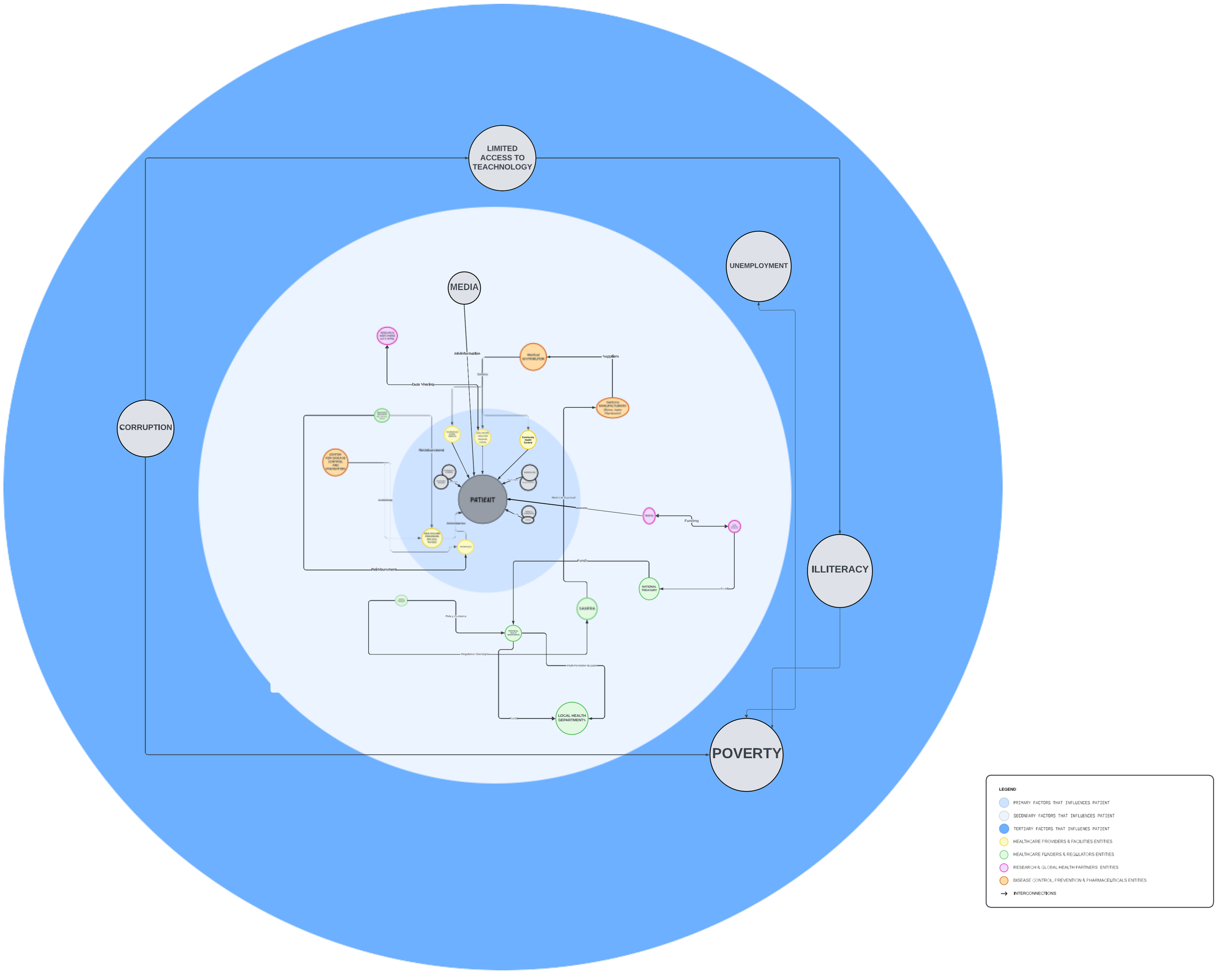
Enhancing Health Systems Through Data Science: An Analysis of Electronic Health Records (EHRs)

This Project Looks at the health care system and explores how data Science Techniques such as Predictive modeling , Clustering and Dimentionality reduction can be used to improve healthcare resource allocation and patient outcomes

REFLECTION

My journey diving into health systems science has greatly deepened my understanding of the complexities within healthcare delivery. I've come to realize just how much social factors, like where someone lives or their economic status, can impact their access to care and health outcomes.
Developing an entrepreneurial mindset in healthcare requires a dual focus on both technical and an understanding of the human side of health systems. This means finding ways to use technology and data to improve patient access and address gaps in care.
Combining data with community-driven initiatives can make a real difference in reducing healthcare disparities. I've learned that health systems require a holistic, multidisciplinary approach, inspiring me to think of myself as an innovator in healthcare, looking for ways to improve healthcare delivery and patient experiences.
I've discovered how data science can support integrated care models, improve population health, and reduce disparities. While data and technology provide powerful tools, solutions must address real-world barriers that patients face.

An Ecosystem Approach in Healthcare



Building blocks of a Health System

Health system are made up of building blocks and key players that depend on each other to create a healthcare system that ensures effective delivery of healthcare services, equitable access as well as improved health outcomes

The building blocks are

1. Service Delivery, this is the way in which healthcare services are provided and it Includes

- Public Health Clinics and Hospitals which are run by the government and their role is to provide healthcare services through immunization campaigns, school-based programs, or routine health services.
- Private Health Providers which are private hospitals, clinics, and family doctors and their role is to services to patients covered by private healthcare schemes or paying out-of-pocket.
- Community Health Workers (CHWs) which are trained healthcare personnels who work at the community level, often in underserved areas, their role is to promote healthcare, educate communities, and help with outreach programs.
- Mobile Health Units which are vehicles or teams that travel to remote areas to ensure that underserved and hard-to-reach populations receive healthcare

2. Health Workforce, this includes individuals responsible for training, deploying, and providing healthcare services. These include,

- Health Professionals such as Doctors, Nurses and Pharmacists
- Medical Training Institutions such as UCT, Wits and Stellenbosch that train doctors, nurses, and other healthcare workers.
- Regulatory bodies that oversee the practice and accreditation of health professionals.

3. Health Information Systems, these play a critical role in tracking, managing patient records, and ensuring accurate reporting of healthcare services. These include,

- Electronic Health Records (EHR)
- In South Africa the District Health Information Software 2 (DHIS2) is used to track health data

3. Access to Essential Medicines, these actors are involved in the procurement, distribution, and availability of medicines. These include,

- The South African Health Products Regulatory Authority (SAHPRA) which is responsible for Regulating medicines, ensuring that they meet safety and efficacy standards.
- Pharmaceutical companies such as Pfizer, Johnson & Johnson, AstraZeneca, etc. Who manufacture and supply medicines.

5. Financing, these actors are responsible for providing the necessary financial support for the procurement, distribution, and administration of healthcare services. These include,

- National Treasury which allocates funds to the Department of Health for the public
- Private Medical Aid Schemes such as discovery Health and Bonitas, who provide financial coverage for healthcare services for their members.
- Global Health Organizations such as the WHO and GAVI who provide financial and technical support for healthcare programs in various countries

6. Leadership and Governance, these actors who set policies, manage public health campaigns, and oversee health systems. These Include,

- The National Department of Health (NDoH) Which Governs the country's healthcare system and sets national health policies.
- Provincial Health Departments Who Oversee the implementation of health care programs at the provincial level and
- South African Health Products Regulatory Authority (SAHPRA) who Regulate and approve the distribution and use of medicines in South Africa.

Other Building Blocks include,

7. Media, NGOs, and Community Organisations, these actors who help inform and influence public attitudes

- Media Outlets such as TV, Radio, social media, communicate health messages to the public.
- Non-Governmental Organizations (NGOs), Organizations like Médecins Sans Frontières (MSF), Right to Care, and Treatment Action Campaign (TAC) engage in promotion, education, and outreach.
- Traditional Leaders and Community Influencers

8. Technology and Innovation, these include entities that provide tools and platforms for improving delivery, outreach and education.

- Telemedicine Platforms such as Vula Mobile, Hello Doctor enable remote consultations and education, providing health information
- Digital Campaign Platforms such as Healthsites.io, GovChat are platforms used for public health campaigns

Health Systems Challenges

Health care systems face challenges that are driven by socio economic inequalities, technological changes as well global dynamics. Some of these key challenges in the Health System include the

- Inequitable access to care with many communities, especially in rural or low-income areas struggling to access healthcare services. To counter this, the is a need for policies and programs that are targeted at ensuring equitable resource allocation and infrastructure development in those underdeveloped areas.

- Inequitable Healthcare delivery, this is persistent across different communities, especially in rural or underdeveloped areas resulting in decreased access to quality healthcare services
- Limited funding for healthcare services and infrastructure resulting in limited ability to deal with emerging and existing health needs. To counter this, there is need for sustainable funding models such as the NHI(National Health Insurance) and collaboration between public and private entities
- Many health systems fail to fully make use of technological innovations like health information systems, data analytics and telemedicine which could help to improve care , delivery and accessibility.
- Poor communication and coordination between various healthcare providers is another challenge which results inefficient use of resources and service gaps
- The South African healthcare system is overburdened and under-resourced. Political interferences, mismanagement, and a lack of accountability further burdens the health system

how data can work to improve the quality of health systems

Data plays an important role in Improving the quality of health systems by supporting evidence-based decisions making. Data collection and analysis helps public health providers and stakeholders to identify risk areas and make informed decisions about care and prioritize interventions. Data Guides the allocation of healthcare resources by monitoring the demand for healthcare services and ensuring that resources are allocated appropriately and responsibly.

Data monitors and provides insights regarding health and patient outcomes, ultimately informing Policy changes and decisions for example in the context of South Africa and the recent law regarding the National Health Insurance (NHI) which aims to provide universal healthcare, Data on population health, service utilization, and cost projections will be crucial in informing NHI policies and financial planning.

The utilization of Data can help improve access to healthcare by addressing the disparities in socioeconomic factors that affect access to healthcare in south Africa; by analysing data on healthcare accessibility and utilization by geographic region, policymakers can identify areas of need and develop targeted interventions aimed at reducing these disparities.

Data can be used as a powerful tool for improving health systems, when used effectively it can address access to equitable healthcare in south Africa as well as inform decision makers to optimize resources, drive policy changes and promote better health outcomes

The role of Data Science in Health Systems Science

Big data in healthcare refers to electronic health data sets so large and complex that they are difficult to manage with traditional software; nor can they be easily managed with traditional or common data management tools and methods. Big data in healthcare is overwhelming not only because of its volume but also because of the diversity of data types and the speed at which it must be managed.8 Furthermore, with health data increasing daily and more health information going digital, specific skills are required to organise, make sense of and manage this data

Data Science plays a role in the improvement and advancement of health systems in many ways through the use of predictive models and the utilizations of Data processes and applications

By making use of predictive models, Data Science can help improve healthcare systems. Predictive models in health science can identify patients at risk of developing diseases by analysing demographic and lifestyle factors for targeted interventions. They can predict disease onset by analysing genetic data and health records thus allowing for early detection and prevention furthermore, by analysing risk profiles and unique characteristics, predictive models can help personalise treatment plans. In South African predictive models can help alleviate the resource burden on the healthcare system, this can be done by using predictive models to identify patients at risk and as a result healthcare providers can allocate resources efficiently allowing for timeous and better health outcomes.

By making use of Data science processes and applications, Data science can help streamline processes for Greater efficiency. Natural Language Processing (NLP) applications automate the extraction of important and relevant information from unstructured data thus reducing the administrative burden on healthcare providers.

By analysing Workflow data and forecasting requirements in supply chain, healthcare institutions and providers can identify and address inefficiencies resulting in reduced waiting periods, reduced patient stress and overall improvement in resource management

Data science in health systems science is integral to building a more effective, efficient, and equitable healthcare system, especially in resource-constrained environments like South Africa. Advancement in Data science will impact overall health outcomes

The competencies of a Health Analyst and My growth areas

To succeed as a health analyst, one needs a combination of technical, analytical and soft skills

The core skills and competencies include Data analysis, statistical Knowledge, health informatics, communication and reporting and technical skills

Data analysis refers to the ability to manipulate, visualize and Explore data. We can manipulate large data sets by making use of tools such as Python (Pandas, NumPy libraries') R or SQL, to visualize the data tools like Tableau, Power BI or python libraries such as matplotlib and seaborn are used and thereafter through Exploratory data analysis patterns and trends are uncovered to gain insights into the health system

Statistical Knowledge makes use of Descriptive, inferential and survival analysis

Descriptive statistics, such as mean, median, variance, and correlation, provide understanding of data distributions and patterns, allowing analysts to summarize and interpret data, uncovering hidden trends and patterns.

Inferential statistics such as hypothesis testing, regression analysis, and p-values enable analysts to obtain actionable insights that informs decisions

Survival analysis is a critical component in healthcare analytics, particularly for studies focused on patient outcomes and treatment effectiveness. By examining time-to-event data, meaningful

patterns and trends are obtained and help to inform treatment strategies and improving patient care.

Health informatics allows for the working and understanding of EHR and workflows

The Technical skills that health analysts use is Programming with Python, R and SQL being the most common, Data storage and management and the know-how on health data standards

Beyond the technical, Analysts need to be able to communicate findings effectively through compelling storytelling and detailed reports and thus communication is a key competency in health data analytics

Growth areas

Key areas for growth that align with the evolving needs of modern health systems

In advanced statistical Methods, using Bayesian statistics allows to make predictions in situations of uncertainty, which is often the case in clinical settings. Time series analysis is essential for studying trends in patient vitals, monitoring disease progression, or managing healthcare resources over time. Longitudinal data analysis, on the other hand, focuses on understanding how patient health evolves, particularly in managing chronic diseases.

Machine learning and artificial intelligence are transforming healthcare the use of Predictive models to forecast outcomes such as patient readmissions, treatment responses, or even disease outbreaks. Deep learning offers powerful tools for analysing medical imaging and extracting meaning from complex datasets, while natural language processing (NLP) can unlock insights hidden in unstructured data like physician notes or patient feedback. These technologies not only enable more precise and timely interventions but also pave the way for personalized and efficient healthcare delivery.

I plan to focus on data engineering skills, starting with mastering ETL (extract, transform, load) processes, which ensure that raw data is cleaned, organized, and ready for use. . Learning to work with cloud computing platforms like AWS, Azure, or Google Cloud will enable me to analyse and store large datasets securely

Tech stacks are a choice of tools and technologies used to address varied analytical requirements, these depend on the specific goals, data characteristics, and organizational needs.

For programming, Python and R stand out due to their vast ecosystems of libraries like Pandas and NumPy for Python, or dplyr and ggplot2 for R. To simplify collaboration and documentation, environments like Jupyter Notebooks and R Markdown are used , allowing one to code, narrate and visualise insights .

because of the unique demands of the healthcare industry dealing with sensitive patient data, large-scale datasets, and complex reporting requirements. stacks used in health analytics must balance precision, scalability, and compliance.

In Health Sciences, a wide range of **tools** are commonly, These tools are used to process, analyse , visualise and document data .

To collect data, Electronic Health Records containing patient data, such as patients' demographics, medical history and lap results are used to improve patient outcomes, reduce cost and inform public health. Surveys, Questionnaires, and digital platforms can also be used to gather patient reported outcomes and real-world evidence.

To process the data, In Most cases with health data, Python is used for data science and machine learning with libraries such as Pandas , SciPy and Scikit learn aiding in analysing the data. R and R studios is used for statistical computing and advanced data modelling

To visualise and predict the data, Machine learning and AI Tools such as TensorFlow and Pytorch are used to predict patient outcomes and optimize clinical Trail. Visual Tools like Matplotlib and seaborn are used for graphical exploration and insights

To Document, Jupiter Notebook an open-source python based tool is used for documenting and running code, Markdowns are used to format text, make notes, and document code in a readable format, often used in Jupyter Notebooks.

For Cloud storage, AWS, Azure, and Google Cloud offer robust environments for secure, scalable, and cost-effective health data analytics and Microsoft Teams and Slack can be used to Collaborate and communicate

GitHub and GitLab are code repositories for version control and collaboration, enabling teams to work together on data analysis and share code.

The **advancements** in healthcare analytics are transforming patient resulting in improved health outcomes. The use of Machine learning and predictive modelling allow for prediction of health outcomes which enable more informed decisions.

There are several data sources in Health, with each providing unique insights to improve patient care and outcomes.

Some of these data sources include clinical data from EHRs, lab tests, and medical devices used for diagnosis and treatment. Administrative data, like claims and hospital records which are helpful in proving insights for operational decision-making. Public health data found from surveillance systems and Registries provide insights on population trends to help address health system challenges. Patients can also provide data wearables and mobile health apps

The analysis of EHRs is beneficial and can assist in the improvement of patient outcomes; by analysing patterns in patient records, healthcare providers can identify risk factors, track treatment effectiveness, and customize care to individual needs.

The reduction of Healthcare Costs by Identifying trends in hospital readmissions, preventable conditions, and high-cost treatments that can help streamline processes and reduce unnecessary expenses.

EHR data can reveal public health trends, predict outbreaks, and support preventive health initiatives.

To demonstrate this Learning an Analysis of a Electronic Healthcare dataset was done and insights were gathered thereafter

The analysis included the loading and cleaning to exploratory data analysis (EDA), dimensionality reduction, and clustering.

Variable Description

The dataset has 42 columns and 4,084 entries, capturing various demographic, medical, and healthcare usage information.

Demographic data – Age, Race, sex

The Hierarchical Condition Category (HCC) score- patient's overall health risk

Average Length of Stay (LOS)- hospital stay duration

Healthcare resource utilization – visits to the ER, hospital outpatient visits, skilled nursing facility (SNF) use, and home health services, Enrolment Months, Total Claims and Labs

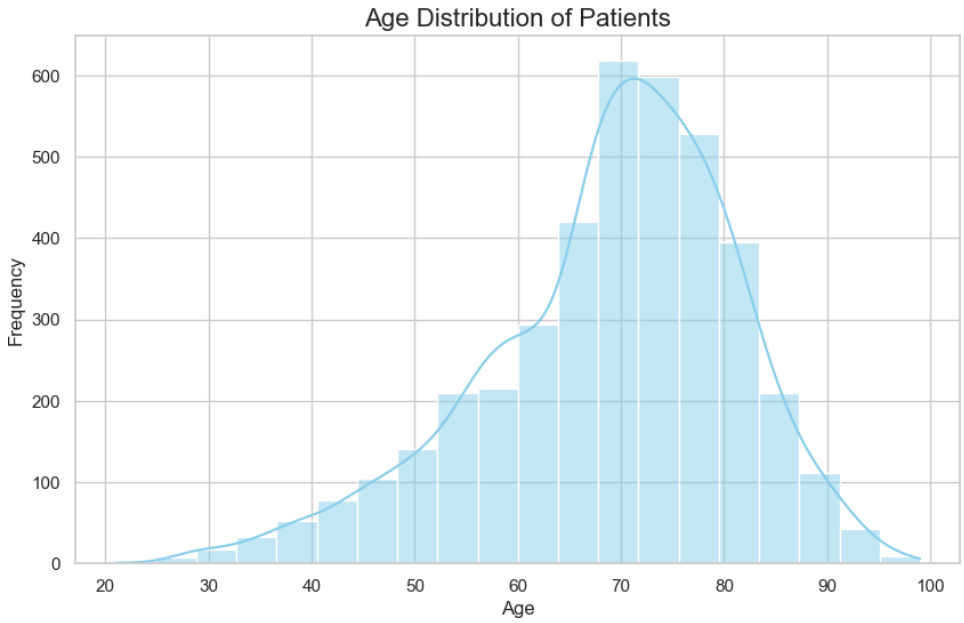
Used libraries like Pandas for data manipulation and NumPy for numerical computations to analyse health data and used *.info() and .describe()* to provides insights into data types, missing values, and statistical distributions.

To **Evaluate the quality of the Data and Prepare data for analysis**

I focused on completeness of the by identifying missing values, I then Imputed the missing values using Imputation with the median for numerical columns in the attempt to reduce the influence of outliers. I made use of the SimpleImputer with the median from scikit-learn.

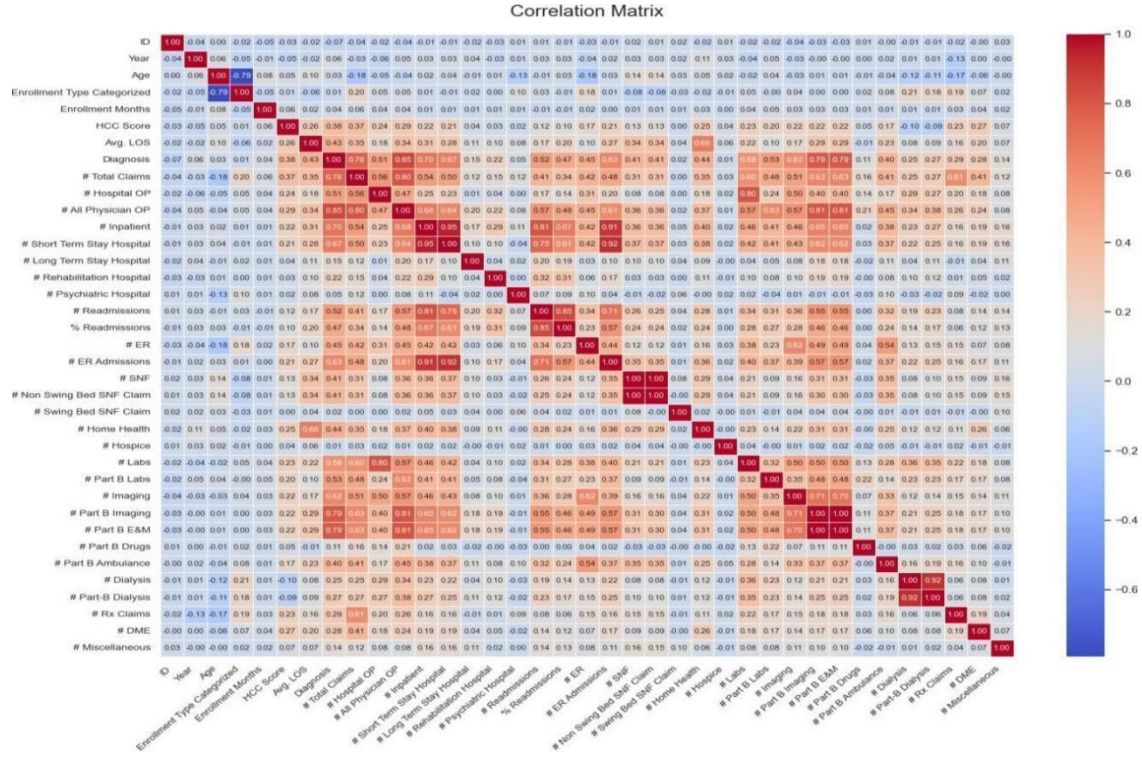
Data Visualization with Matplotlib, Matplotlib aids in presenting data trends and patterns effectively in healthcare.

Histograms was used to examine age distributions; the age distribution is right skewed with most patients being elderly

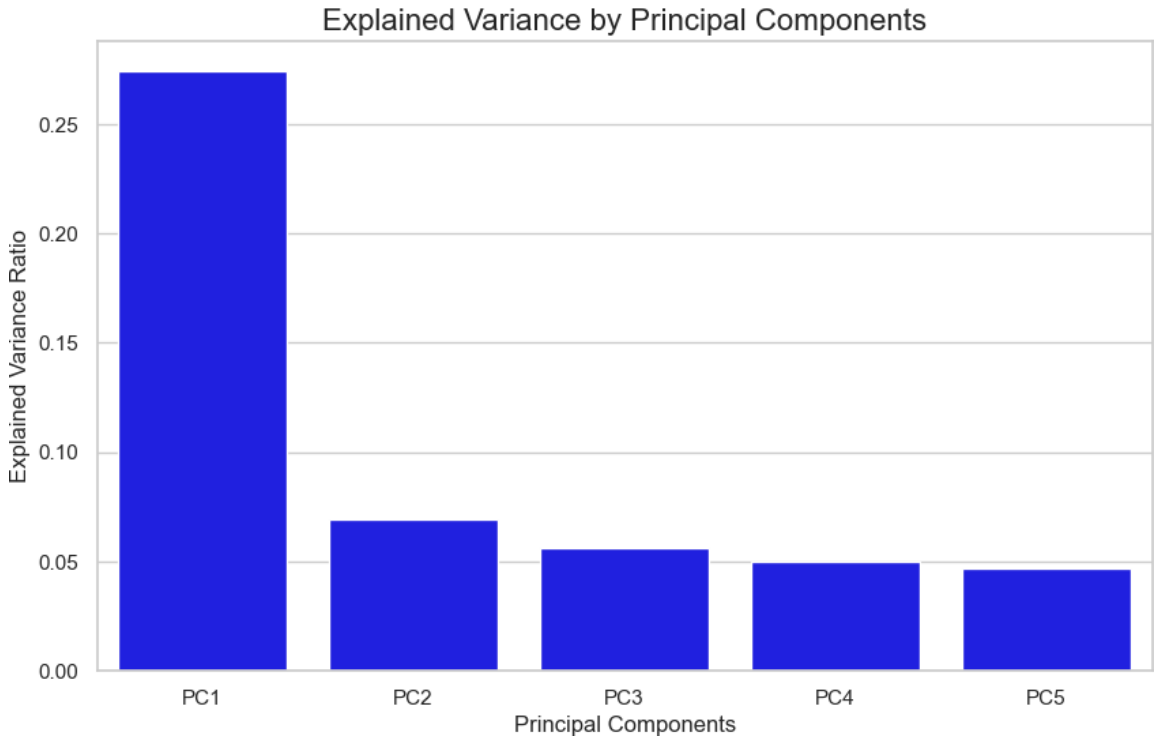


A Scatter plots was used to investigate the relationships between the variables (HCC scores and claims) which informs the utilisation of healthcare, Patients with higher scores have higher healthcare claims

Heatmaps to visualize correlations among numerical variables,



Variables showed strong correlations which could mean that the information is repetitive, Thus advocating for Dimensional reduction,



The variance showed that the data could be summarised using fewer dimension in healthcare records, reducing 43 variables to 2-3 principal components can simplify analysis without losing key patterns and thus the application of PCA from scikit-learn, the data was standardized using a StandardScaler

Supervised Learning Involves labelled data, where the target variable is known, and models learn to predict the target based on input features whereas Unsupervised Learning Involves unlabelled data where the target variable is unknown models predict hidden patterns such as with clustering and PCA

K clustering showed that the optimal number of clusters is 3 meaning that the data could be divided into 3 clusters this was calculated using the Elbow Method, the visualisation of the clusters showed well separated clusters in PCA space

Working with electronic healthcare records (EHRs) comes with the responsibility to protect sensitive patient information. The POPI act Further stipulates this. Data breaches can have serious consequences for the individuals and organizations. With increased use of technology to enhance healthcare it is important to consider and set in data security and safeguarding Practices

References

Nkwanyana A, Mathews V, Zachary I, et al. Skills and competencies in health data analytics for health professionals: a scoping review protocol. BMJ Open 2023 (Skills and Competencies

Munusamy, Ambigavathi & Sridharan, D.. (2018). Big Data Analytics in Healthcare. 269-276. 10.1109/ICoAC44903.2018.8939061.(data Analytics and tools)

Katurura MC, Cilliers L. Electronic health record system in the public health care sector of South Africa: A systematic literature review. Afr J Prim Health Care Fam Med. 2018 Nov 20;10(1):e1-e8. doi: 10.4102/phcfm.v10i1.1746. PMID: 30456963; PMCID: PMC6295973.

Lecture note : Health Analytics 202