*Instructions: Replace all bracketed text with text for your project. Feel free to add. Leave the document structure in place. Delete these instructions.*

***[Hadi Baajour, 96316680]***

**CIS4930 Individual Coding Assignment**

**Spring 2023**

1. **Problem Statement**

*Technology has been advancing exponentially ever since the 1960’s. What was once only a dream in the technology realm has now become a reality. A technological device that was once only a thought in people’s minds is the task-oriented dialogue system. These are devices like Siri, Alexa, Google home, etc. These types of devices are programs used to converse with users to help them with basic tasks, find directions, control other technological devices , and more. For this experiment, we will be looking at how a task-oriented dialogue system, specifically Siri, can assist a user in booking a ticket for a flight. Using different factors, we will determine which of them is most influential when it comes to Siri’s ability to book flights. Conducting this experiment will help us find out and target those factors that most affect Siri with its tasks.*

*It is important to solve this problem in order to make Siri a better device overall for users to actually use when conducting any household tasks. If the device becomes excellent at its tasks, it will eventually become a normal household item and everyone will own it (similar to the development of the television). Conducting this will allow us to find out what the issues with the device are and give us something to focus on when it comes to fixing the problem.*

*To solve this problem, we will be using independent variables and classifying them in order to see how they correlate with the dependent variable. For our experiment, we will have the Purchase as our dependent variable and the ASR\_Error, Intent\_Error, Duration, and Gender as out independent variables. To classify the data, we will be using evaluation metrics to compare the performance of four different ML classification algorithms. They are (1) Logistic Regression, (2) SVM, (3) Naive Bayes, and (4) Random Forest.*

1. **Data Preparation**

*To prepare the data, we must conduct the following steps:*

*We first check if there are any missing values in our data set. If there is, we replace those values with the mean value of the variable itself. Then, the independent and dependent variables will be labeled as x and y, respectively. Following this, we will scale the data to normalize the distribution. Next, we created a scaled version of the data by fitting and transforming it, which involved converting non-numeric features into numeric ones. Additionally, we incorporated SMOTE to address any disparities in the data by evenly distributing the data among the different types.*

1. **Model Development**
   * Model Training
   * *Now with all the values organized and filled in, we can begin the experiment. In this scenario, we have 70 percent of the data to train our classification model and we have the remaining 30 percent of the data to use for testing (in order to compare these data to the ones the model was trained with). These values were then placed in the four different classification models mentioned before. These can be summarized in the following:*
     + *Logistic Regression – This model attempts to estimate the probability of an occurrence happening by using the data provided to it. The dependent variable must be bounded to a binary form, either 0 or 1, since it is essentially predicting the probability: if it happens or it doesn’t. The model calculates the log odds, which is essentially the success probability over the failure probability.*
     + *Support Vector Machine – The SVM is used to locate a hyperplane in an N-dimensional plane, where N is the amount of features there is that categorize the given data values. The line that it creates tries to create the largest separation it possibly can between the N things so that it is easier to classify data fed to it in testing.*
     + *Naïve Bayes: This model will receive information before it can make any judgments on classifying data that is currently being added to it.*
     + *Random Forest – This model works by creating a multitude of decision trees when beginning its training. For this experiment, the output is usually the class that most of the trees select.*
   * Model Evaluation
   * *Logistic Regression*

*Table

Description automatically generated with medium confidence*

* + *Support Vector Machine*

*Table

Description automatically generated*

* + *Naive Bayes Classifier*

*Table

Description automatically generated with medium confidence*

* + *Random Forest Classifier*

*Table

Description automatically generated*

1. **Discussion**
   * *From the experiment we can see that all the models performed pretty good. This experiment allowed me to see how the variables relate to each other and how it affected our problem statement overall.*
   * *I loved doing this experiment because it helped me learn new technologies like Jupyter Notebook and Anaconda, while utilizing a lot of great python tools as well. It took a while to understand the project, but after reading it a lot, I eventually got it working.*
2. **Appendix**
   * *https://github.com/hadiplays/Assignment-1*