Medical cost personal data set in market segmentation
Name:Sushmitha.Akula

Problem Statement

Everyone’s life revolves around their health. Good health is essential to all aspects of our lives. Health refers to a person’s ability to cope up with the environment on a physical, emotional, mental, and social level. Because of the quick speed of our lives, we are adopting many habits that are harming our health. One spends a lot of money to be healthy by participating in physical activities or having frequent health check-ups to avoid being unfit and get rid of health disorders. When we become ill we tend to spend a lot of money, resulting in a lot of medical expenses .So, an application can be made which can make people understand the factors which are making them unfit, and creating a lot of medical expenses, and it could identify and estimate medical expense if someone has such factors.

Objective :Predict the future medical expenses of subjects based on certain features building a robust machine learning model. Identifying the factors affecting the medical expenses of the subjects based on the model output.

Abstruct: For this project, the data has been imported from the machine learning repository. The dataset contains 1338 rows and 7 columns. The columns present in the dataset are ‘age’,’ sex’,’bmi’, ’children’, smoker’, ’region’, and ‘charges’. The charges column is the target column and the rest others are independent columns. Independent columns are those which will predict the outcome.The first column is Age. Age is an important factor for predicting medical expenses because young people are generally more healthy than old ones and the medical expenses for Young People will be quite less as compared to old people. the Next column is sex, which has two Categories in this column: Male and Female. The sex of the person can also play a vital role in predicting the medical expenses of a subject .After that, you have the ‘bmi’ column, then **BMI is Body Mass Index**For most adults, an ideal BMI is in the 18.5 to 24.9 range.For children and young people aged 2 to 18, the BMI calculate

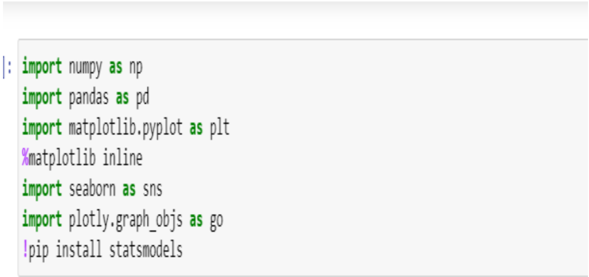
For children and young people aged 2 to 18, the BMI calculation takes into account age and gender as well as height and weight. If your BMI is less than 18.5, you are considered underweight. People with very low or very high ‘bmi’ are more likely to require medical assistance, resulting in higher costs.

The fourth column is the ‘children’ column, which contains information on how many children your patients have. Persons who have children are under more pressure because of their children’s education, and other needs than people who do not have children. The fifth is the ‘smoker’ column. The Smoking factor is also considered to be one of the Most Important factors as the people who smoke are always at risk when their age reaches 50 to 60.

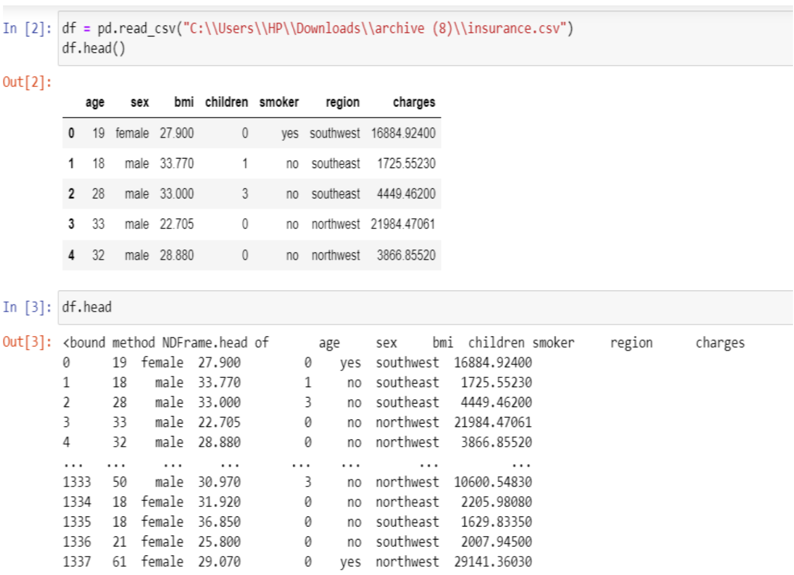
Next is the ‘region’ column. Some Regions are Hygienic, Clean, Neat, and Prosperous, But some Regions are not, and this information affects health which is related to medical expenses.

Coad implementation:

Data preprocessing



```
]: import numpy as np
   import pandas as pd
   import matplotlib.pyplot as plt
   %matplotlib inline
   import seaborn as sns
   import plotly.graph_objs as go
   !pip install statsmodels
```



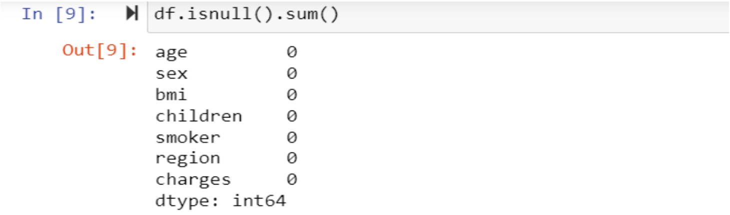
```
In [2]: df = pd.read_csv("C:\\Users\\HP\\Downloads\\archive (8)\\insurance.csv")
        df.head()
```

Out[2]:

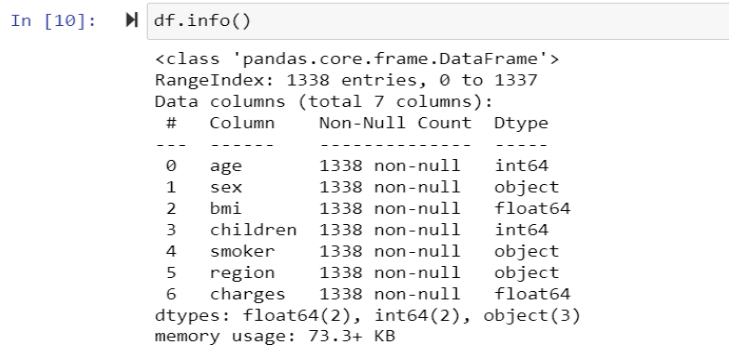
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

```
In [3]: df.head

Out[3]: <bound method NDFrame.head of       age     sex     bmi  children smoker     region      charges
        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
```

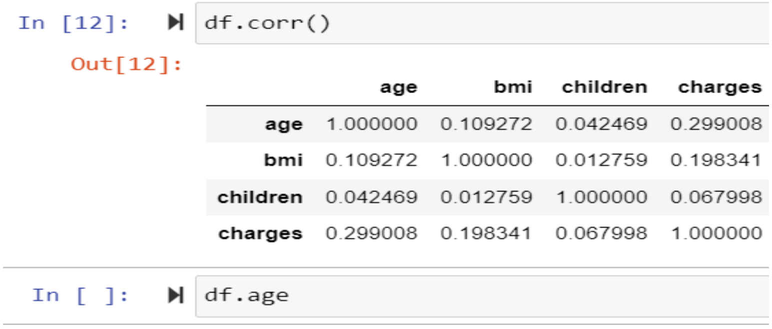
Data cleaning:



```
In [9]:  df.isnull().sum()

Out[9]:  age         0
         sex         0
         bmi         0
         children    0
         smoker      0
         region      0
         charges     0
         dtype: int64
```



```
In [10]:  df.info()

          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 1338 entries, 0 to 1337
          Data columns (total 7 columns):
           #   Column    Non-Null Count  Dtype  
          ---  ------    --------------  -----  
           0   age       1338 non-null   int64  
           1   sex       1338 non-null   object 
           2   bmi       1338 non-null   float64
           3   children  1338 non-null   int64  
           4   smoker    1338 non-null   object 
           5   region    1338 non-null   object 
           6   charges   1338 non-null   float64
          dtypes: float64(2), int64(2), object(3)
          memory usage: 73.3+ KB
```

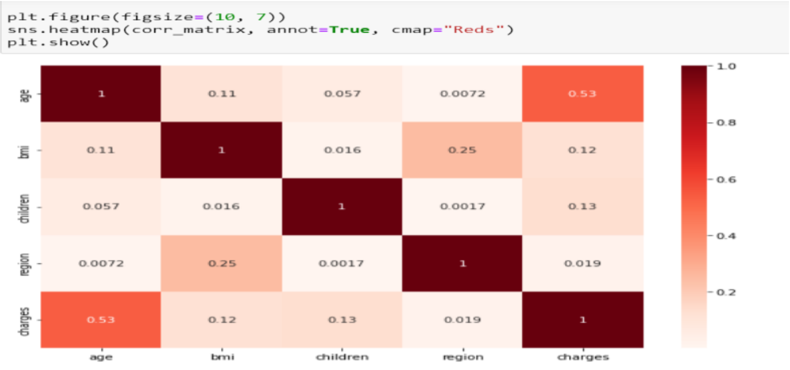


```
In [12]:  df.corr()
Out[12]:
```

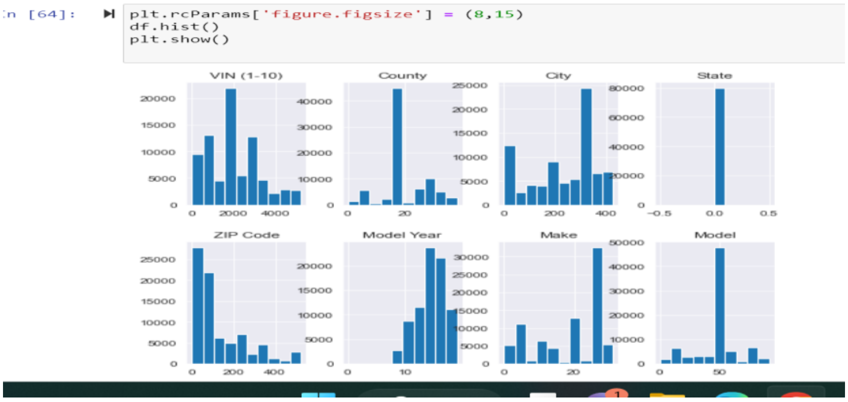
	age	bmi	children	charges
age	1.000000	0.109272	0.042469	0.299008
bmi	0.109272	1.000000	0.012759	0.198341
children	0.042469	0.012759	1.000000	0.067998
charges	0.299008	0.198341	0.067998	1.000000

```
In [ ]:  df.age
```

```
plt.rcParams['figure.figsize'] = (8,15)
df.hist()
plt.show()
```

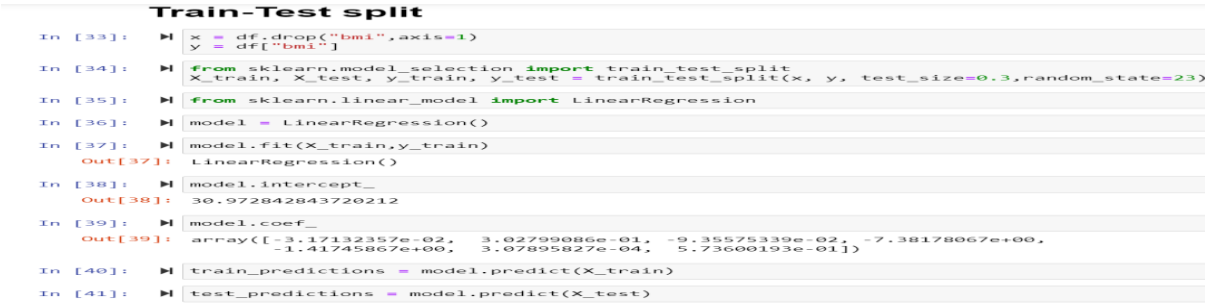


```
plt.figure(figsize=(10, 7))
sns.heatmap(corr_matrix, annot=True, cmap="Reds")
plt.show()
```



```
In [64]:  plt.rcParams['figure.figsize'] = (8,15)
          df.hist()
          plt.show()
```

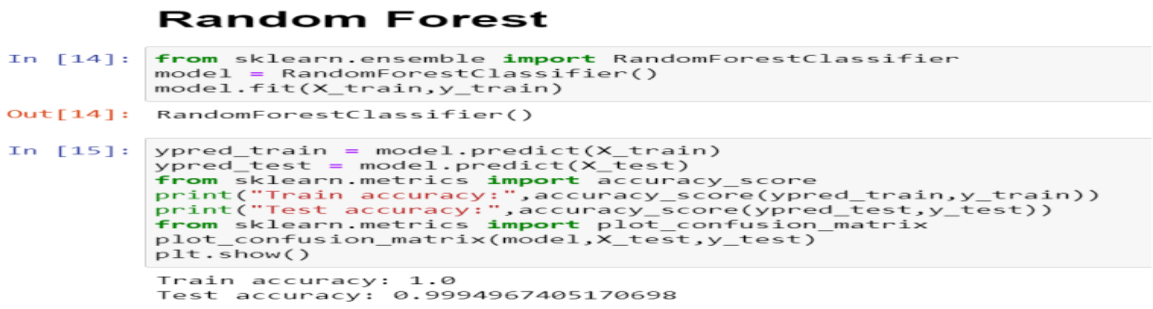
Linear regression:



Train-Test split

```
In [33]:  x = df.drop("bmi",axis=1)
          y = df["bmi"]

In [34]:  from sklearn.model_selection import train_test_split
          X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.3,random_state=23)

In [35]:  from sklearn.linear_model import LinearRegression

In [36]:  model = LinearRegression()

In [37]:  model.fit(X_train,y_train)
Out[37]:  LinearRegression()

In [38]:  model.intercept_
Out[38]:  30.972842843720212

In [39]:  model.coef_
Out[39]:  array([-3.17132357e-02,  3.02799086e-01, -9.35575339e-02, -7.38178067e+00,
                 -1.41745867e+00,  3.07895827e-04,  5.73600193e-01])

In [40]:  train_predictions = model.predict(X_train)

In [41]:  test_predictions = model.predict(X_test)
```

Segment extraction:



Random Forest

```
In [14]:  from sklearn.ensemble import RandomForestClassifier
          model = RandomForestClassifier()
          model.fit(X_train,y_train)
Out[14]:  RandomForestClassifier()

In [15]:  ypred_train = model.predict(X_train)
          ypred_test = model.predict(X_test)
          from sklearn.metrics import accuracy_score
          print("Train accuracy:",accuracy_score(ypred_train,y_train))
          print("Test accuracy:",accuracy_score(ypred_test,y_test))
          from sklearn.metrics import plot_confusion_matrix
          plot_confusion_matrix(model,X_test,y_test)
          plt.show()

          Train accuracy: 1.0
          Test accuracy: 0.9994967405170698
```

AI MARKET SHARE IN INDIA 2021 BY INDUSTRY PUBLISHED SHANGLIOSUN MARCH 2022

The AI market share of the IT services industry in India reached 51.8 percent in 2021. Artificial intelligence has been responsible for drastic changes in the technology sector where it can greatly improve productivity through process simplification and automation. It is also an integral part and one of the fundamental bases of Industry 4.0. In several developed countries, AI could potentially maximize labor productivity by more than 30 percent in the next 15 years.

AI application in India : As India is a country with huge linguistic diversity, it imposes a great challenge to governments and companies when conducting business with people of different linguistic backgrounds. As a result, one of the first applications for AI in India is in the field of customer service. The Indian government has increased public investment to promote the Digital India initiative in the fields of AI, IoT, big data, machine learning, and robotics.

Challenges of AI adoption in India: However, there are several obstacles India faces in the process of AI adoption. India has a comparatively small number of scientists and researchers in the field of machine learning and artificial intelligence. It also lacks sufficient qualified specialists to localize and implement the latest technologies in the field. However, the Ministry of Electronics and Information Technology, along with various industrial bodies have introduced several programs of personnel training and technical infrastructure building to lay the foundation for future AI development in India.

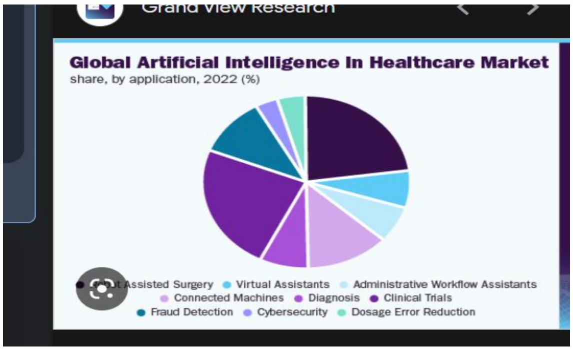
Market insights:

It is predicted that the applications of artificial intelligence in the healthcare space will be worth INR ~431.97 Bn by 2021, expanding at a rate of ~40%. Based on this growth of AI applications in healthcare, the doctor-patient ratio in India is expected to reach ~6.9:1,000 by 2023, from its 2017 ratio of ~4.8:1000.
The adoption of artificial intelligence (AI) is reshaping the Indian healthcare market significantly. AI-enabled healthcare services like automated analysis of medical tests, predictive healthcare diagnosis, automation of healthcare diagnosis with the help of monitoring equipment, and wearable sensor-based medical devices, are expected to revolutionize medical treatment processes in the country. The capability of AI applications to improve doctors efficiency will help in tackling challenges like uneven doctor-patient ratio, by providing rural populations high-quality healthcare, and training doctors and nurses to handle complex medical procedures.

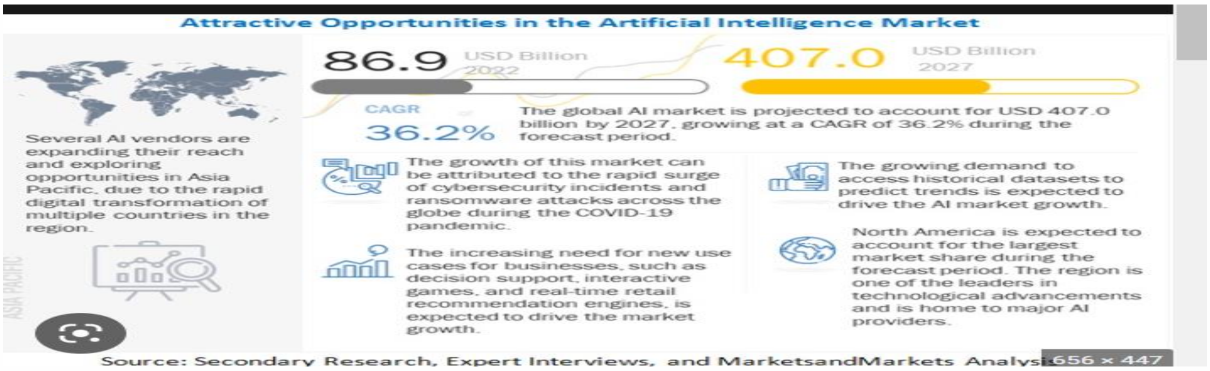
Artificial intelligence is used in six healthcare segments: hospitals, pharmaceuticals, diagnostics, medical equipment and supplies, medical insurance, and tele-medicine.

Benefits of AI in the healthcare sector: The patient doctor ratio in India is as low as 1,700:1. Also, ~70% of the healthcare infrastructure is in cities, which cater to ~30% of the country's population. With the use of artificial intelligence applications, doctors can offer their services to more patients and reduce the existing gap in demand and supply of medical services in the country.AI-enabled healthcare services can be delivered at lower costs with increased efficiency and an emphasis on diagnostics. Moreover, artificial intelligence enables hospitals to implement patient centric plans and eliminate unnecessary hospital procedures, making delivery of healthcare services faster in India.

Key deterrents to the growth of the market: India lacks standardized guidelines for designing AI applications that can be used in healthcare systems. Lack of clarity deters the use of artificial intelligence in the Indian healthcare industry. Also, most AI companies which aid the healthcare sector, are startups. Medical practitioners are not quick to trust startups whose products are not nationally or internationally certified. As a result, start-ups sales get hampered, resulting in the limited implementation of AI in the Indian healthcare industry.



AI MEDICAL COST MARKET SHARE:



Attractive Opportunities in the Artificial Intelligence Market

86.9 USD Billion 2022

407.0 USD Billion 2027

CAGR 36.2%

The global AI market is projected to account for USD 407.0 billion by 2027, growing at a CAGR of 36.2% during the forecast period.

Several AI vendors are expanding their reach and exploring opportunities in Asia Pacific, due to the rapid digital transformation of multiple countries in the region.

The growth of this market can be attributed to the rapid surge of cybersecurity incidents and ransomware attacks across the globe during the COVID-19 pandemic.

The increasing need for new use cases for businesses, such as decision support, interactive games, and real-time retail recommendation engines, is expected to drive the market growth.

The growing demand to access historical datasets to predict trends is expected to drive the AI market growth.

North America is expected to account for the largest market share during the forecast period. The region is one of the leaders in technological advancements and is home to major AI providers.

Source: Secondary Research, Expert Interviews, and MarketsandMarkets Analysis

Opportunities in medical cost persnols:

To develop the best medical insurance products, the insurer need access to historical data approximate the medical costs of each user. With this data, a medical insurer can develop more accurate pricing models, plan a particular insurance outcome, or manage a big portfolios. For all these cases, the objective is to accurately predict insurance costs.This dataset contains 1,339 medical insurance records. The individual medical costs billed by health insurance are the target variable *charges*, and the rest of columns contain personal information such as age, gender, family status, and whether the patient smokes among other features.

Use case: The objective to train and testsplit ML regression model that generates the target column *charges* more accurately. Being a regression model problem, metrics such as the coefficient of determination and the mean squared error are used to evaluate the model.

Privacy This dataset contains personal information about users, making it difficult to work and share this dataset. In Synthesized we can generate a synthetic dataset that preserves statistical information (95% utility across multiple ML tasks compared to original data) in under 10 minutes, while removing all risk of non-compliance with data regulation such as GDPR, HIPAA and CCPA.

Fairness and Biases AI models can be unintentionally (and potentially illegal) discriminative to certain sensitive groups of people, if the underlying training data is biased. This dataset is especially sensitive, as it contains users medical records. Synthesized can help assessing how biased a dataset is, finding where the biases are and flagging them to the user. Read more about discrimination by AI in our blog post.

Global AI Medical Cost persnols:

While healthcare organizations and startups were looking to innovate with AI before the pandemic, they are now doing so more than ever. It's clear that the technology has the potential to revolutionize the industry in at least several areas, such as diagnostics, treatment protocols, and clinical research.

The cost of AI in healthcare, especially when it comes to bespoke solutions, is driven by several factors and needs investigation on a case-by-case basis. In healthcare, custom AI solutions can ensure that specific market problems are addressed, and firms only pay for what they need instead of expensive off-the-shelf products that are not fit for purpose. This article discusses how much AI costs in healthcare and why companies can benefit from a bespoke solution. According the cost of a complete custom AI solution can vary from US$20,000 to US$1,000,000. A minimum viable product (MVP) sets you back between US$8,000 and US$15,000. It's a common misconception that AI costs a fortune and is only for the tech giants like Google, Facebook, or Microsoft. Improving computer power, connectivity, and algorithms have made it affordable to all organizations in the last decade. The variation in costs results from the level of intelligence required, the amount of data applications will consume, and how the algorithms need to perform. As well as the technology itself, various other considerations feed into the cost of implementing artificial intelligence in healthcare. The cost of AI in healthcare, especially when it comes to bespoke solutions, is driven by several factors and needs investigation on a case-by-case basis.

The growth potential of AI in medicalcare in India

AI expenditure in India increased by over 109% in 2018, totaling $665 million and is exp ected to reach $11.78 billion by 2025 , adding $1 trillion to India's economy by 2035.

NITI Aayog, a public policy think-tank linked to the Indian government, has been testing the application of AI in primary care for early detection of diabetes complications, and is currently validating the use of AI as a screening tool in eye care, by comparing its diagnostic accuracy with that of retina specialists. Integrating AI capabilities with portable screening devices, such as 3Nethra, can expand the capacity for eye screenings and early detection, and enable access in remote places aross the country.

Similar applications are possible in oncology. Tata Medical Center and the Indian Institute of Technology recently launched India’s first de-identified cancer image bank: the Comprehensive Archive of Imaging. AI-based tools can use high-quality de-identified images to enable machine learning models to detect biomarkers and improve outcomes for cancer research.

In cardiovascular healthcare, a major and somewhat unique challenge for India, Microsoft’s AI Network for Healthcare and Apollo Hospitals are developing a machine learning model to better predict heart attack risk. Using clinical and lab data from over 400,000 patients, the AI solution can identify new risk factors and provide a heart risk score to patients without a detailed health check-up, enabling early disease detection.

India's way forward

AI maturity in health requires critical investments in the capacity of the workforce, data and infrastructure, governance and regulatory mechanisms, design and processes, partnerships and stakeholders as well as innovative business models.

Integrating AI into healthcare systems also requires an understanding of AI in nationalcurricula for medical and public health students, both academic and practical.

Similarly, the Indian government will need to make appropriate investments in data infrastructure, such as interoperability, unified EMR and data stewardship. This is essential to build trust and long-term integration of AI into India’s healthcare system.

The government must also invest in and build public-private partnerships across the healthcare industry to facilitate coordination between academia, government, industry, NGOs and patient advocacy organizations. They should scale governance and regulatory mechanisms to provide appropriate oversight for privacy, fairness and transparency.

NITI Aayog’s National Strategy for AI prioritizes principles of privacy, ethics, security, fairness, transparency and accountability, as well as alignment with the rights afforded by the Indian Constitution. India is a founding member of the Global Partnership on AI alliance and has thus far adopted a measured approach for integration of AI, in keeping with ethical and responsible standards. These principles must be applied in practice as the technology scales.

The way in which AI systems are integrated, too, will be crucial. Human-in-the-loop and

human-centric designs that empower healthcare staff to understand how a decision is made and
how to incorporate this knowledge into treatment will minimize risk.
Investments in expanding the healthcare workforce and data literacy will build an informed
workforce capable of leveraging AI in healthcare. India's measured adoption of this technology
can enable it to bridge rural-urban disparities without leaving anyone behind, while becoming a
leader among other emerging markets on the road to meeting the Sustainable Development
Goals.

Summary

Extracted segments and profiling of potential segments: Here is a brief description and profiling of each of the segments mentioned earlier in the AI in medical diagnosis market in India.Overall, these segments provide insights into the potential markets and opportunities for growth and investment in AI in medical medical cost in India.:

2 Technology:

Machine Learning: Machine learning algorithms can analyze large amounts of medical data to identify patterns and predict treatment responses, enabling personalized
treatment plans.

Deep Learning: Deep learning neural networks can analyze complex medical data such as
brain images to identify patterns and provide insights into disease mechanisms and treatment responses.

Natural Language Processing (NLP): NLP can help analyze medical text and patient data to identify patterns and insights, enabling faster and more accurate diagnosis and treatment planning.

Others: Other AI technologies such as computer vision and robotics can help improve the accuracy and efficiency of medical cost persons and treatment.

3 End-User:

Hospitals: Hospitals are the primary end-users of AI in medical cost ,persnols as they use AI
to improve the accuracy and speed of medical cost persnolsnals.

Region: AI models can be unintentionally (and potentially illegal) discriminative to certain sensitive groups of people, if the underlying training data is biased. This dataset is especially sensitive, as it contains users medical records. Synthesized can help assessing how biased a dataset is, finding where the biases are and flagging them to the user. Read more about discrimination by AI in our blog post.

age: age of primary beneficiary

sex: insurance contractor gender, female, male

bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height,
objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9

children: Number of children covered by health insurance / Number of dependents

smoker: Smoking

region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

charges: Individual medical costs billed by health insurance

Customizing the Marketing Mix

The marketing mix is a critical tool for businesses to plan and execute their marketing strategies
effectively. It consists of four major components - product, price, promotion, and place
(distribution). These components must be designed to work together seamlessly to achieve a
business's marketing objectives. In this task, we will explore how the given dataset can be used to
customize the marketing mix for a pharmaceutical company.

Product

The first component of the marketing mix is the product. For a pharmaceutical company, the
product is the drug or medicine they manufacture. The dataset provided can be used to customize
the product in several ways. For instance, the company can focus on manufacturing drugs that
target the most common symptoms and conditions listed in the dataset. For example, drugs that
relieve symptoms of itching, skin rash, joint pain, stomach pain, acidity, vomiting, fatigue, cough,
high fever, headache, abdominal pain, and diarrhea. The company can also focus on manufacturing
drugs for conditions that are prevalent in the dataset, such as fungal infections, urinary tract
infections, and liver failure.

For business marketing:

In the world business one of the key factors determiune the success in ability to identity and capitalize of early market opportunities for business operataing in the business to business this   means identitfy potential customer bases and calculating potential in the early market
To begin with, it is important to define what we mean by the "early market". The early market
refers to the initial stage of a new product or service launch, where the product or service is still
relatively unknown and untested in the market. During this stage, the customer base is likely to be

smaller and more niche, consisting of early adopters and innovators who are willing to take a risk on a new product or service. Identifying the potential customer base in the early market is a crucial step in determining the potential for sales and profits. This involves researching the market and identifying the key players, such as industry leaders, competitors, and potential customers. It also involves analyzing market trends and identifying emerging needs and opportunities.

Once the potential customer base has been identified, the next step is to calculate the potential sale (profit) in the early market. This can be done by multiplying the potential customer base by your target price range. For example, if your target price range is $100-$150 per unit and your potential customer base is 500, your potential profit in the early market would be $50,000-$75,000.

Of course, it is important to keep in mind that these calculations are only estimates and there arr many factors that can impact sales and profits in the early market. For example, the level of competition, the quality and features of the product or service, and the effectiveness of marketing and sales strategies can all play a role in determining success.

It is also important to consider the long-term potential of the product or service beyond the early market. While the early market may provide a good starting point for sales and profits, it is important to have a long-term strategy in place to continue to grow and expand the business.

In order to succeed in the early market, businesses must be agile and adaptable, able to respond quickly to changing market conditions and customer needs. This requires a strong focus on innovation, as well as a willingness to take risks and experiment with new ideas.

Overall, identifying potential customer bases and calculating potential profits in the early market is a key step in building a successful B2B business. By doing thorough research and analysis, and by remaining agile and adaptable, businesses can capitalize on early market opportunities and set themselves up for long-term success.

Market research and segmentation

Market research and segmentation are essential tools for businesses looking to understand their
target audience and create effective marketing strategies. In the healthcare industry, understanding
the symptoms and conditions that people experience can help healthcare providers better diagnose

Market research and segmentation Market research and segmentation are essential tools for businesses looking to understand their target audience and create effective marketing strategies. In the healthcare industry, understandingthe symptoms and conditions that people experience can help healthcare providers better diagnose and treat patients. In this passage, we will discuss a wide range of symptoms and conditions,exploring their prevalence and impact on individuals. Itching is a common symptom experienced by many people, and it can be caused by a variety of factors, including dry skin, insect bites, and allergies.

Business Modelling Specialists are required to correctly label the collected data. The dataset has to be regularly updated.Data must be of good quality so that the model predictions will be accurate.Maintenance of the web application has to be done.

KEY PARTNERS:Age Sex Region Smoker Bmi Charges children.

MONETISATION SCHEME: The best suited model for the product would be the subscription-based model. In this model, access to the product will be provided for a fixed amount of time for a fixed amount of fee.

FINANCIAL MODELLING:

For financial modelling, we need to consider the following variables:
Number of north regions and mid-size hospitals which have subscribed to the product: Number of south regions hospitals which have subscribed to the product: Subscription charge for north regions to mid-size hospitals: Rs. 60000 per year per hospital. Subscription cost for large hospitals: Rs. 100000 per year per hospital. Cost of data collection = Rs. D lakh/year. Our team will have 2 ML engineers, 2 full-stack developers and 1 product manager. Cost to company for each ML engineer: 15 lakh per annum.Cost to company for each developer: 15 lakh per annum. Cost to company for the product manager: 20 lakh per annum. Total cost to company for the team: 90 lakh per annum. Therefore, the financial equation for the product will be:
This equation doesn’t take into account the initial one-time costs required to set the product,
for example we have to collect a large amount of data in the beginning and there is a large
cost associated with it, but it is done only once. This financial equation is for calculating thevalues
y=500000*x1t=50000*x2t-d*100000=1000000-900000

Conclusion:

To develop the best medical insurance products, the insurer need access to historical data to approximate the medical costs of each user. With this data, a medical insurer can develop mo*re accurate pricing models, plan a particular insurance outcome, or manage a big portfolios. 5For all these cases, the objective is to accurately predict insurance costs. This dataset contains persona*l information about users, making it difficult to work and share this dataset. In Synthesized we can generate a synthetic dataset that preserves statistical information (95% utility across multiple ML tasks compared to original data) in under 10 minutes, while removing all risk of non-compliance with data regulation such as GDPR, HIPAA and CCPA. AI models can be unintentionally (and potentially illegal) discriminative to certain sensitive groups of people, if the underlying training data is biased. This dataset is especiall.

Github: https://github.com/SushmithaAkula/medical-cost-persnol-data-in-market-sigmentation.git t