भारतीय सूचना प्रौद्योगिकी संस्थान वडोदरा
★ INDIAN INSTITUTE OF INFORMATION TECHNOLOGY VADODARA ★
|| उद्यमस्य समं किमपि न भवति ||

# Predictive Modeling for Healthcare and Business Applications

Name : Kudithi Krishna Kishore
Student ID : 202111044
Institute name : IIIT-Vadodara - International Campus Diu (IIITV-ICD)
Organisation : Coincent.ai
Mentor Name : Kishore Kumar

***Abstract*—This report presents an overview of two machine learning projects undertaken during the summer internship: Diabetes Prediction Part and Lead Scoring Case Study . The diabetes prediction model, implemented using Python libraries like Streamlit, Pandas, and Scikit-learn, aims to classify individuals as diabetic or non-diabetic based on clinical features. The lead scoring model, built with Pandas, NumPy, Seaborn, Statsmodels, and Scikit-learn, predicts the likelihood of lead conversion for a marketing campaign. Both projects involved data preprocessing, feature engineering, model training, and evaluation. The internship provided valuable experience in data analysis, model development, and problem-solving.**

***Index Terms*—Diabetes prediction, Lead scoring, Machine learning, Data analysis, Data preprocessing, Model evaluation, Python programming, Scikit-learn, Pandas, NumPy, Streamlit, Statsmodels, Logistic regression, Classification, Prediction**

## I. Introduction

Myself, Kudithi Krishna Kishore, 4th year student pursuing a BTech degree in Department of Computer Science at IIIT-Vadodara - International Campus Diu (IIITV-ICD), undertook a summer internship at Coincent.ai from 01st Jun, 2024 to 31st Jul, 2024. Coincent.ai is an organization that offers live industrial training, live projects and internships, and placement preparation. The internship provided an invaluable opportunity to apply theoretical knowledge to real-world challenges and gain practical experience in Machine Learning using Python. The organization's focus on machine learning and data analytics aligned seamlessly with my academic pursuits, making it an ideal platform to explore the industry's current trends and challenges.

## II. Project Description

The internship involved two projects: Diabetes Prediction Part and Lead Scoring Case Study.

### A. Diabetes Prediction Part

This project focused on developing a machine learning model to predict the potential onset of diabetes in individuals. The model utilized patient data from a publicly available dataset ("diabetes.csv"). My key responsibilities encompassed:

*1) Data Preprocessing:* I cleaned and transformed the data to ensure its quality and consistency for model training. This included handling missing values, identifying outliers, and potentially encoding categorical features. (Specific actions based on the code might be added here)

*2) Feature Engineering:* I analyzed the data and potentially created new features to enhance the model's ability to learn relationships between variables and the target variable (diabetes diagnosis). This could involve feature scaling or creating interaction terms based on domain knowledge.

*3) Model Selection and Training:* I experimented with various machine learning algorithms, such as the Random Forest Classifier implemented in the code, to identify the one that best predicts diabetes based on patient information. The code demonstrates splitting the data into training and testing sets, training the model, and evaluating its performance.

*4) Visualization:* I created visualizations (scatter plots) to compare the user's input data points against the overall data distribution, allowing for a visual interpretation of potential health risks.

This project Diabetes Prediction Part contributed to the internship's focus on healthcare solutions by developing a model that could potentially be used for early diabetes detection and preventive measures.

### B. Lead Scoring Case Study

*1) Data Cleaning and Preparation:*

- The code imports necessary libraries like pandas, NumPy, and scikit-learn.
- It reads the lead data from a CSV file (Leads.csv).
- It explores the data by checking for missing values, data types, and value counts.
- It performs data cleaning steps like:
  - Dropping columns with too many missing values.
  - Dropping irrelevant columns like City and Country.

- Handling missing values by dropping rows or filling them (depending on the column).

*2) Feature Engineering:*

- The code creates dummy variables for categorical features using pd.get_dummies.
- It uses power transformers to normalize skewed numeric features.
- It performs feature selection using Recursive Feature Elimination (RFE) to select the most relevant features for the model.

*3) Model Building:*

- The code uses statsmodels to build a logistic regression model on the selected features.
- It checks for variables with high p-values (indicating weak statistical significance) and high VIF (Variance Inflation Factor) which can cause multicollinearity.
- It iteratively removes features with high p-values and VIFs, refitting the model after each removal.
- It checks the final model summary to ensure most features have p-values below a threshold 0.05 and VIFs below a certain value 5.

*4) Model Evaluation:*

- The code uses the trained model to predict the conversion probability for the training data.
- It creates a dataframe with actual conversion flags and predicted probabilities.
- It calculates various evaluation metrics like accuracy, confusion matrix, sensitivity, and specificity.
- It plots an ROC curve to visualize the model's performance and calculates the Area Under the Curve (AUC).
- It explores different probability cutoffs for classifying leads as ”converted” and analyzes the trade-off between sensitivity (correctly identifying true positives) and specificity (correctly identifying true negatives).

This project Lead Scoring case study demonstrates a process for building and evaluating a lead scoring model using logistic regression. It highlights the importance of data cleaning, feature selection, and model evaluation to achieve a reliable model for predicting lead conversion.

## III. Work Carried Out

### *A. Diabetes Prediction Part*

My role in the diabetes prediction project encompassed several key responsibilities:

*1) Data Acquisition and Preprocessing:* I sourced the diabetes dataset and performed essential preprocessing steps including handling missing values, outlier detection, and feature scaling using Python libraries like Pandas and NumPy.

*2) Exploratory Data Analysis (EDA):* I conducted exploratory data analysis to gain insights into the dataset, visualized data distributions, and identified potential correlations between variables.

*3) Model Development:* I implemented the Random Forest Classifier algorithm using Scikit-learn to predict diabetes onset based on patient attributes. I experimented with different hyperparameter configurations to optimize model performance.

*4) Model Evaluation:* I evaluated the model's performance using metrics such as accuracy, precision, recall, and F1-score. Additionally, I employed techniques like cross-validation to assess model robustness.

*5) Visualization and Interpretation:* I created visualizations to understand the model's predictions and identify potential areas for improvement. I developed a Streamlit-based interactive application to allow users to input patient data and receive diabetes risk predictions.

### *B. Lead scoring case study*

My contributions to the lead scoring project involved:

*1) Data Exploration and Cleaning:* I explored the lead dataset to understand its structure and identify potential issues like missing values and outliers. I implemented data cleaning techniques to ensure data quality.

*2) Feature Engineering:* I created new features from existing data to improve model performance. This included transforming categorical variables into numerical representations and combining relevant features.

*3) Model Building:* I developed a logistic regression model using Statsmodels to predict the likelihood of lead conversion. I employed feature selection techniques (RFE) to identify the most impactful variables.

*4) Model Evaluation:* I evaluated the model's performance using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. I performed model tuning to optimize the classification threshold.

Overall, both projects Diabetes Prediction Part and Lead scoring case study provided opportunities to apply data analysis, machine learning techniques, and Python programming skills to address real-world problems.

## IV. Experience and Learning Outcomes

The internship at Coincent.ai provided invaluable practical experience in applying theoretical knowledge to real-world challenges. Working on the Diabetes Prediction Part and Lead Scoring Case Study projects allowed me to develop a strong foundation in data analysis, machine learning, and problem-solving.

Key skills acquired during the internship include:

- Technical Skills: Proficiency in Python programming, data manipulation using Pandas, data visualization with Matplotlib and Seaborn, model development and evaluation using Scikit-learn and Statsmodels.
- Soft Skills: Effective communication, Personal responsibility, time management, and problem-solving abilities.

The internship significantly enhanced my understanding of the machine learning lifecycle, from data exploration to model evaluation. I gained insights into the importance of data quality, feature engineering, and model selection in achieving optimal results. Additionally, I developed a strong foundation in statistical analysis and model evaluation techniques.

## V. Future Prospects

The projects undertaken during the internship offer potential avenues for further collaboration with Coincent.ai. The Diabetes Prediction Part model could be refined by incorporating additional patient data and exploring advanced machine learning techniques. Similarly, the Lead Scoring Case Study model can be enhanced by incorporating real-time data and implementing a more sophisticated lead nurturing strategy.

The skills and knowledge acquired during the internship are directly applicable to my future studies and career goals. A strong foundation in data analysis, machine learning, and Python programming will be invaluable in pursuing advanced degrees or roles in data science and analytics. The ability to build and evaluate predictive models will be essential for addressing complex challenges in various industries.

## VI. Conclusion

The internship at Coincent.ai provided a comprehensive learning experience in the field of data science and machine learning. Through involvement in the Diabetes Prediction Part and Lead scoring Case Study projects, I gained practical exposure to data analysis, model development, and evaluation. The experience has significantly enhanced my technical and analytical skills, preparing me for future challenges in the industry.

I am immensely grateful to the organisation Coincent.ai for providing this invaluable opportunity. I would like to express my sincere appreciation to my mentor, Kishore Kumar sir, for their guidance and support throughout the internship. Their expertise and mentorship were instrumental in my growth as a data scientist.

## References

[1] "Benefits of Switching to an Electronic Health Record (EHR)," Practice Fusion, Nov. 22, 2016. https://www.practicefusion.com/healthinformatics-practical-guide-page-1 (accessed Sep. 21, 2022).

[2] "4 Health Care Data Challenges and How to Overcome Them," Corporate Compliance Insights, Jun. 28, 2018. https://www.corporatecomplianceinsights.com/4-health-care-datachallenges-overcome (accessed Sep. 21, 2022).

[3] HealthIT.gov, "What are the advantages of electronic health records?," HealthIT.gov, Mar. 08, 2022. https://www.healthit.gov/faq/what-areadvantages-electronic-health-records

[4] E. Brynjolfsson and K. McElheran, "The Rapid Adoption of Data-Driven Decision-Making," American Economic Review, vol. 106, no. 5, pp. 133–39, May 2016, doi: 10.1257/AER.P20161016.

[5] G. Shmueli and O. R. Koppius, "Predictive analytics in information systems research," MIS Quarterly: Management Information Systems, vol. 35, no. 3, pp. 553–572, 2011, doi: 10.2307/23042796.

[6] W. K. Lin, S. J. Lin, and T. N. Yang, "Integrated Business Prestige and Artificial Intelligence for Corporate Decision Making in Dynamic Environments," Cybernetics and Systems, vol. 48, no. 4, pp. 303–324, May 2017, doi: 10.1080/01969722.2017.1284533.