PROPOSAL

Kartik Sharma

MSc. Data Analytics 2022-23

x21125813

Data Mining and Machine Learning – 1

*Motivation*

**Flight Price Prediction:**

World has again started gaining pace after a long time of recovering from the pandemic. With the ease in the restrictions, the offices, tourist attractions, restaurants, educational institutions have the process of reopening, the employees are returning, tourists have started boarding their flight, international student are looking forward to meeting their mates in college. In this all the most important question that comes to a travelling individual is “when and how much should I spend on a flight ticket?”. The Flight dataset have details ranging from the source city, price, many other attributes from the 6 major cities in India. The motive is to find the critical features affecting the price of a flight.

**Superstore Sales:**

Today every entrepreneur wants to make his/her business reach new heights and every customer (product consumer) expects an ease and comfort of buying products with the highest possible discounts. A right amount of research and analysis can uncover the hidden factors that derive the sales of a particular product. It can also help the product owner to decide right time and discounts on products in order to gain more customers without hampering the profits.

**Heart Disease Data:**

In today’s times, when every want to be successful, we tend to ignore our health. There is no meaning to success without a healthy body. 32% of global deaths in 2019 were due to cardiovascular disease [1]. Clearly there’s an urgent need to put technology to its right purpose and develop a system that can predict the cardiovascular disease, before it takes a toll on an individual’s life.

*Research Question*

**Flight Price Prediction:**

The objective of this research is to study on the attributes of flights in order to predict most legible prices, using the Multiple Regression and using the Logistic regression whether individual should opt for the flight or not.

**Superstore Sales:**

Profits and Sales are key factors in making a business grow to new heights, with the use of KNN and Multiple Regression we can predict both the factors, thus can help in determining the critical features that derive these factors, indeed helping a business to take data-driven decisions.

**Heart Disease Data:**

The research is to find the most critical factor affecting the cardiovascular health and to predict a cardiovascular disease before a person suffers its ill effects. Random forest or XgBoost can help in predicting the disease while Multiple Regression will determine critical features affecting cardiovascular health.

*Initial Literature Review*

**Flight Price Prediction:**

As per an estimate by FlightRadar24, there are almost an average 115,000 taking off and landing in a day. Everybody from a tourist to a student, is traveling by air, a system can be built to help these individuals, or in some case the corporates to save expenses on flight tickets. The Literatures/Papers discuss an approach of how predictive model created by applying machine learning algorithms on the historical flight data can help in achieving this objective.

By using Linear Regression algorithm [2], we can easily predict the price of a flight based on various attributes like duration of flight and season (like: - Holidays). Using Logistic regression [3], we can predict whether the flight should be booked or not, in order to minimize the expenses. The evaluation is done, and final predictor model is selected.

**Superstore Sales:**

A true business success is not based on how much profit it made, but its how much it learned from the past. This paper discusses the approach to handle the Big Data and using the previous sales data to predict the future sales. Linear Regression algorithm [2] is used to achieve the goal along with various metrics to evaluate the models.

Also, by using Random Forest algorithm [5] we can predict the profits for future.

**Heart Disease Data:**

A healthy lifestyle leads to a healthy life, but in this everchanging world, sometimes we forget to spend time looking after our health. Today a major percentage of population is suffering from cardiovascular diseases. With the power a machine learning, a solution can be developed that can help doctors and individual to detect any heart related disease based on some attributes like cholesterol and Thalassemia.

This research uses various classifiers like KNN [6] and Support Vector Classifier [7], to predict whether a patient is suffering from a cardiovascular disease based on the data collected. Vote [8] with a hybrid approach of Linear Regression and Naive Bayes [9] is used, to achieve 87.4% accuracy in predicting heart diseases.

*Data Sources*

**Flight Price Prediction:**

<https://www.kaggle.com/shubhambathwal/flight-price-prediction>

**Superstore Sale:**

[*https://www.kaggle.com/rohitsahoo/sales-forecastings*](https://www.kaggle.com/rohitsahoo/sales-forecastings)

**Heart Disease Data:**

[*https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset*](https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset)

*Machine Learning Methods:*

***Flight Price Prediction:***

1. **Logistic Regression:**
2. **Multiple Regression:**

***Superstore Sale:***

1. **K-Nearest Neighbours:**
2. **Multiple Regression:**

***Heart Disease Data:***

1. **Random Forest:**
2. **XgBoost:**
3. **Multiple Regression:**

*Evaluation Methods:*

1. **Performance Metrics:**
2. **R2(Coefficient of determination):**
3. **Precision:**
4. **Recall:**
5. **Confusion Matrix:**
6. **Root Mean Squared Error:**

*Bibliography*

[1] Cardiovascular diseases (CVDs)

<https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=Cardiovascular%20diseases%20(CVDs)%20are%20the,%2D%20and%20middle%2Dincome%20countries>.

[2] Linear Regression

By Rohith Gandhi

<https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a>

[3] Logistic Regression

By Jason Brownlee

<https://machinelearningmastery.com/logistic-regression-for-machine-learning/>

[4] Multiple Regression

By Adam Hayes

<https://www.investopedia.com/terms/m/mlr.asp>

[5] Random Forest

By Sruthi E R

<https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/>

[6] KNN

By Joos Korstanje

<https://realpython.com/knn-python/>

[7] Support Vector Classifier

By Bruno Stecanella

<https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/>

[8] Vote

By Aashish Nair

<https://towardsdatascience.com/combine-your-machine-learning-models-with-voting-fa1b42790d84>

[9] Naive Bayes

By Sunil

<https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>