Cal

****

CS – 4661 Introduction to Data Science

**Identifying High-Risk Patients for Targeted Prevention and Intervention**

**Prepared by**

Mitali Purohit

Tammy Xaypraseuth

Carlos Ramirez

**Supervised By**

Dr. Mohammad Pourhomayoun

Ashkan Aledavoud(Teaching Assistant)

Ly Jacky Nhiayi (Teaching Assistant)

**November 10, 2025**

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| 1. **Introduction** | **X** |
| 1. **Methodology** | **X** |
| 1. **Findings & Discussion** | **X** |
| 1. **Conclusion** | **X** |
| 1. **Recommendations** | **X** |

1. **Introduction**

The "Identifying High-Risk Patients for Targeted Prevention and Intervention" project is a data-driven initiative within the healthcare domain. It seeks to leverage advanced data science and machine learning techniques to address a critical challenge facing the healthcare industry.

1. **Problem Statement**

The healthcare industry faces a growing need to identify high-risk patients who are more likely to develop medical complications. This project aims to address this challenge by leveraging data science and machine learning techniques to predict and target prevention and intervention efforts effectively. Our primary objective is to develop a predictive model that can:

* Identify patients at high risk of medical complications.
* Prioritize and tailor prevention and intervention strategies based on individual patient characteristics.
* Contribute to improved patient outcomes and resource allocation within the healthcare system.

This project will play a crucial role in supporting healthcare providers and policymakers in their efforts to improve patient care and reduce healthcare costs.

1. **Data Source: Patient Treatment Classification**

|  |  |  |
| --- | --- | --- |
| HAEMATOCRIT | HAEMOGLOBINS | ERYTHROCYTE |
| 33.8 | 11.1 | 4.18 |
| 44.6 | 14.0 | 6.86 |

**Feature Importance**

**Methods**: KNN regression, logistic regression, linear regression, and random forest regression

We evaluated the performance of the four models using the following metrics: Accuracy Precision Recall F1 score

**Correlation Analysis**

**Descriptive Statistics**

**Data Visualization**

Scatter Plot

Pie Chart

1. **Findings & Discussion**

|  |  |  |
| --- | --- | --- |
|  | **HAEMATOCRIT** | **HAEMATOCRIT** |
| 33.8 | 33.8 | 33.8 |
| 33.8 | 33.8 | 33.8 |
| 33.8 | 33.8 | 33.8 |

1. **Conclusion**

The "Identifying High-Risk Patients for Targeted Prevention and Intervention" project represents an academic endeavor that holds substantial promise within the healthcare domain. By leveraging the power of data science and machine learning, the project aims to improve patient care, reduce healthcare costs, and enhance resource efficiency.

This project is available [at GitHub Repository](https://github.com/MitaliP001/Identifying-High-Risk-Patients-for-Targeted-Prevention-and-Intervention).

1. **References**
2. Kaggle Datasets. (n.d.). "Patient Treatment Classification Dataset." Available online at: <https://www.kaggle.com/datasets/manishkc06/patient-treatment-classification>
3. M. Pourhomayoun, M. Shakibi, “[Predicting Mortality Risk in Patients with COVID-19 Using Artificial Intelligence to Help Medical Decision-Making](https://www.sciencedirect.com/science/article/pii/S2352648320300702?via%3Dihub#bib9),” the Journal of Elsevier Smart Health, 2020.
4. Python Software Foundation. (n.d.). "Python Programming Language." Available online at: <https://www.python.org/>
5. GitHub. (n.d.). "GitHub Documentation." Available online at: <https://docs.github.com/>