Ira A. Fulton Schools of Engineering
Arizona State University

PROJECT PORTFOLIO COVER SHEET

Instructions: Students must complete all sections below. The completed form must be submitted to the Project Portfolio Canvas course with the project portfolio by the posted semester deadline.

Section I: Student Information: Student must complete this section

Last Name Wenzhe	First Name Zheng	ASU ID 1216729826
Graduation Term Fall/2021	All GPAs above 3.0? ☒ Yes ☐ No	Date submitted 11/14/2021

Section II: Project Information: Student must complete the sections below. Faculty signature or email approval is required for courses taken prior to Spring 2020. Courses taken Spring 2020 and later do not need faculty approval. Online students are not required to submit faculty approval.

PROJECT 1: Instructor name –Samira Ghayekhloo ☒ Project worth at least 30% OR ☐ Project under 30% with additional work done Instructor attestation (initials) ______________		
Course CSE578 Data Visualization	Semester completed Spring/2020	Final grade in course A
Instructor Statement: My signature below certifies that the project is applicable for the CS portfolio. I have also read the project description for this course and approve the content. **Instructor signature:** _____________________________ **Date:** ________ ☒ **Course taken Spring 2020 or later**		

PROJECT 2: Instructor name - Ayan Banerjee ☒ Project worth at least 30% OR ☐ Project under 30% with additional work done Instructor attestation (initials) ______________		
Course CSE535 Mobile Computing	Semester completed Spring/2020	Final grade in course A
Instructor Statement: My signature below certifies that the project is applicable for the CS portfolio. I have also read the project description for this course and approve the content. **Instructor signature:** _____________________________ **Date:** ________ ☒ **Course taken Spring 2020 or later**		

School of Computing and Augmented Intelligence
PO Box 878809 Tempe, AZ 85287-8809
p: 480-965-3190 **f:** 480-965-2751 **email:** scai@asu.edu **web:** scai.engineering.asu.edu

Summary of Final Portfolio

Wenzhe Zheng
Arizona State University
Tempe, American
wzheng41@asu.edu

I. PROJECT1:CSE578 DATA VISUALIZATION

Data visualization is the graphical representation of information and data. By using visual elements such as charts, graphs, and maps, data visualization tools provide an accessible way to view and understand data trends, outliers, and patterns. In the world of big data, data visualization tools and techniques are essential for analyzing large amounts of information and making data-driven decisions. Data visualization is also the last step of big data output, so this lesson is very important for big data.Visualization is an increasingly important tool for understanding the trillions of rows of data generated every day. Data visualization helps tell stories by organizing data into a more understandable form, highlighting trends and outliers. A good visualization tells a story. The process of data visualization can be divided into: obtaining data, cleaning the data (eliminating the noise of the data), using a specific model to analyze the data and making predictions, and using a specific table or image to show the data. Help people to read information at a glance.

In this team project, we used the population income data provided to us by the U.S. Census. The purpose of this team project is to use the income threshold of $50,000 to help UVW University find suitable candidates and help them analyze humans which attributes have the greatest impact on personal income data.

At this time, the team members started from multi-dimensional research in order to display the results in a more perfect form. In the first part, the research team made the corresponding 8 attributes [Education Num, Relationship, Age, Hours Per Week, Sex, Capital-Gain, Marital-Status, Capital Loss] in turn to visualize the impact of different forms of data on income Images, such as: Bar Plot of Education-Num VS Income, Mosaic Plot of Sex VS Income and Scatter Plot of Age and Capital-Gain VS Income, etc. In the second part, the research team used a variety of machine learning classification methods to predict Results, for example: SVM machine module,LogisticRegression machine module and Decision Tree machine machine module,ect.

After the data in the first part was visualized, the research team used the weight calculation method to draw the Heatmap and concluded that the first six attributes that affect personal income are: 1) Education-num 2) Captial-gain 3) Age 4)Sex 5)Marital-Status 6)Capital-Loss, for the second part of the prediction of the data by different machines, the research team also came up with the best two machine models, which are the SVM Machine Module with a prediction accuracy of 84.5%. and LogisticRegression Machine module with an accuracy rate of 84.4%. For the detailed process, please refer to the separate report below.

II. PROJECT1:CSE535 MOBILE COMPUTING

GPS (Global Positioning System), originally the NAVSTAR global positioning system, is a satellite-based radio navigation. It is a type of global navigation satellite (GNSS). Now it has been popularized for civilian use. People will have this GPS software on their mobile phones. Whether it's a Google map or an app map, these navigation software bring great convenience to human life. When you go out, directly enter the destination and there will be a route to lead us forward, but in general, these GPS software will guide customers Choosing the shortest route or the most time-saving route does not take into account the fuel of the vehicle. Today, with the rapid increase in fuel prices, it is not good if the shortest route or the most time-saving route consumes more gasoline. Therefore, this team project will develop a GPS software that calculates the most fuel-efficient route for customers to use.

In this team project, the researchers will design a novel calculation method in the Android Studio environment to help customers find the first four routes with the fewest friends from the start point to the end point. This program uses the Google Map API as the basis for the map. It can read the real-time traffic situation on the road, the name of the building on the roadside, the distance, and the traffic signal timing. What is indispensable is the calculation of the fuel consumption algorithm of each road and the reading parameters of the intersection. Here We used BFS as the search algorithm and calculated the top four most fuel-efficient routes. Of course, in the context of these algorithms, it is also essential to create a beautiful and exquisite user interface. The researchers used the entire tempe area as the test area for this GPS software. In order to facilitate the calculation of each intersection, the researchers divided the entire Tempe area into a 6*6 matrix, which is convenient for BFS to search, and the test uses Tempe station as the starting point. , Take Mcclintock high school as the end point for multiple tests, and conduct on-site inspections to verify whether the fuel consumption is correct or not. The results prove that the fuel consumption algorithm is basically consistent with the actual situation. Please check the detailed report below for specific practices.

CSE578 Portfolio Report

Wenzhe Zheng
Arizona State University
Tempe, American
wzheng41@asu.edu

***Abstract*—Data visualization is the most important part of learning big data. After the data is cleaned up, it is presented to the audience in various graphs or tables, such as: Bar Chart, Pie Chart, QQ Plot, Mosaic Plot, Scatter Plot, Parallel Coordinate Plot, Star Plot, Box Plot, Line Chart, Histograms, Heatmap, etc. In this project, we will use these charts and data predictive analysis together, and at the end we will predict the future trend.**

***Index Terms*—Bar Chart, Pie Chart, Q-Q Plot, Mosaic Plot, Scatter Plot, Parallel Coordinate Plot, Star Plot, Box Plot, Line Chart, Histograms, Heatmap, Decision Tree, KNN [1], PCA (deduce all features into two features) and apply classification method, Linear Regression, Navie Bayes, SVM [3]**

I. INTRODUCTION

This project is a team project of CSE578. The purpose of this project is to exercise our ability to visualize and analyze data. In this project, the data we use is the income data provided to us by the United States Recognition Census. Our goal is to use US$50,000 the income threshold helps UVW University to help them find suitable candidates, thereby increasing the enrollment rate, so we have to determine which attributes have the greatest impact on personal income and the relationship between all attributes and income. We will use many machine learning classification methods,to predict the income of a person by inputting important characteristics.

This project is divided into two parts. The first part is called: *Exploratory Data Analysis*. This part uses a variety of graphics to support data visualization to find the relationship between various characteristics, trends, and variables. To this end, we thoroughly train the entire database, perform various data prepossessing steps, such as separating elements and features, transforming categorical variables into 0 and 1, and dividing the data into training and testing data sets, and then transforming them into standardized Digital form and implement visualization. According to the visualization that has been carried out, we will analyze the relationship between various variables and characteristics and the proportion of income.

In the second part of the project, we called: *Results Analysis*. In this part, we used several data classification models such as Naive Bayes, SVM [3], Decision Tree, Logistic Regression, and KNN to predict the relationship between variables.

A. About the dataset [1]:

- Data Source:US Adult Census
- Label:">50K" and "<50K"
- RangeIndex: 32561 entries
- Data Training:Remove unknown value ('?')
- Memory Usage: 3.7+ MB
- Number of data in dataset after removing unkown value: 30162

B. Project Package Used:

TABLE I
PACKAGES

Language	Python
IDE	Jupyter Notebook
Packages	Pandas, Numpy, Matplotlib, Sklearn, Seaborn
Data Visualizations	Bar Chart, Pie Chart, Box Plot, Mosaic Plot, Scatter Plot, Heatmap
Prediction Models	Naive Bayes, SVM [3], Decision Tree, Logistic Regression, KNN [1]

In the following section called Solution, I will explain in detail the relationship between each graph and the results obtained after using each model to analyze the data.

II. SOLUTION

A. Phase One: Exploratory Data Analysis

Since we created a correlation heatmap(in Figure 1) based on the whole 14 attributes and the income label, we could assume that the top 8 darkest squares would be the attributes we wanted. However, the correlation can be negative so even though Relationship has a large negative effect on Income, it still is a key factor to be considered. Therefore, we assumed that we found the top 8 largest absolute values of the income row. After this operation we can have a sorted list of attributes which is [Education Num, Relationship, Age, Hours Per Week, Sex, Capital-Gain, Marital-Status, Capital Loss]

From this Heatmap, we have selected several interesting data and analyzed them. The data are as follows:

1)Bar Plot of Education-Num VS Income: From the heatmap visualization we can see that the Education-Num attribute plays the most important role when it contributes to the Income

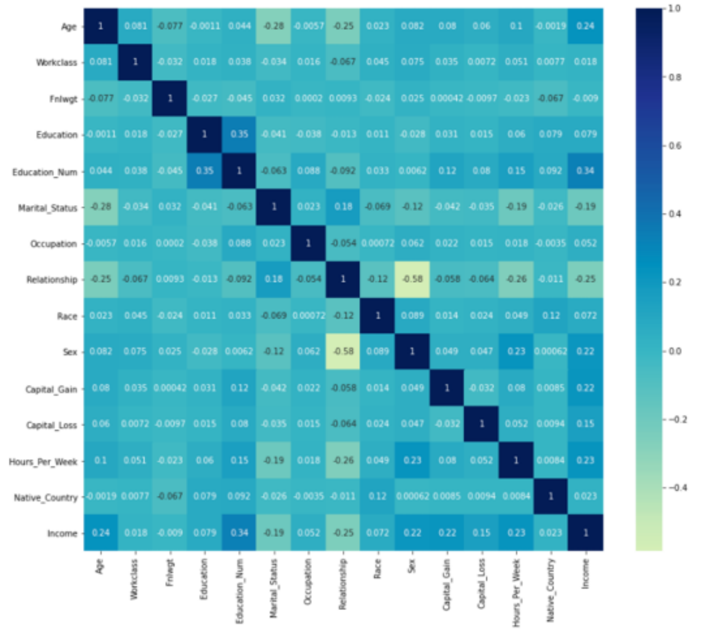

Fig. 1. Data Visualization Heatmap

label and it makes sense because the higher education a person gets, the more likely he/she will get high income. From Fig 1 we can see that if people get higher than level 13 education level, they have a higher chance to get over 50K of income.

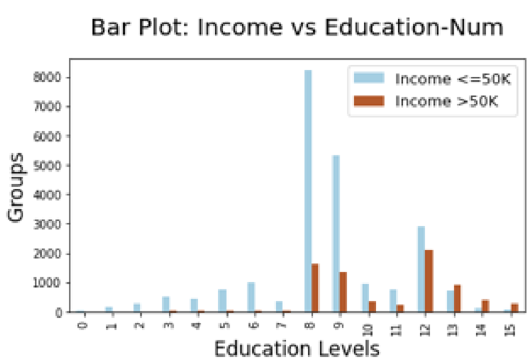

Fig. 2. Bar Plot of Education-Num VS Income

*2)Box Plot of Education-Num VS Income:*A Box Plot can always show the percentage of a group. As Fig 3 shows, there is a 75% of chance that a person gets under Education Level 10 when the income is under 50K while only a 25% of chance that a person gets under 10 when the income is over 50K. Besides, Education Level over 12 is considered an outlier when the income range is under 50K while there are no outliers of high education level in the income range of over 50K. This means that the higher education one gets, the more likely he/she has a high paid job.

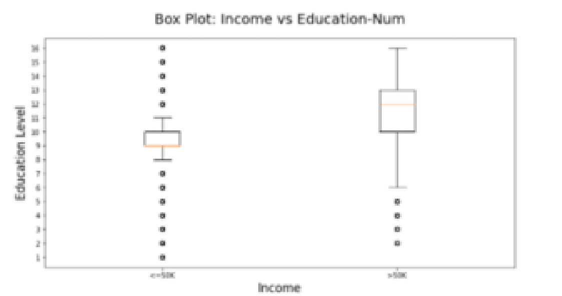

Fig. 3. Box Plot of Education-Num VS Income

*3)Mosaic Plot of Sex VS Income:*A Mosaic Plot is suitable for Sex VS Income plot because there are only 4 regions that should be visualized. As the plot shows, it infers that if one gets over 50K, the person is more likely to be a Male. Besides, there is more Male data than Female in the original dataset.

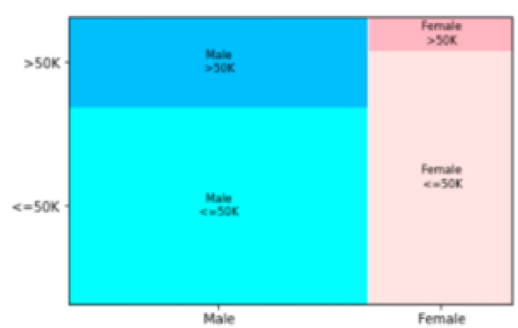

Fig. 4. Mosaic Plot of Sex VS Income

*4)Bar Plot of Marital Status VS Income:*Based on the heatmap, the marital status shows very low coefficient of income, however, it is an attribute that is always been asked or considered in real job seeking situations. So it is still a key factor of income which makes the plot meaningful. Bar plot is suitable with marital status VS income which can show clearly the number of people earning different incomes in different marital status. Based on the plot, we can see that no matter which marital status is, the number of people earning less than 50K is larger than the number of people who earn more than 50K. However, a civilian spouse can have much higher probability to earn more than 50K than other marital status.

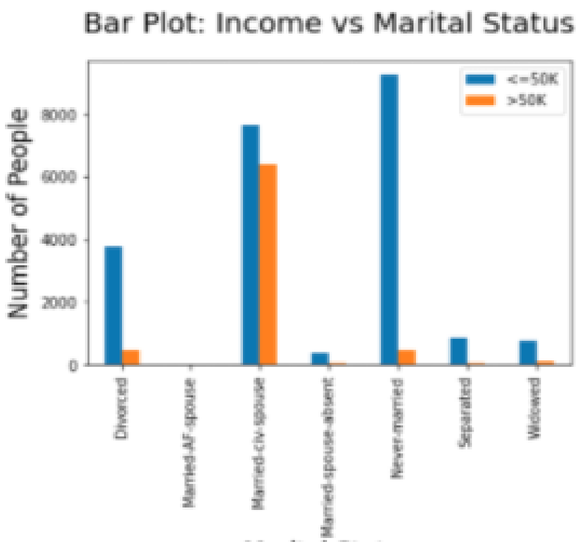

Fig. 5. Bar Plot of Marital Status VS Income

*5)Pie Plot of Workclass VS Income:*Based on the heatmap, the workclass has a low coefficient of income. But based on the real situation, workclass is always a determinant of how much income you can earn. The plot shows that a private has both very high probabilities to earn less than or more than 50K compared to other work classes. For other work classes, they all have higher probabilities to earn more than 50K.

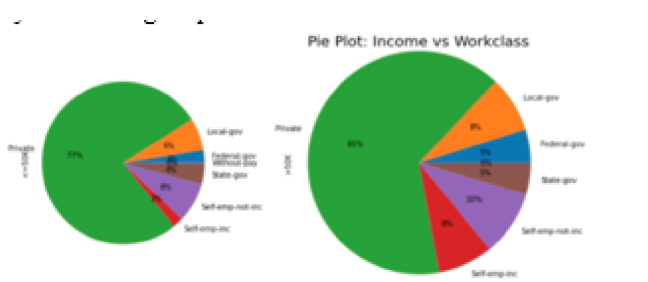

Fig. 6. Pie Plot of Workclass VS Income

*6)Scatter Plot of Age and Capital-Gain VS Income:*From the heatmap visualization we can see that the Capital-Gain attribute plays the most important role.It can be seen from this chart that capital gains are not closely related to age, but capital gains are closely related to income. Only when the income is greater, the capital investment and capital gains can be greater.

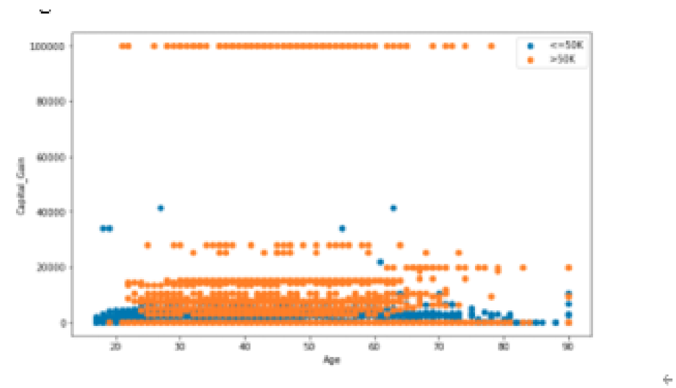

Fig. 7. Scatter Plot of Age and Capital-Gain VS Income

7)Bar Plot of Hours-Per-Week VS Income: Based on the heatmap, the Hours-Per-Week attribute also occupies a certain proportion of the income. After all, any work requires the work of working hours, and only work can be rewarded. We can see from the figure that, regardless of their income, those who work 36-45 hours a week account for the highest proportion of all people, indicating that 36-45 hours a week work is recognized by everyone and the best reporting ratio can be derived.

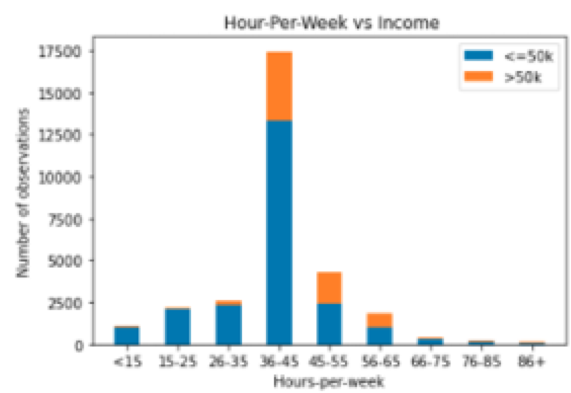

Fig. 8. Bar Plot of Hours-Per-Week VS Income

*8)Line Plot of Education VS Income:*Education is one of the most significant factors that influence people's earnings. As the heatmap visualization shows, education has a strong correlation with income. It's worth noting in the below graph that people with higher education levels, such as master's, professional school, and doctorate degrees, are more likely to earn higher wages than those with lower education. Also, we can see that people with lower educational levels have a lower chance to get higher salaries (¡=50 K).

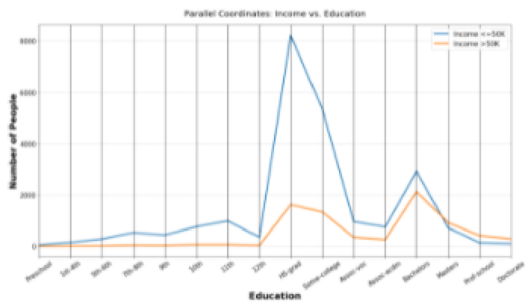

Fig. 9. Line Plot of Education VS Income

B. Phase Two: Results Analysis

In this part, we used NaiveBayes, SVM [3], LinearSVM, LogisticRegression, KNN [1], DecisionTree to train the data, and got different accuracy, the data is as follows:

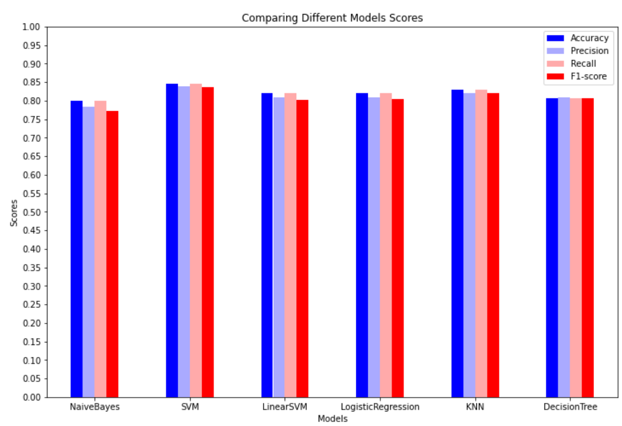

Fig. 10. All models accuracy rate and other comparisons

From the above figure, we can conclude that the accuracy rates of SVM [3] and LogisticRegression are very high, respectively 84.5 and 84.4.
Based on the above demonstrations of various pictures and our research using six machine learning models for predictive analysis, we have summarized the six most important features for predicting whether personal income is higher or lower than 50K. They are:

1)Education-num
2)Captial-gain
3)Age
4)Sex
5)Marital-Status
6)Capital-Loss

This project was done by ourselves but had not been reviewed by other teams. Besides, we did not have a chance to review other teams' results so the final exploratory analysis and predictive analysis were not authorized by either the marketing department (our customer) or other authorities. Therefore, our team's plan is to get someone that has experience in this marketing and business analysis field and let the professional people check on the final results in order to acknowledge the real world marketing system.

III. CONTRIBUTIONS

This is a team project. I have the honor to complete the analysis and sorting of this project with Wei Xin, Zian Zhang, Xinyi Liu, DongAo Ma, Mohammad Reza Hosseinzadeh Taher. Thank them very much. I have also participated a lot in this project. Things like:

- I drew Scatter Plot of Age and Capital-Gain VS Income(Figure 7) and Bar Plot of Hours-Per-Week VS Income(Figure 8)
- Organize a zoom meeting. Since our team members have all returned to their own countries due to the epidemic, we have three time zones. I used the meeting planner website to find the time when we can unify the meeting.
- Create task lists and assign them during discussions
- Create Github and share with group members
- Participate and discuss with the group members how to train data
- Use SVM machine model to analyze and predict data
- Pass the request of the member's merge code on github, and combine the code
- View the progress of other team members' data visualization
- Participated in and compiled project execution reports and system reports

IV. LESSONS LEARNED

The following list is what I learned from this course

- Learned how to train data more scientifically
- Learned a variety of graphics to better express the situation of the data
- Create task lists and assign them during discussions
- Create Github and share with group members
- Learned to use machine learning models such as NaiveBayes, SVM, LinearSVM, Lo-gisticRegression, KNN [1], Decision Tree to predict data
- Learned to express mine opinions more effectively in group discussions
- Learned how to efficiently carry out team work and manage other people's works on github

REFERENCES

[1] Harrison, O. (2019, July 14). Machine learning basics with the k-nearest Neighbors ALGORITHM. Retrieved May 03, 2021, from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761
[2] https://archive.ics.uci.edu/ml/machine-learning-databases/adult/
[3] andhi, R. (2018, July 05). Support vector machine - introduction to machine learning algorithms. Retrieved May 03, 2021, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

CSE535 Portfolio Report

Wenzhe Zheng
Arizona State University
Tempe, American
wzheng41@asu.edu

***Abstract*—This project developed an Android mobile phone application software as required. Customers can choose the starting point and ending point in a selected area, and can choose to add a stop between the two points, and then the mobile phone program will use the user's starting point position to predict the distance to the end point, fuel consumption and time. Finally, the most optimized route is generated and the most fuel-efficient route is sent to the customer.At the same time, this ascending order allows the customer to map the time of the specified route to the Google map [2] for comparison, so as to allow the user to choose the most convenient option**

***Index Terms*—Google map API [2],Android Studio,Java,Fuel consumption,GPS**

I. INTRODUCTION

The map is the most important part of the mobile phone software, and the navigation system is one of the most important functions of the map. There are many companies in the world that provide global positioning system (GPS) functions and provide them to customers who need them. For example: Google, Baidu, Gaode, etc. These systems will help users find the starting point and ending point according to their needs. The most commonly used routes. The most common recommended route is the shortest route. Regardless of the traffic conditions on the road, fuel consumption and how the time is calculated, the distance from the starting point to the destination is the smallest. As fuel is used as the driving force for vehicle driving, the factors in calculating the road should also be taken into consideration. So we designed a novel calculation method to help customers find the route recommended by the vehicle with the least fuel consumption. This is an application we developed in Android. This program uses the Google Map API as the basis for the map. At the same time, we take distance, real-time traffic conditions on the road, traffic signal timing, and fuel consumption as the key elements for calculating each road. Then we use BFS as the search algorithm and calculate the top 4 routes that use the least fuel for all key elements. The map scope of our application this time is set to Tempe. Because the City of Tempe has added an adaptive system [3] for traffic lights, we cannot obtain real-time traffic signal data from the city government database or Google Api. So we took a hypothetical measure to help us process traffic signal data, and we will introduce this method in detail below.

A. System requirements:

- IDE: Android Studio 3.4
- Map API: Google Directions API [2],Google Distance Matrix API,Google Maps SDK for Android,Google Roads API
- Emulator(AVD): API 28+,Huawei Mate Pro10
- Program language:Java
- Packages:Android.Manifest,Android.location, Android.graphics, Com.google.maps

II. SOLUTION

In this part, we are divided into three parts to complete, the first is about the UI interface part of the phone:

A. Create UI element for Android Phone

The first page is the login interface. This project only provides the illusion of an algorithm, so the main page does not have too much design. You can see that after directly entering the page, there is a button to activate the GPS map:

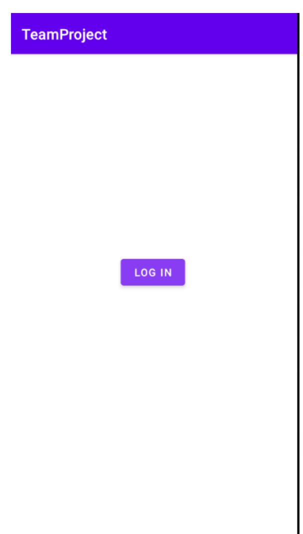

Fig. 1. Login Page

After clicking the Button, you will enter to the second page, and the GPS map function will be activated at the same time. The latitude and longitude that appears directly below the page

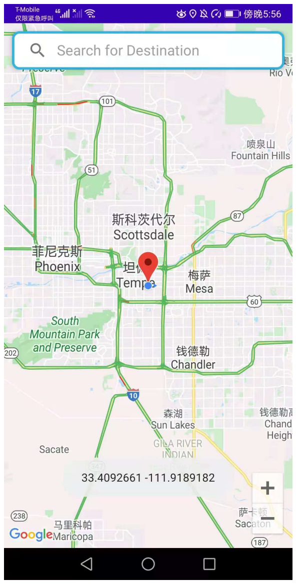

Fig. 2. Map UI with Search Box

In this interface, you can see that when we zoom in on the map, we can see the real-time traffic flow on the road. Blue means everything is unimpeded, yellow means some congestion, and red means particularly congested. At the same time, we can also see the names and functions of the buildings on both sides of the road. By looking at the icon, we can distinguish whether it is a supermarket or a gas station

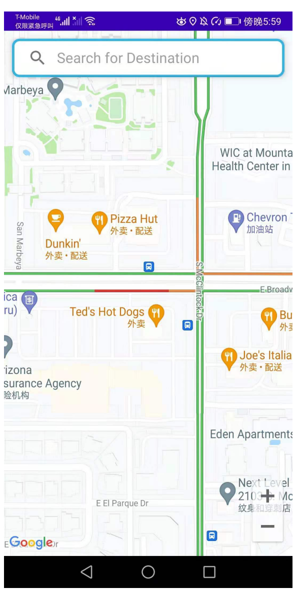

Fig. 3. Traffic conditions and Building name

We can see the traffic light after continue zoom in:

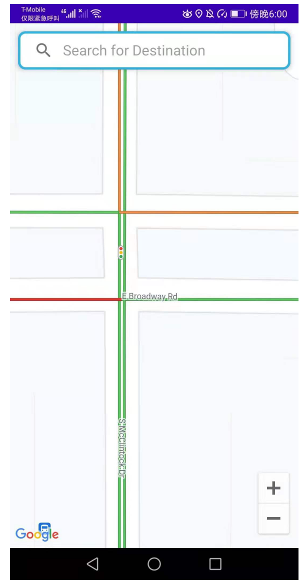

Fig. 4. Traffic Light

We can add a stop point by touching the screen with our finger, and we can also click the stop again to cancel it

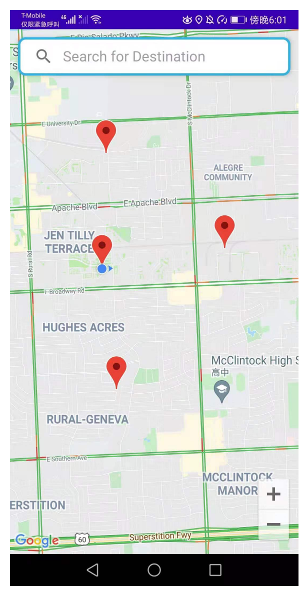

Fig. 5. Stop point

We can write the destination we want to go to in the search form and click on the search in the lower right corner. Here we enter Mcclintock high school

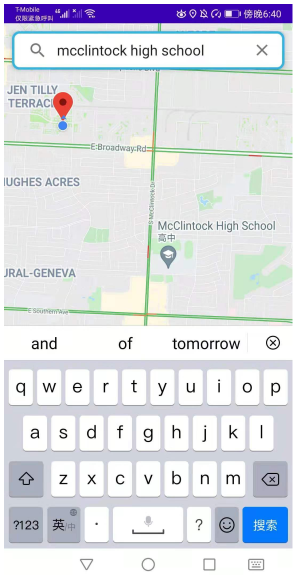

Fig. 6. Search

Then the map will calculate the four feasible roads to reach Mcclintock high school, and will also display the fuel consumption, time required, total distance and number of intersections passed by each corresponding road:

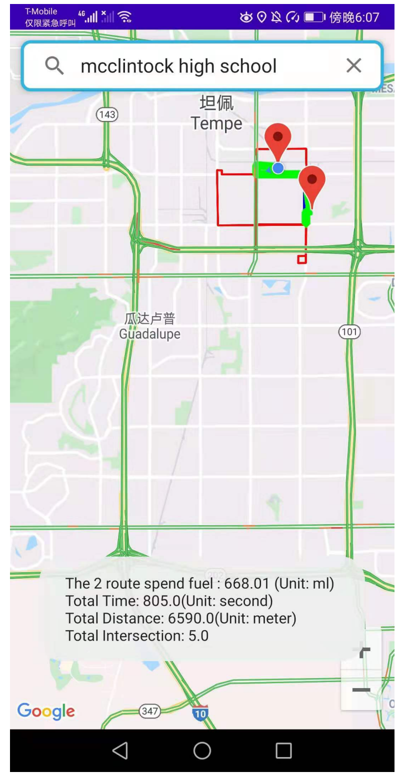

Fig. 8. Route2

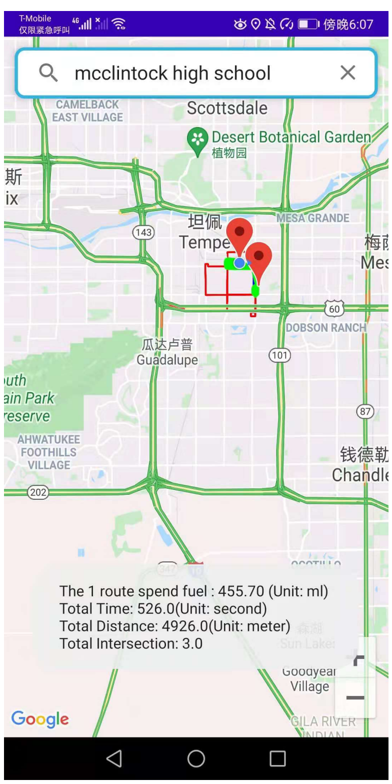

Fig. 7. Route1

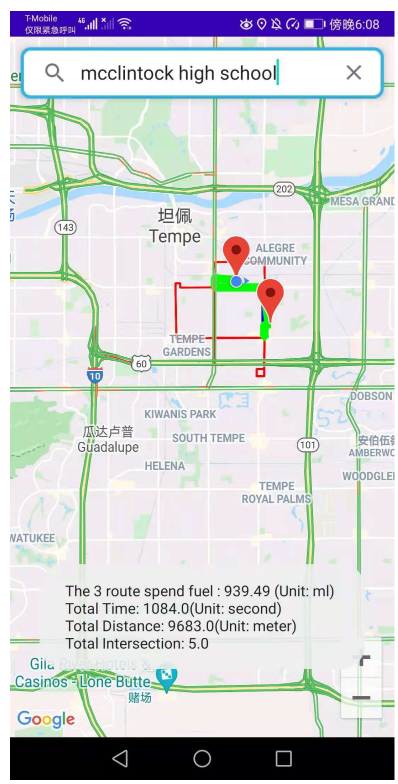

Fig. 9. Route3

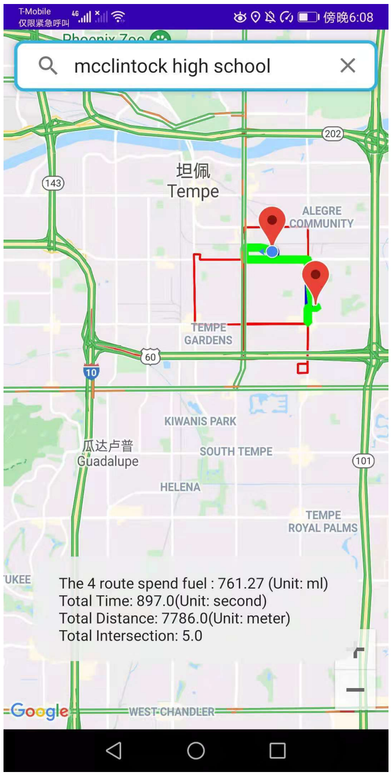

Fig. 10. Route4

The blue route in the last picture is the best route chosen by the app for users based on fuel consumption, distance, time and road conditions. The blue route is the best route generated by Google [2], and the green route is the best route generated by the reference program. We can see that these two routes are basically the same.

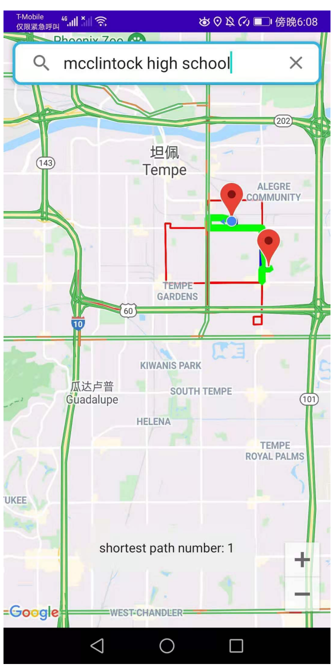

Fig. 11. Best Route

B. Map operation logic

The logic of this part is quite critical. It solves the problem of not getting real-time traffic light data and finding the path. I will explain it in detail below. We first applied for the Google Map API [2] to embed a map for the program. When we click the button on the main page, we will get our current location through the GPS provided by Google and display it in the reference program. Later, we also added Google Directions API, Google Distance Matrix API, Google Maps SDK for Android, Google Roads API provides us with a series of requirements such as road condition information. Then, we set the scope of the project to Tempe, and manually entered the latitude interval (33.383582, 33.439985) and longitude interval (-111.983661, -111.887204) of this location. All other coordinates and positions that are not in these intervals are regarded as invalid input, and the reason: no invalid input can be found. Displayed on the screen. At the same time, as shown in the figure below, we roughly divide the entire Tempe into a 7*7 matrix according to the direction of the road, manually input the longitude and latitude of each intersection, and at the same time we manually record the traffic signal time of each intersection, these All the information is included in 2.txt, which also solves the problem of not getting traffic signal information because the Tempe city government added an adaptive system [3] to the traffic lights. At present, we set the signal time of each intersection as 30 seconds of red light time for calculation.

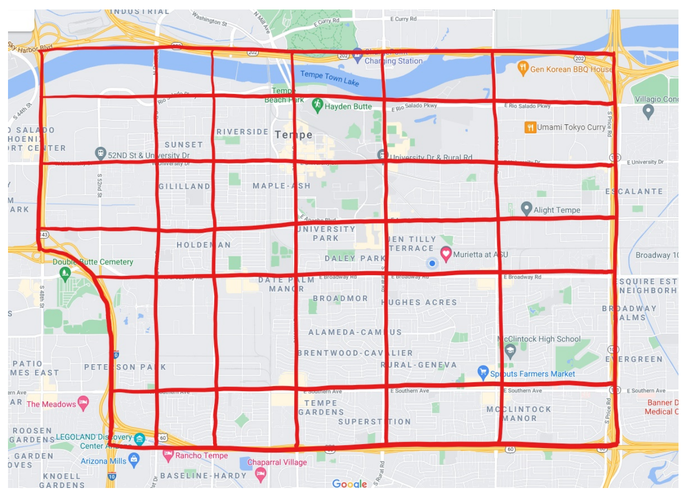

Fig. 12. Map Matrix

When the user confirms the starting point, stop point and destination, we will call Google Map API to determine the latitude and longitude corresponding to each point, and then we will find the closest point and The corresponding intersection on the map. From the starting point, we used the logical algorithm of BFS search. But we slightly changed the algorithm to achieve the goal i we want to find the path. We use the priority queue as the queue data structure, and each node in the queue will be calculated according to the following formula:

$$Value(P,S) = dist(P,S) + 0.8 * [T_{light}(P) + T_{light}(S)]$$

$where:$

P : *The current intersection*

S : *The next intersection along the road*

$dist(P,S)$: *Distance between intersection P and intersection S*

$T_{light}(X)$: *Traffic light timing for intersection X* Our algorithm this time uses priority queues and BFS, and searches every possible path until four paths are found. In the algorithm, we only consider the distance and the time of the traffic signal. Below we will explain the fuel consumption algorithm in this project.

C. Calculate Fuel Consumption

After we have four different routes, the next step is to calculate the fuel consumption of each route. Fuel economy (miles/gallon) is a common indicator of fuel consumption, but because each car has a different model and power, the economy of each car is different, so we find another way to Calculate the average fuel consumption of a random vehicle. [4] Another way to calculate fuel consumption is to use the speed shown in the figure. From the figure, we can see the relationship between fuel consumption and speed: When average car speed is less than 25 mph: [3]

$$AverageSpeed < 25(mph) \tag{1}$$

$$FuelEconomy = 5.75 + 0.85 \times AverageSpeed \tag{2}$$

When average car speed is greater than 25 mph but less than 57 mph:

$$25(mph) < AverageSpeed < 57(mph) \tag{3}$$

$$FuelEconomy = 24.5 + 0.1 \times AverageSpeed \tag{4}$$

When average car speed is greater than 57 mph:

$$57(mph) < AverageSpeed(mph) \tag{5}$$

$$FuelEconomy = 47.1 - 0.3 \times AverageSpeed \tag{6}$$

By using three formulas from above, the application is able to approximate fuel economy using only average route speed.

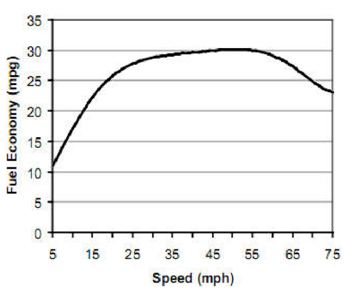

Fig. 13. Relationship between fuel economy(mpg) and speed(mph)

III. CONTRIBUTIONS

This is a team project. I am honored to complete the development of this project with Zian Zhang, Kekang Shu and Shanna Peterson. Thank them very much. I also participated in this project. Like:

- I manually input all the coordinates of 7*7Matrix and record the time of each signal light and input it into the database file
- Responsible for organizing remote meetings
- Create task form and assign it to everyone
- Use Github to manage the code of team members
- Apply for various Google APIs [2]
- Establish and use DFS to find the path
- Record and edit demo video
- Participate and write final group report

IV. LESSONS LEARNED

I learned a lot from this lesson, for example

- Understand and be familiar with using Android Studio
- Learned how to develop a GPS software logic
- Understand the process of developing an Android mobile phone software
- Learn to find the knowledge we need from Research
- Learned how to apply for API
- Learned how to use Github to manage the team
- Learned how to reasonably express one's point of view in the team

REFERENCES

[1] Kamga, C., Tario, J. D., Ancar, R., Yazici, A., Almotahari, S. Mudigonda, S. (2018). Reducing incident-induced emissions and energy use in transportation: use of social media feeds as an incident management support tool: final report. New York State Energy Research and Development Authority.

[2] Google API: https://developers.google.com/maps/documentation/javascript/reference

[3] Adaptive signal control technology. (n.d.). Retrieved May 02, 2021, from https://www.maricopa.gov/4553/Adaptive-Signals

[4] Gross, R. (2018). Standby Diesel Generator Fuel Consumption Calculation. https://doi.org/10.2172/1466196