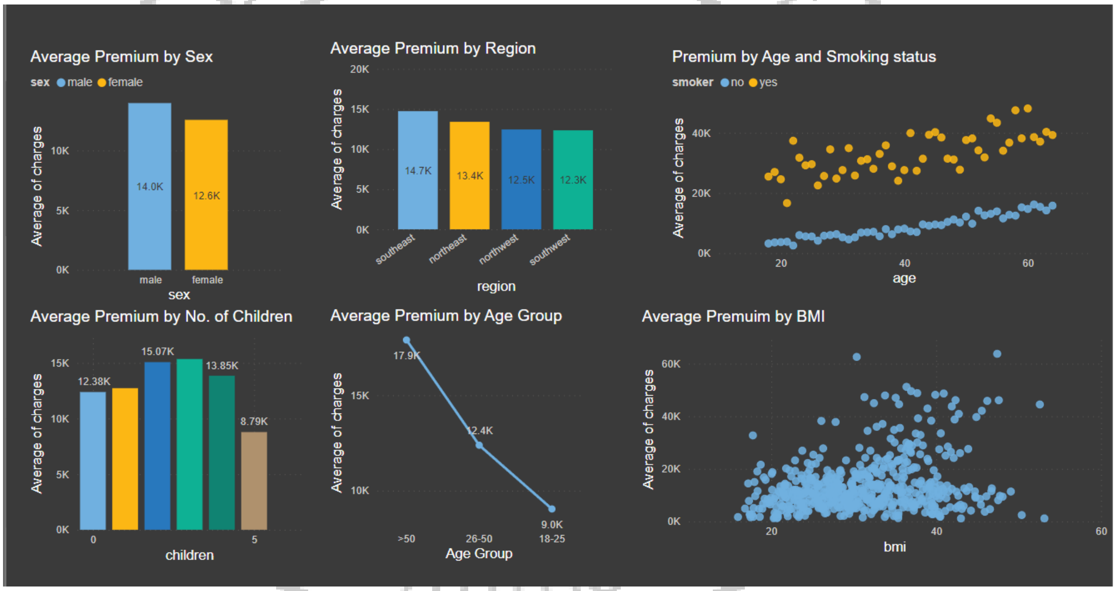
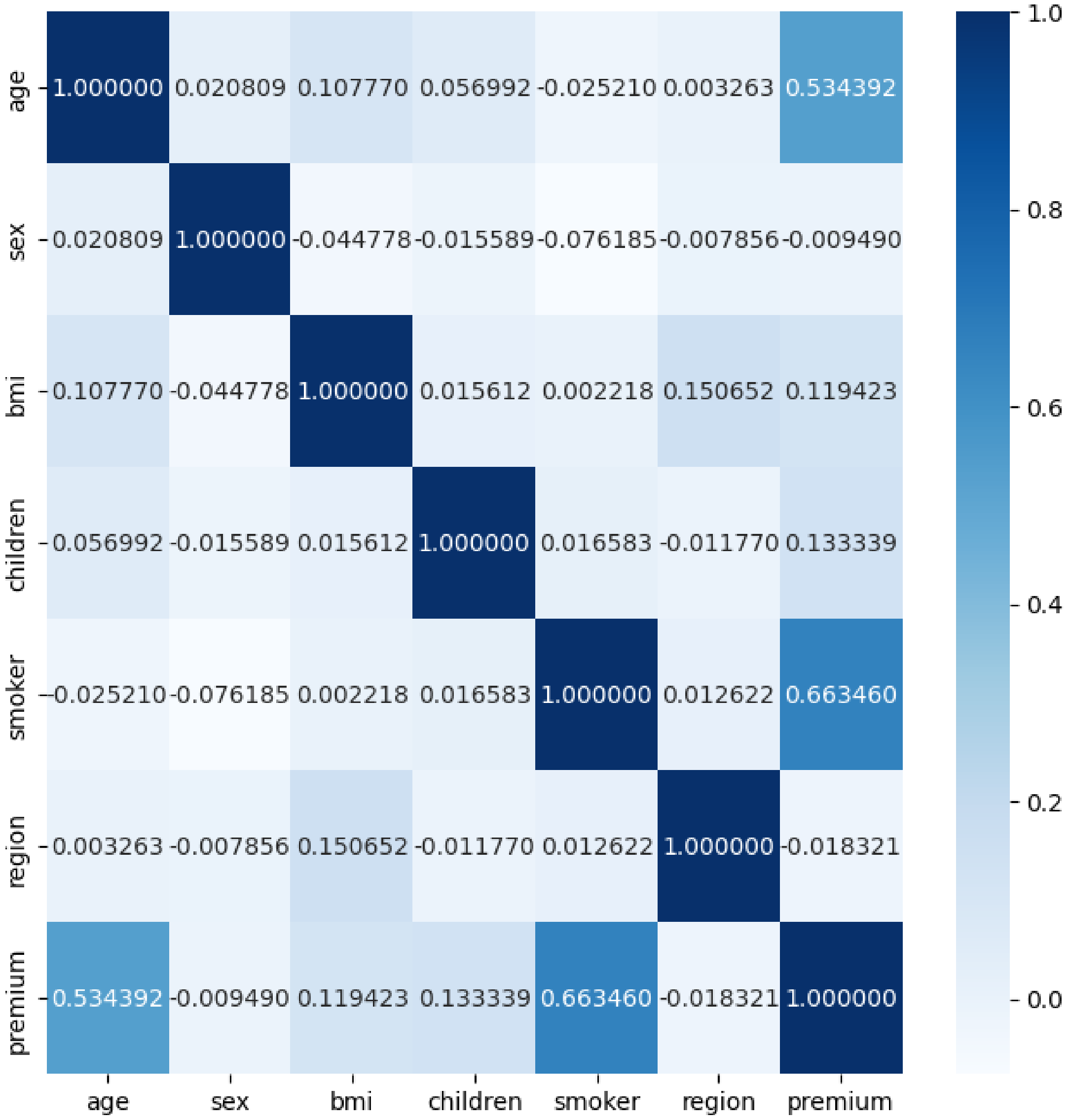


Fig. 3.5.4 Data Visualization using Power BI

The correlation between features in machine learning refers to the statistical measure of how closely two variables move in relation to each other. Specifically, it quantifies the degree to which a change in one variable corresponds to a change in another.

To get the correlation between all the features, python's library Seaborn is used to visualized in Fig. 5.2.5.



	age	sex	bmi	children	smoker	region	premium
age	1.000000	0.020809	0.107770	0.056992	-0.025210	0.003263	0.534392
sex	0.020809	1.000000	-0.044778	-0.015589	-0.076185	-0.007856	-0.009490
bmi	0.107770	-0.044778	1.000000	0.015612	0.002218	0.150652	0.119423
children	0.056992	-0.015589	0.015612	1.000000	0.016583	-0.011770	0.133339
smoker	-0.025210	-0.076185	0.002218	0.016583	1.000000	0.012622	0.663460
region	0.003263	-0.007856	0.150652	-0.011770	0.012622	1.000000	-0.018321
premium	0.534392	-0.009490	0.119423	0.133339	0.663460	-0.018321	1.000000

Fig. 3.5.5 Correlation Matrix

3.6 Implementation

Now, the final stage of implementing machine learning models. Since, this project uses 4 different models to get the better understanding of machine learning and how different perform with this dataset. Finalizing the model with the best accuracy.

1. **Support Vector Machine**:

   Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It is particularly effective in high-dimensional spaces and is widely used for tasks such as image classification, text classification, and hand-written character recognition.

   ```
   from sklearn.metrics import r2_score
   svrmodel=SVR()
   svrmodel.fit(xtrain,ytrain)
   ypredtrain1=svrmodel.predict(xtrain)
   ypredtest1=svrmodel.predict(xtest)
   ```

2. **Random Forest Regression**:

   It is an ensemble learning algorithm in machine learning that is widely used for both classification and regression tasks. It belongs to the family of tree-based models and is known for its robustness and high predictive accuracy.

   ```
   rfmodel=RandomForestRegressor(random_state=21)
   rfmodel.fit(xtrain,ytrain)
   ypredtrain2=rfmodel.predict(xtrain)
   ypredtest2=rfmodel.predict(xtest)
   print(r2_score(ytrain,ypredtrain2))
   print(r2_score(ytest,ypredtest2))
   print(cross_val_score(rfmodel,X,Y,cv=5,).mean())
   from sklearn.model_selection import GridSearchCV
   estimator=RandomForestRegressor(random_state=42)
   param_grid={'n_estimators':[10,40,50,98,100,120,150]}
   grid=GridSearchCV(estimator,param_grid,scoring="r2",cv=5)
   ```

```
grid.fit(xtrain,ytrain)
print(grid.best_params_)
rfmodel=RandomForestRegressor(random_state=42,n_estimators=120)
rfmodel.fit(xtrain,ytrain)
ypredtrain2=rfmodel.predict(xtrain)
ypredtest2=rfmodel.predict(xtest)
```

3. **Gradient Boosting Regression**:
   It is an ensemble machine learning technique that builds a predictive model in the form of an ensemble of weak learners, typically decision trees.

```
gbmodel=GradientBoostingRegressor()
gbmodel.fit(xtrain,ytrain)
ypredtrain3=gbmodel.predict(xtrain)
ypredtest3=gbmodel.predict(xtest)
print(r2_score(ytrain,ypredtrain3))
print(r2_score(ytest,ypredtest3))
print(cross_val_score(gbmodel,X,Y,cv=5,).mean())
from sklearn.model_selection import GridSearchCV
estimator=GradientBoostingRegressor()
param_grid={'n_estimators':[10,15,19,20,21,50],'learning_rate':[0.1,0.19,0.2,0.21,0.8,1]}
grid=GridSearchCV(estimator,param_grid,scoring="r2",cv=5)
grid.fit(xtrain,ytrain)
print(grid.best_params_)
gbmodel=GradientBoostingRegressor(n_estimators=19,learning_rate=0.2)
gbmodel.fit(xtrain,ytrain)
ypredtrain3=gbmodel.predict(xtrain)
ypredtest3=gbmodel.predict(xtest)
```

4. **XGBoost Regressor**:
   XGBoost, or Extreme Gradient Boosting, is a popular and highly efficient machine learning algorithm that belongs to the family of gradient boosting methods. Developed by Tianqi Chen, XGBoost is known for its speed, scalability, and high predictive performance

```
xgmodel=XGBRegressor()
xgmodel.fit(xtrain,ytrain)
ypredtrain4=xgmodel.predict(xtrain)
ypredtest4=xgmodel.predict(xtest)
print(r2_score(ytrain,ypredtrain4))
print(r2_score(ytest,ypredtest4))
print(cross_val_score(xgmodel,X,Y,cv=5,).mean())
from sklearn.model_selection import GridSearchCV
estimator=XGBRegressor()
param_grid={'n_estimators':[10,15,20,40,50],'max_depth':[3,4,5],'gamma':[0,0.15,0.3,0.5,1]}
grid=GridSearchCV(estimator,param_grid,scoring="r2",cv=5)
grid.fit(xtrain,ytrain)
print(grid.best_params_)
xgmodel=XGBRegressor(n_estimators=15,max_depth=3,gamma=0)
xgmodel.fit(xtrain,ytrain)
ypredtrain4=xgmodel.predict(xtrain)
ypredtest4=xgmodel.predict(xtest)
```

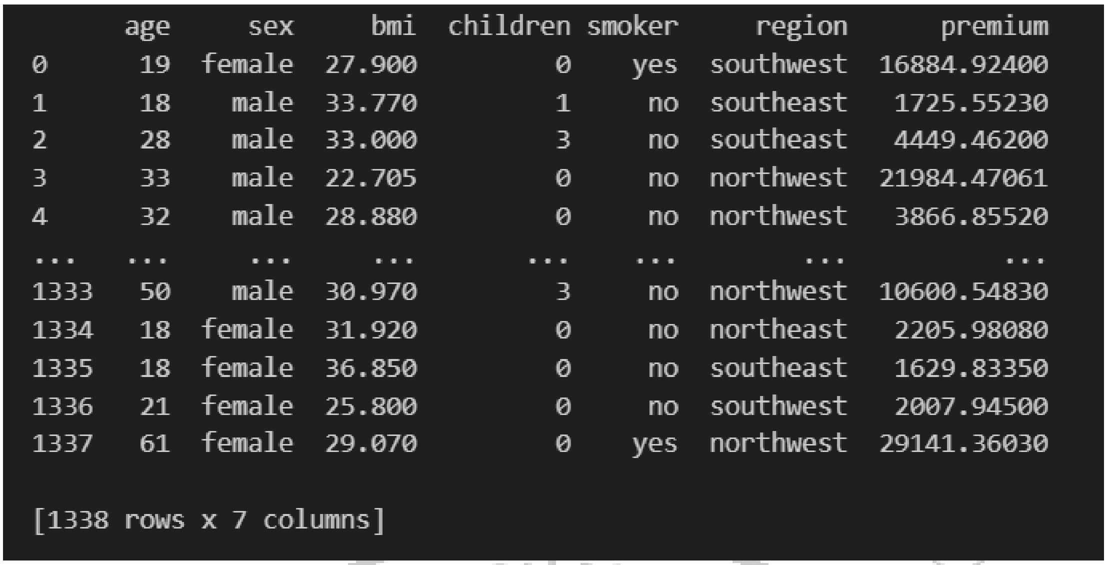
3.7 **Results Obtained**

Result 1

For this project, we have implemented 4 different kinds of Multivariable Linear Regression models:

1. Support Vector Machine (SVR) Regression
2. Random Forest Regression
3. Gradient Boost Regression
4. Extreme Gradient Boost (XGBoost) Regression

Among these 4 models, we have compared the model that gives us the best combination of testing and training accuracy on our dataset shown in Fig. 6.1.1.

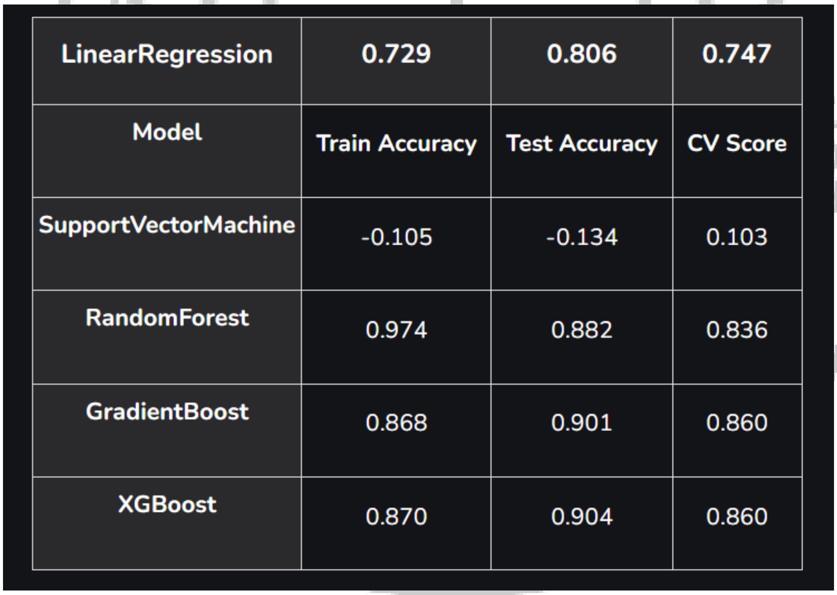


```
      age     sex     bmi  children smoker     region      premium
0      19  female  27.900         0    yes  southwest  16884.92400
1      18    male  33.770         1     no  southeast   1725.55230
2      28    male  33.000         3     no  southeast   4449.46200
3      33    male  22.705         0     no  northwest  21984.47061
4      32    male  28.880         0     no  northwest   3866.85520
...   ...     ...     ...       ...    ...        ...          ...
1333   50    male  30.970         3     no  northwest  10600.54830
1334   18  female  31.920         0     no  northeast   2205.98080
1335   18  female  36.850         0     no  southeast   1629.83350
1336   21  female  25.800         0     no  southwest   2007.94500
1337   61  female  29.070         0    yes  northwest  29141.36030

[1338 rows x 7 columns]
```

Fig. 3.7.1 Health Insurance Dataset

The comparison of the testing and training accuracy of these models in shown in the Fig 3.7.2



LinearRegression	0.729	0.806	0.747
Model	Train Accuracy	Test Accuracy	CV Score
SupportVectorMachine	-0.105	-0.134	0.103
RandomForest	0.974	0.882	0.836
GradientBoost	0.868	0.901	0.860
XGBoost	0.870	0.904	0.860

Fig. 3.7.2 Testing & Training Accuracy Comparison

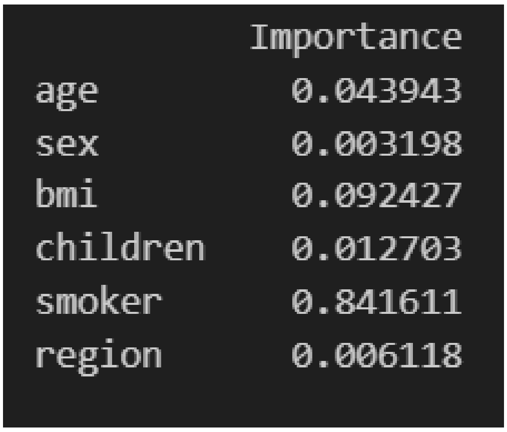
Supervised machine learning uses algorithms to train a model to find patterns in a dataset with labels and features and then uses the trained model to predict the labels on a new dataset's features.

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. Decision trees create a model that predicts the label by evaluating a tree of if-then-else true/false feature questions, and estimating the minimum number of questions needed to assess the probability of making a correct decision. Decision trees can be used for classification to predict a category, or regression to predict a continuous numeric value. In the simple example below, a decision tree is used to estimate a house price (the label) based on the size and number of bedrooms (the features). A Gradient Boosting Decision Trees (GBDT) is a decision tree ensemble learning algorithm similar to random forest, for classification and regression. Ensemble learning algorithms combine multiple machine learning algorithms to obtain a better model.

The term “gradient boosting” comes from the idea of “boosting” or improving a single weak model by combining it with a number of other weak models in order to generate a collectively strong model. Gradient boosting is an extension of boosting where the process of additively generating weak models is formalized as a gradient descent algorithm over an objective function. Gradient boosting sets targeted outcomes for the next model in an effort to minimize errors. Targeted outcomes for each case are based on the gradient of the error (hence the name gradient boosting) with respect to the prediction. XGBoost is a scalable and highly accurate implementation of gradient boosting that pushes the limits of computing power for boosted tree algorithms, being built largely for energizing machine learning model performance and computational speed. With XGBoost, trees are built in parallel, instead of sequentially like GBDT. It follows a level-wise strategy, scanning across gradient values and using these partial sums to evaluate the quality of splits at every possible split in the training set.

Result 2

With the help of Exploratory Data Analysis (EDA), the most important features for the prediction of health insurance prices are smoking status of the customers and their age. The Fig. 6.2.1 shows the importance of all the variable features in the prediction of the health insurance price.

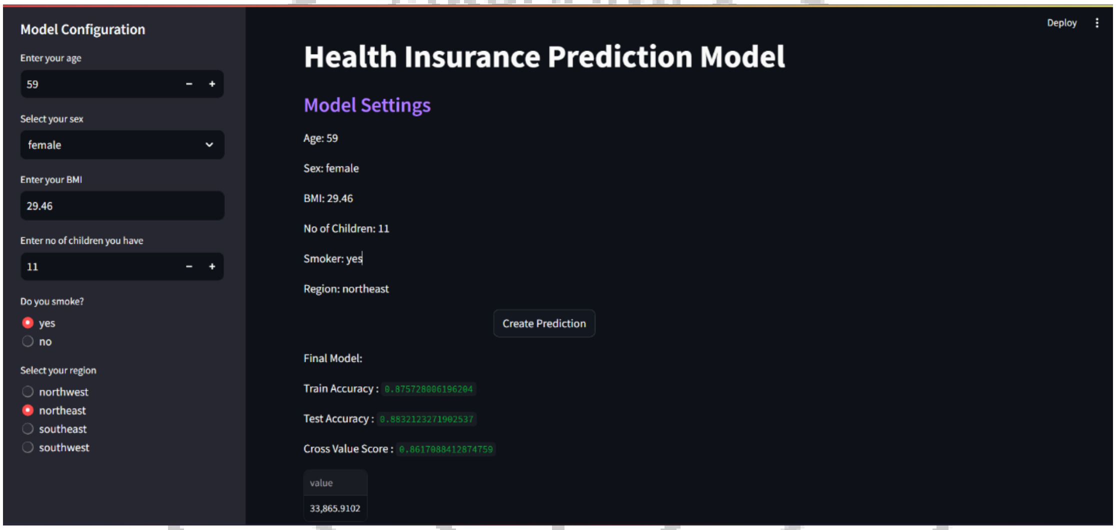


```
          Importance
age         0.043943
sex         0.003198
bmi         0.092427
children    0.012703
smoker      0.841611
region      0.006118
```

Fig. 3.7.3 Importance of Features

Result 3

The health insurance price predicting model build using the Multivariable Linear Regression concept specifically Extreme Gradient Boost or XGBoost has the highest combination of test and training accuracy. Thus, in our final model we are using XGBoost as our prediction model. Using XGBoost, a Graphical User Interface (GUI) or an app has been build. This app is built using streamlit. This app can be used by the customers to predict the price of health insurance by themselves. Customers would need to input their data according to their requirements and the predicting model will give them the price of health insurance need by them to pay. Fig. 6.3.1 shows the GUI developed.



Model Configuration

Enter your age: 59

Select your sex: female

Enter your BMI: 29.46

Enter no of children you have: 11

Do you smoke?
● yes
○ no

Select your region
○ northwest
● northeast
○ southeast
○ southwest

Health Insurance Prediction Model

Model Settings

Age: 59

Sex: female

BMI: 29.46

No of Children: 11

Smoker: yes

Region: northeast

Final Model:

Train Accuracy : 0.875728006196204

Test Accuracy : 0.8832123271902537

Cross Value Score : 0.8617088412874759

value
33,865.9102

Fig. 3.7.4 GUI Developed

CHAPTER 4

REFLECTIONS

CHAPTER 4
REFLECTIONS

4.1 Technical Skills Acquired

Throughout the project, I acquired an array of technical skills vital for machine learning (ML) and data science. This included a deepened understanding of Python programming, where I extensively used libraries like Pandas, NumPy, and Scikit-learn for tasks such as data preprocessing, feature engineering, model training, and evaluation. I also delved into data visualization techniques using Matplotlib and Seaborn to create insightful visualizations that aided in the interpretation of complex data patterns. Additionally, I became proficient in applying various ML algorithms such as logistic regression, decision trees, and ensemble methods to classify health insurance price prediction. This hands-on experience with implementing ML models allowed me to navigate through the intricacies of model selection, hyperparameter tuning, and performance evaluation effectively.

4.2 Soft Skills Acquired

In addition to technical skills, this project facilitated the development of several soft skills crucial for success in collaborative and dynamic environments. Effective communication emerged as a cornerstone skill as I regularly interacted with team members, stakeholders, and end-users to gather requirements, provide progress updates, and solicit feedback. I refined my ability to ll complex technical concepts into easily understandable insights, ensuring alignment among team members with diverse backgrounds and expertise. Furthermore, the project emphasized the importance of proactive problem-solving, as I encountered challenges such as data inconsistencies, model performance issues, and user feedback integration. By approaching these challenges with resilience and adaptability, I cultivated a robust problem-solving mindset, enabling me to iteratively refine solutions and drive project progress effectively. Additionally, I honed my project management skills by prioritizing tasks, managing timelines, and coordinating efforts to meet project milestones and deliverables consistently.

4.3 Conclusion and Future Scope

4.3.1 Conclusions

This project is aimed at the dark side of health insurance companies. This project can be used to get an estimate of the premium need to be paid for health insurance in accordance to the general features. This helps customers that aren't aware about how the premium for health insurance are calculated by these companies. There are some instances where the seller would coax the customer into buying higher premium health insurance while telling white lies and exaggerating the profits gained from the plan but the customer requirements would have been satisfied by another cheaper plan of health insurance. Since the customer is not aware of the reality of his plan, he might claim it under circumstances not covered by the insurance.

4.3.2 Future Scope

The future scope of the health insurance price predictor can encompass various aspects, driven by advancements in technology, changes in the healthcare landscape, and evolving data science methodologies. These can be listed as but not limited to :

- Incorporation of Advanced Features: Explore the inclusion of additional and more granular features, such as genetic information, wearable device data, or lifestyle factors. Advancements in data collection methods and technology can contribute to a more comprehensive understanding of individuals' health profiles.

- Collaboration with Experts: Collaboration with healthcare professionals and domain experts can provide valuable insights for feature selection and model enhancement. Overall, continuous iteration and refinement are essential to ensure the health insurance price predictor remains robust and aligns with evolving industry standards and societal expectations.

- Integration of Continuous updating dataset: incorporating real-time data updates and continuous model training would ensure adaptability to evolving healthcare trends. Further research into the ethical considerations of using certain features in premium determination is warranted to address fairness and regulatory compliance.

- Machine Learning Model Enhancements: Experiment with more sophisticated machine learning models beyond linear regression. Models like ensemble methods, gradient boosting, or deep learning architectures may capture complex relationships within the data and improve prediction accuracy.

- Dynamic Pricing Models: Explore the development of dynamic pricing models that adjust insurance premiums in real-time based on changing health conditions, lifestyle choices, or other relevant factors. This could create a more responsive and personalized pricing structure.

- Blockchain for Data Security: Explore the use of blockchain technology to enhance data security and privacy. Blockchain can provide a decentralized and tamper-resistant system for storing and sharing sensitive health-related information, addressing concerns about data integrity and privacy.

- Personalized Care Plans: Develop personalized recommendations for individuals to optimize their insurance coverage and premiums. This could involve providing insights on how specific lifestyle changes or preventive measures could positively impact insurance costs.

- Consumer Education and Engagement: Implement strategies for educating consumers about how the prediction model works and how certain choices may impact their insurance premiums. Enhance engagement through user-friendly interfaces, interactive tools, and clear communication.

REFERENCES

1. Md. Shohel Rana , Jeff Shuford “AI in Healthcare: Transforming Patient Care through Predictive Analytics and Decision Support Systems” , Jan 22, 2024
2. Ugochukwu Orji, Elochukwu Ukwandu,“Machine Learning for an explainable cost prediction of medical insurance”, 24 November 2023
3. Abhyudaya Bora; Ritika Sah; Alabhya Singh; Deepak Sharma; “Interpretation of machine learning models using XAI - A study on health insurance dataset”, 13-14 October 2022
4. Yang Xie, C.W. Chang, Sandra Neubauer, “Predicting Days in hospital using health Insurance Claims” , 2020.

5. H. Chen Jonathan, M. Asch Steven, “Medical Expense Prediction System using Machine learning Techniques and Intelligent Fuzzy Approach” , 2020