Medical Insurance Cost Prediction

Predicting insurance costs using demographics and health data.

Team Details

1. Om Prakash Mahato — 10800222098
2. Rahul Kumar Lal — 10800222100
3. Md Ebad — 10800222120
4. Karan Kumar Nonia — 10800222101

Guided By : Dr. Arnab Chakraborty Sir

Introduction

What is the project about?

Predicting medical insurance costs based on features like age, BMI, smoking status, and region using a machine learning model.

- **Why is this important?**
  - Helps insurance companies set premiums fairly.
  - Enables customers to understand their expected costs.
- **Key tools and technologies:**
- Python, Pandas, Scikit-learn, Matplotlib, and Seaborn.

Dataset Overview

```
data = pd.read_csv(r"C:\Users\karan\OneDrive\Desktop\FSP PROJECT ML\medical_insurance.csv")
data.sample(10)
```

[80] ✓ 0.1s

...

	age	sex	bmi	children	smoker	region	charges
1810	48	male	30.200	2	no	southwest	8968.33000
1904	35	female	31.000	1	no	southwest	5240.76500
1562	38	male	27.835	2	no	northwest	6455.86265
58	53	female	22.880	1	yes	southeast	23244.79020
980	54	male	25.460	1	no	northeast	25517.11363
1666	40	female	28.120	1	yes	northeast	22331.56680
497	45	male	28.700	2	no	southwest	8027.96800
148	53	female	37.430	1	no	northwest	10959.69470
2077	47	male	36.200	1	no	southwest	8068.18500
1791	52	female	38.380	2	no	northeast	11396.90020

Data Distribution

- **Age**

Most between 20-40, with a tail towards older ages.

- BMI

Normally distributed with a slight right skew, with a few outliers.

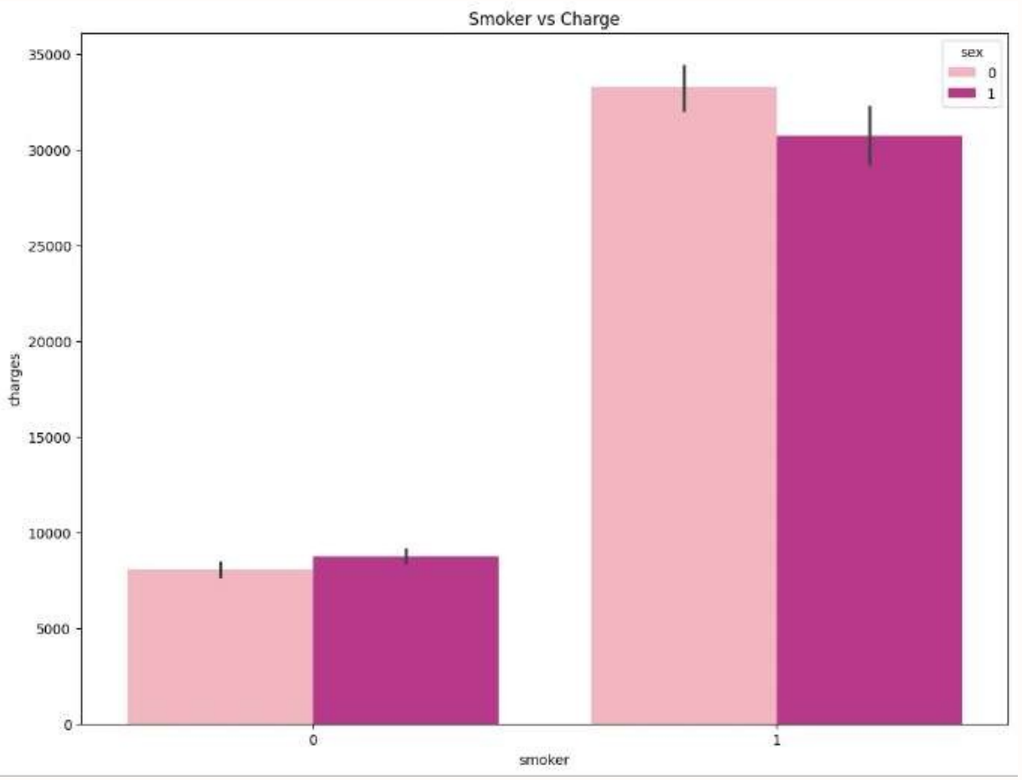
- Charges

Right skewed, indicating a few individuals with high costs.

Feature Relationships

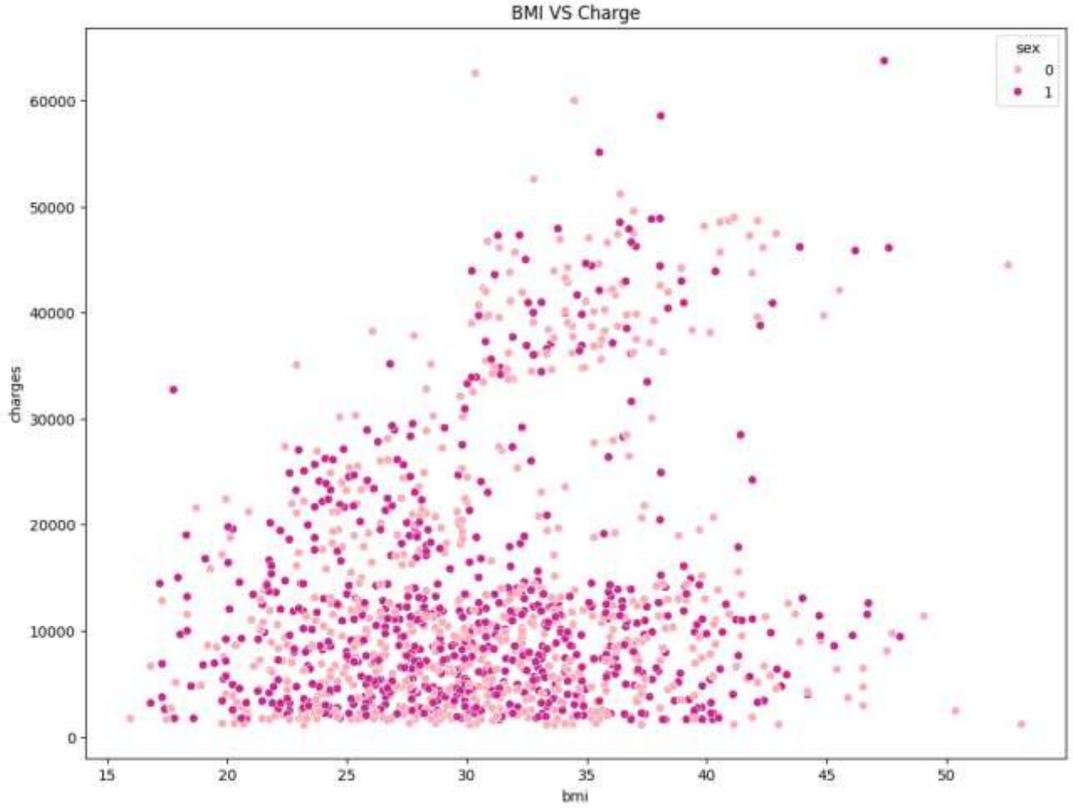
1 1. Charges vs Smoker

Strongest positive correlation, smokers pay significantly more.

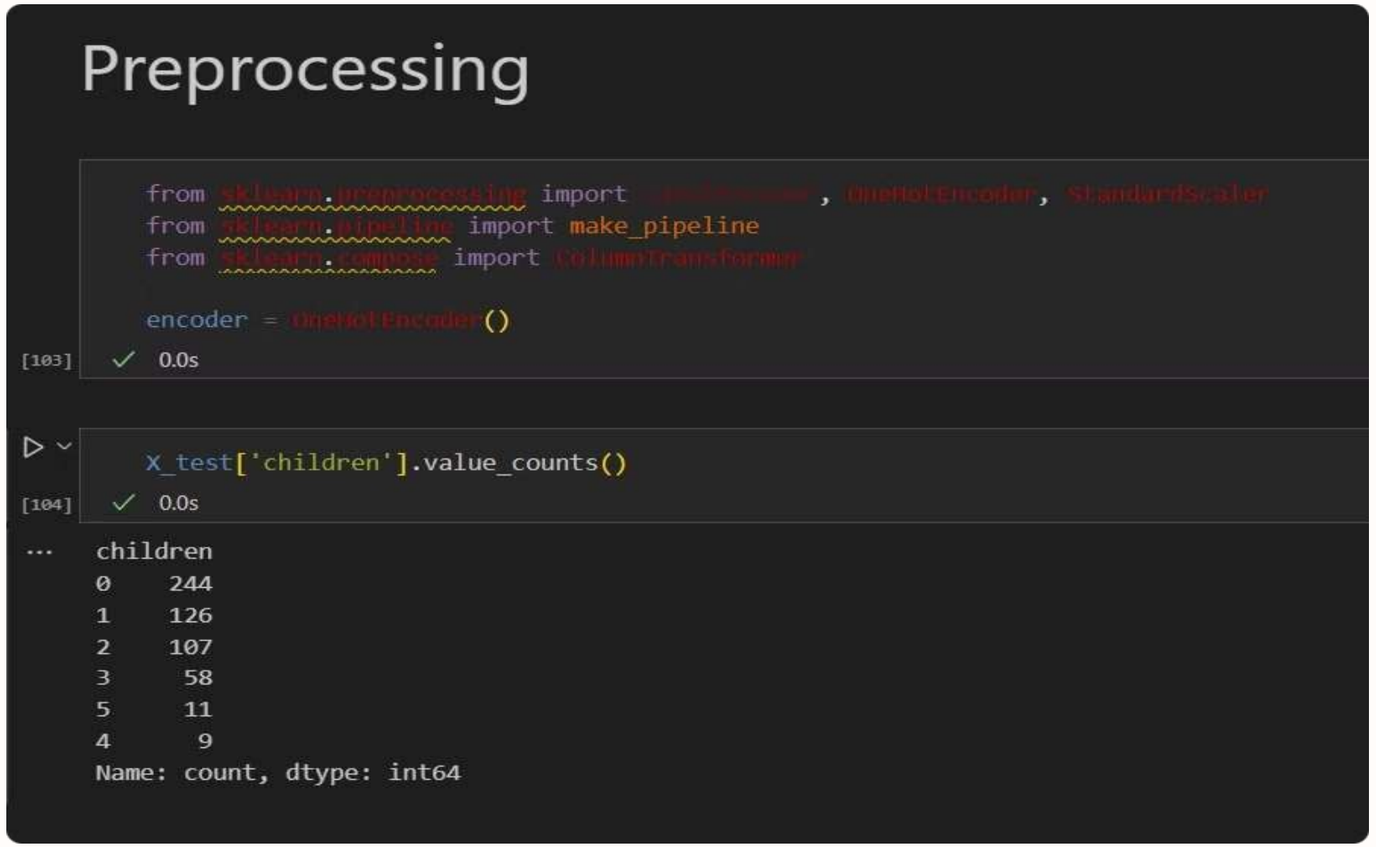


2 2. Charges vs BMI

Moderate positive correlation, higher BMI, higher costs.



Data Preprocessing



Encoding

Convert categorical features to numerical using one-hot encoding.

Splitting

Divide dataset into training (80%) and testing (20%) sets.

Normalization

Scale features to a similar range for model training.

Model Selection

Linear Regression

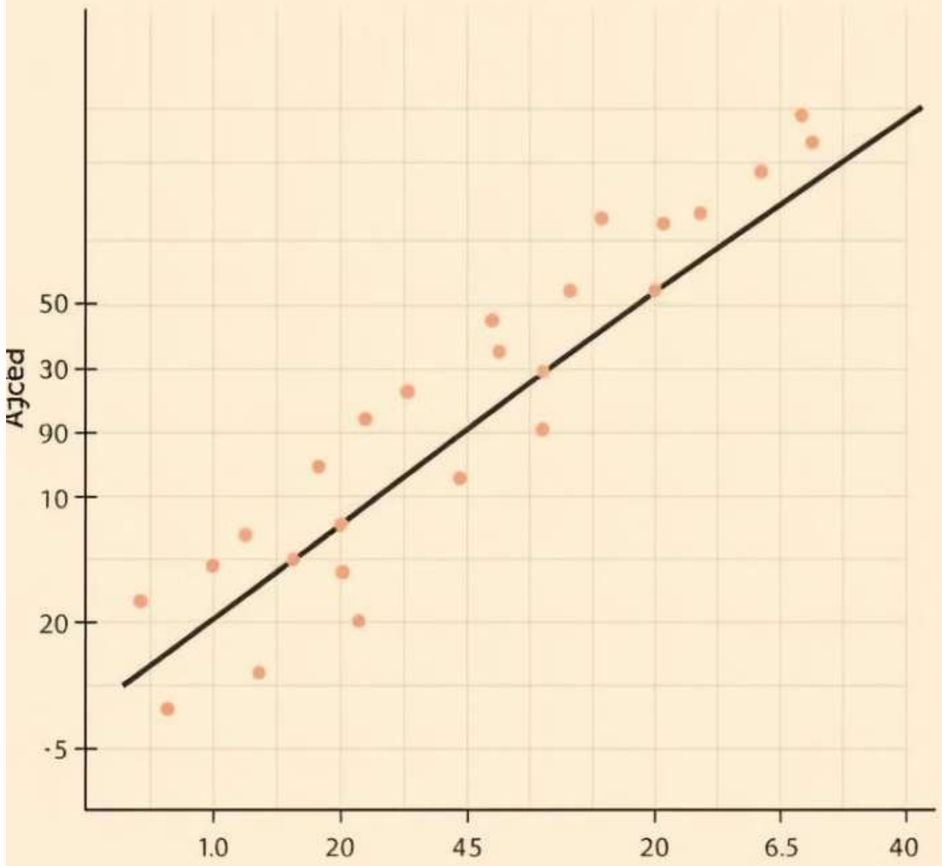
Simple, interpretable, and widely used.

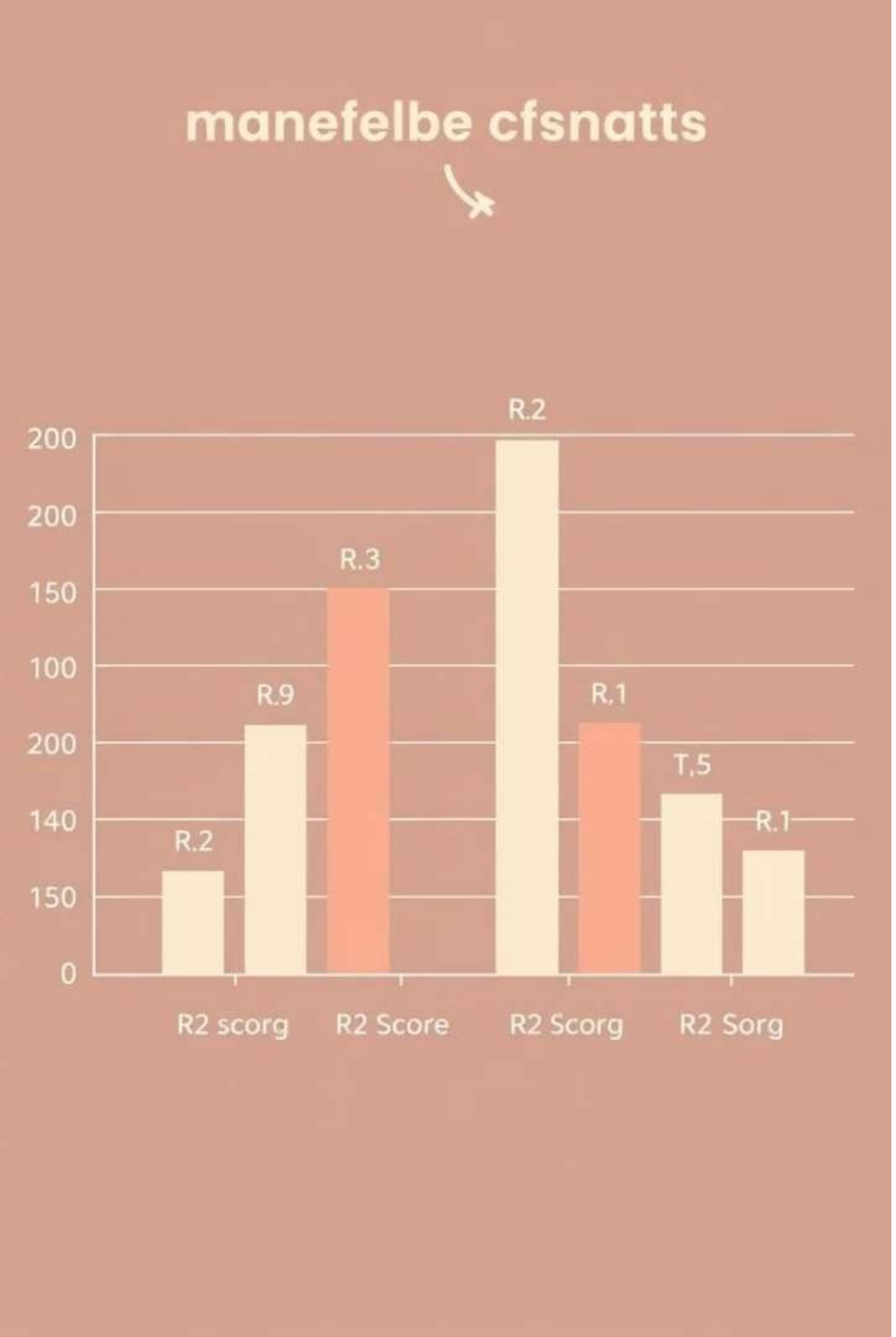
Train/Test Split

80% train, 20% test, to assess model generalization.

R² Score

Metric for evaluating model performance (higher is better).





Model Performance

1. Training R^2: 0.75
2. Testing R^2: 0.74
3. Consistent performance, reliable model.

Prediction Example

1

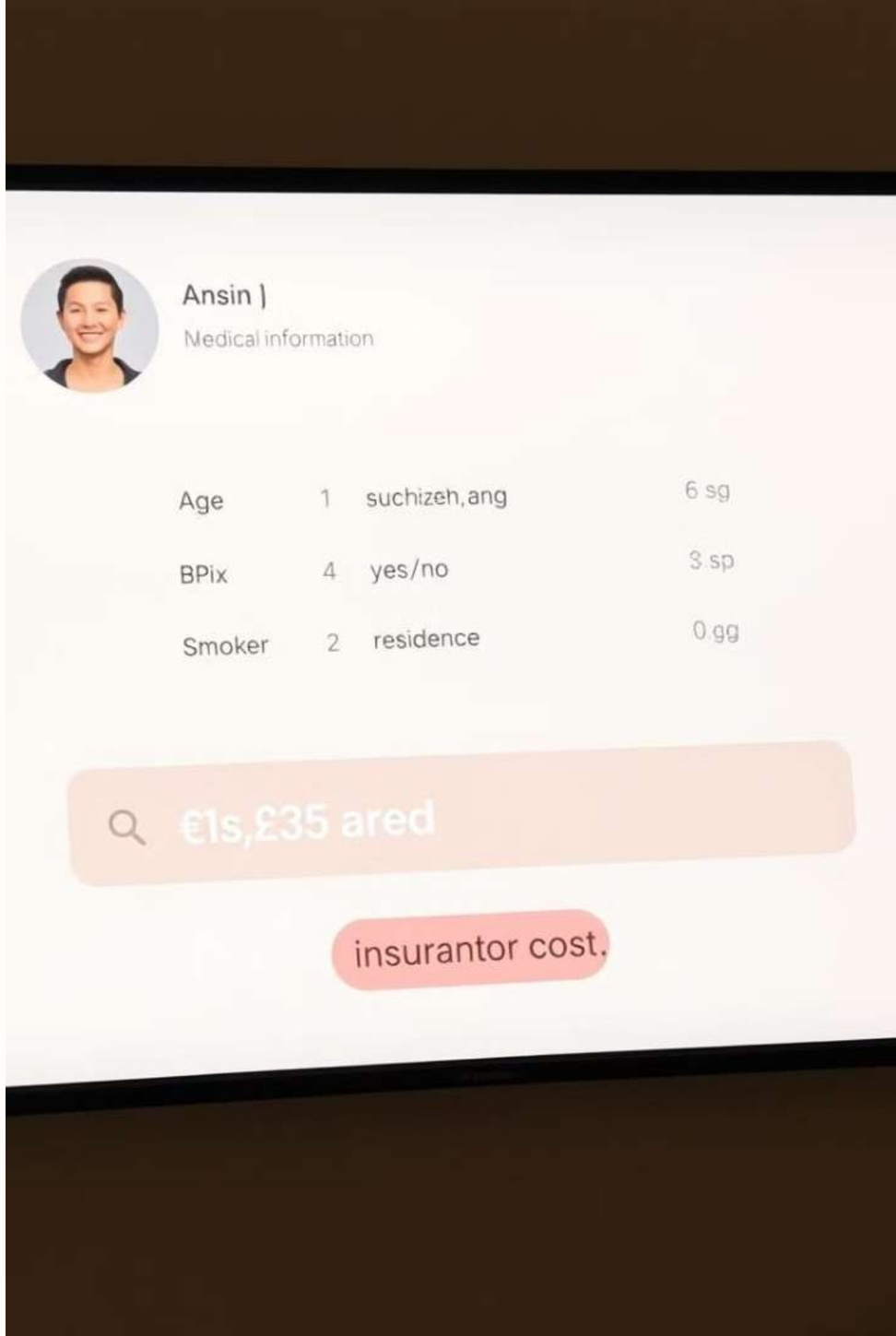
Input

Age: 37, Sex: Female, BMI: 30.8, Children: 2, Smoker: No, Region: Southeast

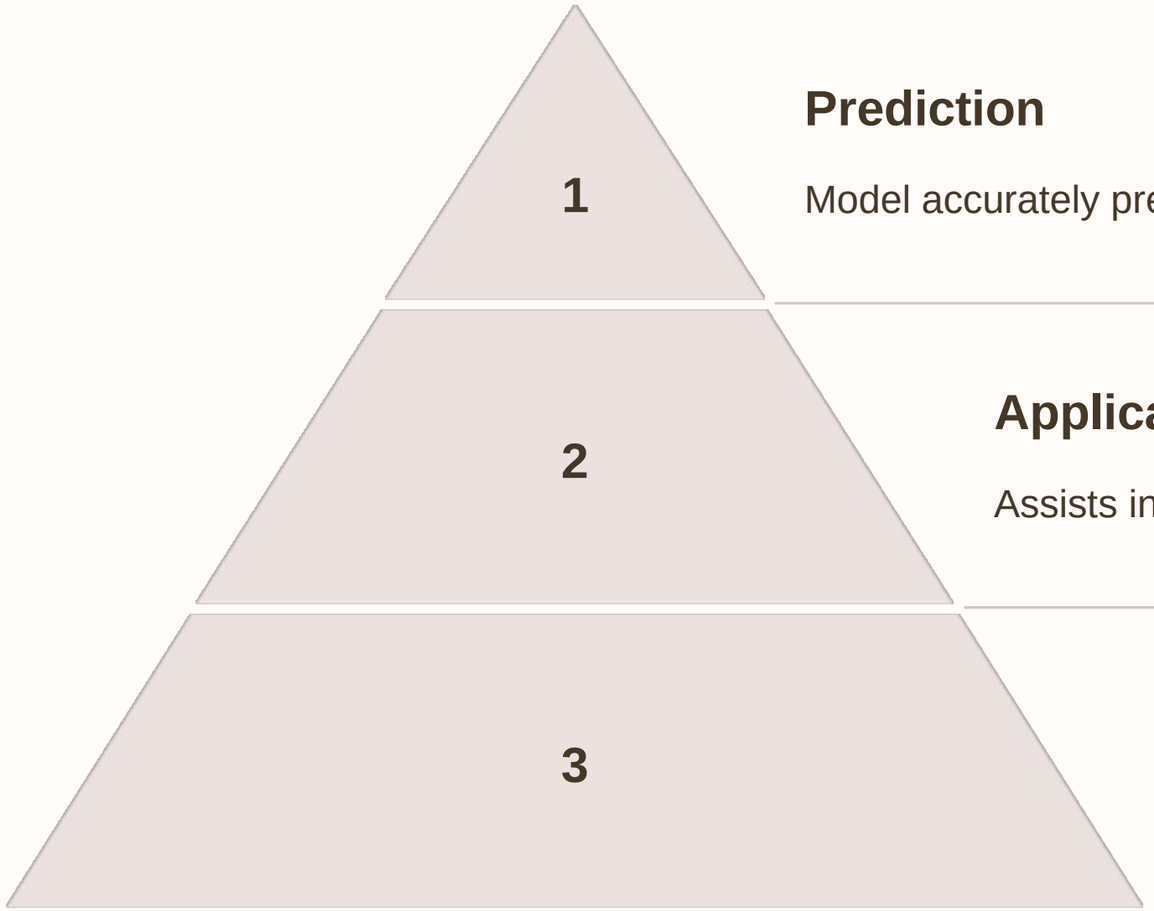
2

Output

Estimated insurance cost: $8102.13



Conclusion



Prediction

Model accurately predicts insurance costs using relevant features.

Applications

Assists insurance companies in risk assessment and pricing.

Future Work

Explore more complex models, improve accuracy, and incorporate new features.