IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Predicting Gender Recognition by Voice and Speech Analysis

M.Swapna

Assistant Professor

Matrusri Engineering College,Hyderabad

Abstract :

This paper talks about predicting speech recognition technology, on this gift global machines are dominating and there are anywhere on this type of situation we seeking to train machines the way to interact with human beings or even with each other this is the coming technology, this paper concentrate at training system via constructing predictive version and supporting it to categorize the given voice parameters into male or female voice. The paper offers with the outline of speech recognition process, its applications in distinct sectors and additionally it offers with Analytics .Then it covers approximately predictive analytics later a part of record covers the constructing predictive version in Logistic Regression process, Finally the paper concludes on the distinct potentials makes use of the utility and in addition upgrades and considerations. The predominant purpose of the paper is Predicting Speech Recognition structures offers with predictive modelling, system learning, exploratory information evaluation and categorizing voice into male or female primarily based totally upon ancient statistics. Various voice residences are used which can be already gathered in a database as enter for the modelling.

Key Words: Logistic Regression, Machine Learning, Data Analysis, Speech Recognition, Predictive Analytics.

1.Introduction

The idea of speech recognition commenced someplace in Nineteen Forties, almost the primary speech popularity software become seemed in 1952 on the bell labs, that become approximately popularity of a digit in a noise loose surroundings. Nineteen Forties and Fifties recollect because the foundational length of the speech popularity generation, in this era paintings become performed at the foundational paradigms of the speech popularity this is automation and data theoretic models. In the 1960's we have been capable of understand small vocabularies (order of 10-a hundred phrases) of remote phrases, primarily based totally on easy acoustic-phonetic homes of speech sounds. The key technology that have been evolved at some point of this decade have been, clear out banks and time normalization techniques. In Seventies the medium vocabularies (order of a hundred-one thousand phrases) the use of easy template-primarily based totally, sample popularity techniques have been diagnosed. In Nineteen Eighties huge vocabularies (one thousand-unlimited) have been used and speech popularity issues primarily based totally on statistical, with a huge variety of networks for managing language systems have been addressed. The key invention of this period have been hidden markov model (HMM) and the stochastic language model, which collectively enabled effective new techniques for managing non-stop speech popularity trouble efficaciously and with excessive overall performance. In Nineteen Nineties the important thing technology evolved at some point of this era have been the techniques for stochastic language information, statistical mastering of acoustic and language models, and the techniques for implementation of huge vocabulary speech information systems. After the 5 a long time of research, the speech popularity generation has ultimately entered marketplace, reaping rewards the customers in kind of ways. The task of designing a gadget that definitely features like an smart human remains a primary one going forward.

1.1.An review of Speech Recognition

Speech popularity is a generation that capin a position a pc to seize the phrases spoken through a human with a assist of microphone. These phrases are in a while diagnosed through speech recognizer, and within side the end, gadget outputs the diagnosed phrases. An best scenario within side the technique of speech popularity is that, a speech popularity engine acknowledges all phrases uttered through a human but, almost the overall performance of a speech popularity engine relies upon on wide variety of elements. Vocabularies, a couple of customers and noisy surroundings are the essential elements which can be counted in because the relying elements for a speech popularity engine.

2. Predictive Analytics

2.1. Predictive Analytics

Predictive analytics is a place of Data Science that offers with extracting facts from records and the use of it are expecting tendencies and behaviour patterns. Often the unknown occasions of hobby is with inside the destiny, however predictive analytics may be carried out to any kind of unknown whether or not it's with inside the beyond, gift or destiny. For example, figuring out suspects after against the law has been committed, or credit score card fraud because it occurs (Strickland J., 2013). The middle of predictive analytics is predicated on taking pictures relationships among explanatory variables and the expected variables from beyond occurrences, and exploiting they are expecting the unknown final results. It is vital to note, however, that the accuracy and usefulness of effects will rely significantly on the extent of records evaluation and the excellent of assumptions.

2.2.Applications of Predictive Analytics

Although predictive analytics may be placed to apply in lots of applications, we define some examples wherein predictive analytics has proven tremendous effect in latest years.

- Analytical purchaser courting management (CRM)
- Clinical selection assist systems
- Collection analytics
- Fraud detection

2.3.Different strategies of Predictive Analytics

2.3.1.Linear Regression

One of the maximum not unusual place statistical techniques is linear regression. At its maximum simple, it's used while you need to specific the mathematical courting among variables or attributes. When you operate it, you're making the belief that there may be a linear courting among an final results variable (every now and then additionally known as the reaction variable, structured variable, or label) and a predictor (every now and then additionally known as an unbiased variable, explanatory variable, or characteristic); or among one variable and numerous different variables, wherein case you're modeling the connection as having a linear structure. Assuming that there may be a linear courting among an final results variable and a predictor is a large assumption, however it's additionally the only one You can make linear features are greater simple than nonlinear ones in a mathematical experience so in that experience it's an amazing beginning factor. In a few cases, it makes experience that modifications in a single variable correlate Linearly with modifications in any other variable, It's tougher to justify the belief of linearity besides locally: with inside the spirit of calculus, the entirety may be approximated through line segments so long as features are non-stop. Let's lower back up. Why could you even need to construct a linear version with inside the first place? You would possibly need to apply this courting to are expecting destiny outcomes, or you would possibly need to apprehend or describe the connection to get a hold close at the situation. Let's say you're reading the connection among a business enterprise's income and what sort of that business enterprise spends on advertising, or the quantity of buddies a person has on a social networking web page and the time that character spends on that web page daily. These are all numerical outcomes, which suggest linear regression could be a sensible choice, as a minimum for a primary by skip at your trouble. One access factor for considering linear regression is to reflect on consideration on deterministic strains first. We found out lower back in grade faculty that we should describe a line with a slope and an intercept, y = f x = β0 +β1 *x. But the placing there has been usually deterministic. Even for the maximum mathematically state-of-the-art amongst us, in case you haven't achieved it earlier than, it's a brand new attitude to begin considering stochastic features. We nevertheless have the identical components: factors indexed out explicitly in a table (or as tuples), and features represented in equation shape or plotted on a graph. So let's increase to linear regression beginning from a deterministic function.

2.3.2.Decision Trees

Decision tree is a kind of supervised studying set of rules (having a pre-described goal variable) this is primarily utilized in class problems. It works for each specific and non-stop enter and output variables. In this approach, we cut up the populace or pattern into or greater homogeneous sets (or sub-populations)
primarily based totally on maximum extensive splitter / differentiator in enter variables.

2.3.3.Random Forest

Random Forest is taken into consideration to be a panacea of all records technology problems. On a humorous note, while you can't think about any set of rules (no matter situation), use random forest. Random Forest is a flexible device studying technique able to appearing each regression and class tasks. It additionally undertakes dimensional discount techniques, treats lacking values, outlier values and different vital steps of records exploration, and does a reasonably properly job. It is a kind of ensemble studying technique, wherein a set of susceptible fashions integrate to shape a effective version

2.3.4.KNN

The motive of a bias closer to class fashions is that maximum analytical trouble entails creating a selection. For example will a purchaser attrite or now no longer, ought to we goal purchaser X for virtual campaigns, whether or not purchaser has a excessive capacity or now no longer etc. These evaluation are greater insightful and without delay hyperlinks to an implementation roadmap. We will communicate approximately any other broadly used class approach known as K-nearest neighbours (KNN).

2.3.5.SVM

Support Vectors are without a doubt the co-ordinates of man or woman observation. For example, (45,150) is a assist vector which corresponds to a female. Support Vector Machine is a frontier which first-class segregates the Male from the Females. In this case, the 2 lessons are nicely separated from every different, as a result it's far less difficult to discover a SVM.

2.3.6.Principal Component Analysis

In easy words, foremost element evaluation is a technique of extracting vital variables (in shape of components) from a big set of variables to be had in a records set. It extracts low dimensional set of functions from a excessive dimensional records set with a cause to seize as a lot facts as possible. With fewer variables, visualization additionally turns into a lot greater meaningful. PCA is greater beneficial whilst coping with three or better dimensional records. It is usually done on a symmetric correlation or covariance matrix. This way the matrix ought to be numeric and feature standardized records.

2.3.7.K-way Clustering

K-way clustering is a kind of unsupervised studying, that is used if you have unlabelled records (i.e., records without described classes or companies). The aim of this set of rules is to locate companies with inside the records, with the quantity of companies represented through the variable K. The set of rules works iteratively to assign every records factor to certainly considered one among K companies primarily based totally at the functions which can be provided. Data factors are clustered primarily based totally on characteristic similarity. The effects of the K-way clustering set of rules are: 1. The centroids of the K clusters, which may be used to label new records 2. Labels for the schooling records (every records factor is assigned to a unmarried cluster) Rather than defining companies earlier than searching on the records, clustering permits you to locate and examine the companies which have shaped organically. The "Choosing K" segment beneath describes how the quantity of companies may be determined. Each centroid of a cluster is a set of characteristic values which outline the ensuing companies. Examining the centroid characteristic weights may be used to qualitatively interpret what sort of institution every cluster represents. The K-way clustering set of rules is used to locate companies that have now no longer been explicitly labelled with inside the records. This may be used to verify commercial enterprise assumptions approximately what styles of companies exist or to discover unknown companies in complicated records sets. Once the set of rules has been run and the companies are described, any new records may be effortlessly assigned to the best institution.

3.Logistic Regression

3.1. Logistic Regression

Until 1972, human beings failed to recognize how to research facts which has a non-everyday mistakes distribution with inside the established variable. Then, in 1972, got here a step forward via way of means of John Nelder and Robert Wedderburn with inside the shape of Generalized Linear Models. Generalized Linear Models are an extension of the linear version framework, which incorporates established variables which can be non-everyday also. In general, they own 3 characteristics:

• These fashions incorporate a linear aggregate of enter functions.
• The suggest of the reaction variable is associated with the linear aggregate of enter functions thru a hyperlink feature.
• The reaction variable is taken into consideration to have an underlying chance distribution belonging to the own circle of relatives of exponential distributions together with binomial distribution, Poisson distribution, or Gaussian distribution. Practically, binomial distribution is used whilst the reaction variable is binary. Poisson distribution is used whilst the reaction variable represents count. And, Gaussian distribution is used whilst the reaction variable is continuous. Logistic Regression belongs to the own circle of relatives of generalized linear fashions. It is a binary class set of rules used whilst the reaction variable is dichotomous (1 or zero). Inherently, it returns the set of chances of goal elegance. But, we also can achieve reaction labels the use of a chance threshold fee. Following are the assumptions made via way of means of Logistic Regression:
• The reaction variable need to comply with a binomial distribution.
• Logistic Regression assumes a linear dating among the unbiased variables and the hyperlink feature (logit).
• The established variable need to have together exceptional and exhaustive categories. In R, we use glm() feature to use Logistic Regression. In Python, we use sklearn linear version feature to import and use Logistic Regression. We do not use Linear Regression for binary class due to the fact its linear feature consequences in chances outside [0, 1] interval, thereby making them invalid predictions.

3.2.Types of Logistic Regression Techniques

Logistic Regression isn't always simply constrained to fixing binary class issues. To clear up issues which have more than one instructions, we are able to use extensions of Logistic Regression, which incorporates Multinomial Logistic Regression and Ordinal

Logistic Regression. Let's get their primary idea: 1. Multinomial Logistic Regression: Let's say our goal variable has K = four instructions. This approach handles the multi-elegance trouble via way of means of becoming K-1 unbiased binary logistic classifier version. For doing this, it randomly chooses one goal elegance because the reference elegance and suits K-1 regression fashions that evaluate every of the final instructions to the reference elegance. Due to its restrictive nature, it isn't always used extensively as it does now no longer scale thoroughly with inside the presence of a big quantity of goal instructions. In addition, because it builds K - 1 fashions, we might require a far large facts set to obtain affordable accuracy. 2. Ordinal Logistic Regression: This approach is used whilst the goal variable is ordinal in nature. Let's say, we need to expect years of labor experience (1,2,3,four,5, etc). So, there exists an order with inside the fee, i.e., 5>four>3>2>1. Unlike a multinomial version, whilst we teach K -1 fashions, Ordinal Logistic Regression builds a unmarried version with more than one threshold values. If we've K instructions, the version would require K -1 threshold or cutoff points. Also, it makes an vital assumption of proportional odds. The assumption says that on a logit (S shape) scale, all the thresholds lie on a immediately line. Logistic Regression isn't a awesome desire to clear up multi-elegance issues. But, it is right to be aware about its types.

3.3.Working style of Logistic Regression

Logistic Regression assumes that the established (or reaction) variable follows a binomial distribution. Binomial distribution may be recognized via way of means of the subsequent characteristics:

• There need to be a set quantity of trials denoted via way of means of n, i.e. with inside the facts set, there need to be a set quantity of rows.

• Each trial may have best outcomes; i.e., the reaction variable may have best precise categories.

• The final results of every trial need to be unbiased of every different; i.e., the precise degrees of the reaction variable need to be unbiased of every different.

• The chance of success (p) and failure (q) need to be the equal for every trial. Let's recognize how Logistic Regression works. For Linear Regression, in which the output is a linear aggregate of enter feature(s),we write the equation as: `Y = βo + β1X + ∈`-------(1) In Logistic Regression, we use the equal equation however with a few adjustments made to Y. Let's reiterate a truth approximately Logistic Regression: we calculate chances. And, chances usually lie among zero and 1. In other words, we can say:

• The reaction fee need to be fantastic.

• It need to be decrease than 1.

First, we're going to meet the above criteria. We recognize the exponential of any fee is usually a fantastic quantity. And, any quantity divided via way of means of quantity + 1 will usually be decrease than 1. Let's put in force those findings:

$$P(Y=1|X) = \frac{e^{(\beta_o+\beta_1 x)}}{e^{(\beta_o+\beta_1 x)}+1} \quad \text{---------(2)}$$

This is the logistic function.

Now we can say that the probability value will be always between 0 and 1 . To determine the link function, follow the algebraic calculations carefully. P(Y=1|X) can be read as "probability that Y =1 given some value for x. Y can take only two values, 1 or 0. For ease of calculation, let's rewrite P(Y=1|X) as p(X).

$$\implies p(X) = \frac{e^{(\beta_o+\beta_1 x)}}{e^{(\beta_o+\beta_1 x)}+1}$$
$$\implies p(e^{(\beta_o+\beta_1 x)}+1) = e^{(\beta_o+\beta_1 x)}$$
$$\implies p.e^{(\beta_o+\beta_1 x)} + p = e^{(\beta_o+\beta_1 x)}$$
$$\implies p = e^{(\beta_o+\beta_1 x)} - p.e^{(\beta_o+\beta_1 x)}$$
$$\implies p = e^{(\beta_o+\beta_1 x)}(1-p)$$
$$\implies \frac{p}{1-p} = e^{(\beta_o+\beta_1 x)}$$
$$\implies \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x$$

As you might recognize, the right side of the (immediate) equation above depicts the linear combination of independent variables. The left side is known as the log - odds or odds ratio or logit function and is the link function for Logistic Regression. This link function follows a sigmoid (shown below) function which limits its range of probabilities between 0 and 1.

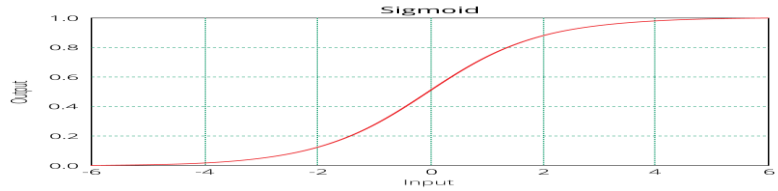

Figure 1: a sigmoid function with its range of probabilities between 0 and 1.

We can interpret the above equation as, a unit increase in variable x results in multiplying the odds ratio by ε to power β. In other words, the regression coefficients explain the change in log(odds) in the response for a unit change in predictor. However, since the relationship between p(X) and X is not straight line, a unit change in input feature doesn't really affect the model output directly but it affects the odds ratio.

This is contradictory to Linear Regression where, regardless of the value of input feature, the regression coefficient always represents a fixed increase/decrease in the model output per unit increase in the input feature.

In Multiple Regression, we use the Ordinary Least Square (OLS) method to determine the best coefficients to attain good model fit. In Logistic Regression, we use maximum likelihood method to determine the best coefficients and eventually a good model fit. Maximum likelihood works like this: It tries to find the value of coefficients (βo,β1) such that the predicted probabilities are as close to the observed probabilities as possible. In different words, for a binary classification (1/0), most probability will try and discover values of βo and β1 such that the ensuing possibilities are closest to both 1 or 0.The likelihood function is written as

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \quad \text{---------(3)}$$

3.4.Evaluating Logistic Regression model

In Linear Regression, we check adjusted R^2, F Statistics, MAE, and RMSE to evaluate model fit and accuracy. But, Logistic Regression employs all distinctive units of metrics. Here, we deal with probabilities and categorical values. Following are the evaluation metrics used for Logistic Regression:

1. **Akaike Information Criteria (AIC)**

You can have a take a observe AIC as counterpart of adjusted r square in more than one regression. It's an important indicator of model fit. It follows the rule: Smaller the better. AIC penalizes growing quantity of coefficients with in the model. In different words, including extra variables to the version would not permit AIC increase. It facilitates to keep away from over fitting. Looking on the AIC metric of 1 version would not surely help. It is extra beneficial in evaluating fashions (version selection). So, construct 2 or three Logistic Regression fashions and evaluate their AIC. The model with the lowest AIC will be relatively better.

2. Null Deviance and Residual Deviance

Deviance of an observation is computed as -2 times log likelihood of that observation. The importance of deviance can be further understood using its types: Null and Residual Deviance. Null deviance is calculated from the model with no features, i.e. only intercept. The null model predicts class via a constant probability.

Residual deviance is calculated from the model having all the features. On comparison with Linear Regression, think of residual deviance as residual sum of square (RSS) and null deviance as total sum of squares (TSS).The difference between null and residual deviance, is better in the model.

Also, you can use these metrics to compare multiple models: whichever model has a lower null deviance, means that the model explains deviance pretty well, and is a better model. Also, lower the residual deviance, better the model. Practically, AIC is constantly given desire above deviance to assess model fit.

3. Confusion Matrix

Confusion matrix is the most crucial metric commonly used to evaluate classification models. The skeleton of a confusion matrix looks like this:

	1 (Predicted)	0 (Predicted)
1 (Actual)	True Positive	False Negative
0 (Actual)	False Positive	True Negative

Figure 2: Confusion Matrix

As you can see, the confusion matrix avoids "confusion" by measuring the actual and predicted values in a tabular format. In above table, Positive class = 1 and Negative class = 0. Following are the metrics we can derive from a confusion matrix:

Accuracy - It determines the accuracy of overall predicted model. It is calculated as Accuracy = (True Positives + True Negatives)/(True Positives + True Negatives + False Positives + False Negatives)

True Positive Rate (TPR) - It indicates how many positive values, out of all the positive values, have been correctly predicted**.** The formula to calculate the true positive rate is (TP/TP + FN). Also, TPR = 1 - False Negative Rate. It is also known as **Sensitivity** or **Recall**.

False Positive Rate (FPR) - It indicates how many negative values, out of all the negative values, have been **incorrectly predicted**. The formula to calculate the false positive rate is (FP/FP + TN). Also, FPR = 1 - True Negative Rate.

True Negative Rate (TNR) - It indicates how many negative values, out of all the negative values, have been **correctly predicted**. The formula to calculate the true negative rate is (TN/TN + FP). It is also known as Specificity.

False Negative Rate (FNR) - It indicates how many positive values, out of all the positive values, have been incorrectly predicted. The formula to calculate false negative rate is (FN/FN + TP).

Precision: It indicates how many values, out of all the predicted positive values, are actually positive. It is formulated as:(TP / TP + FP).

F Score: F score is the harmonic mean of precision and recall. It lies between 0 and 1. Higher the value, better the model. It is formulated as 2((precision*recall) / (precision+recall)).

4. Receiver Operator Characteristic (ROC)

The ROC determines the classification model accuracy at a user defined threshold value. It determines the model's accuracy using Area under Curve (AUC). The area under the curve (AUC), also referred to as index of accuracy (A) or concordant index, represents the performance of the ROC curve. Higher the area, better the model. ROC is plotted between True Positive Rate (Y axis) and False Positive Rate (X Axis). In this plot, our intention is to push the red curve (proven below) in the direction of 1 (left corner) and maximize the area under curve. Higher the curve, higher the model. The yellow line represents the ROC curve at 0.5 threshold. At this point, sensitivity = specificity.

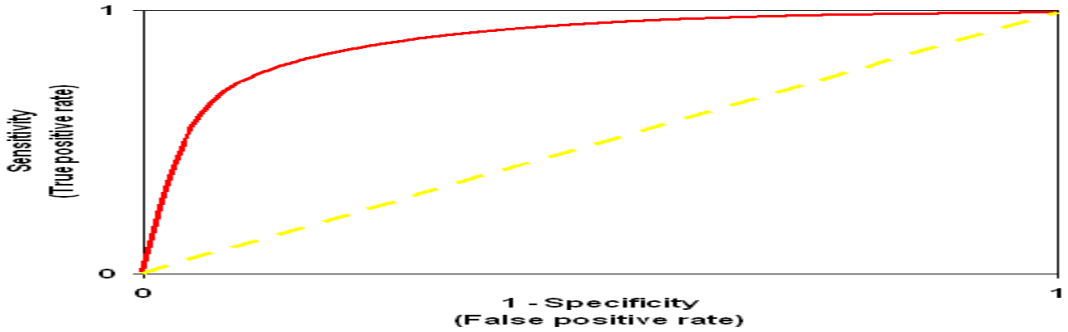

Figure 3: ROC Curve

For Model 1

Total Observations in Table: 634

test$label	test$pred_class: female	male	Row Total
female	293	9	302
	151.656	139.708	
	0.970	0.030	0.476
	0.964	0.027	
	0.462	0.014	
male	11	321	332
	137.953	127.084	
	0.033	0.967	0.524
	0.036	0.973	
	0.017	0.506	
Column Total	304	330	634
	0.479	0.521	

Figure 4: Confusion Matrix for Model 1

Accuracy: 0.968
Recall / True Positive Rate: 0.003
False Positive Rate: 0.029
True Negative Rate: 0.970
False Negative Rate: 0.033
Precession: 0.972
F1-Score: 0.0059

For Model 2

Total Observations in Table: 634

test$label	test$pred_class2: female	male	Row Total
female	292	10	302
	148.162	137.354	
	0.967	0.033	0.476
	0.957	0.030	
	0.461	0.016	
male	13	319	332
	134.774	124.943	
	0.039	0.961	0.524
	0.043	0.970	
	0.021	0.503	
Column Total	305	329	634
	0.481	0.519	

Figure 5: Confusion Matrix for Model 2

Accuracy: 0.963
Recall / True Positive Rate: 0.960
False Positive Rate: 0.033
True Negative Rate: 0.966
False Negative Rate: 0.039

Precession: 0.969
F1-Score: 0.964

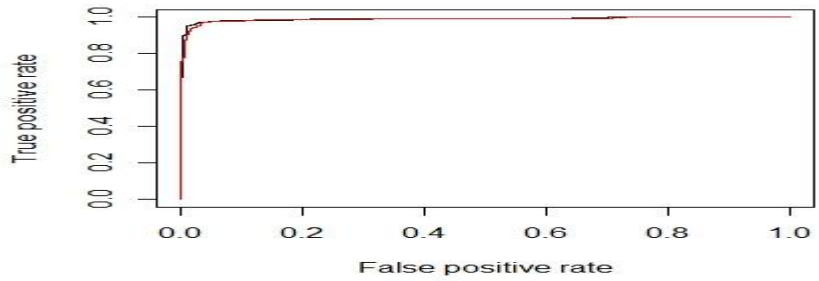

Figure 6: Roc Curve for model 1 (black line) and model 2 (red line)

Area under curve (AUC) for model 1: 0.9911
Area under curve (AUC) for model 2: 0.9899
Therefore the model 1 is better model than model 2 for prediction

4.Conclusion

In this paper i've efficaciously predicted (i.e. 97% of precision) the given voice parameters into male and female. We have constructed 2 fashions the use of Logistic Regression Technique, amongst them we've selected one version with maximum precision, keep in mind, AUC and F1-rating as our predictive version. Further there's a scope for extending values of precision and keep in mind via way of means of the use of diverse different technologies.

References

1.NurAdlinNazifa[1], ChongYenFook[2], LimCheeChin[3], Vikneswaran Vijean[4] and EngSweeKheng5 (2018) "Gender Prediction by Speech Analysis" *Journal of Physics: Conference Series*, Volume 1372**,** International Conference on Biomedical Engineering 26–27 August 2019, Penang island, Malaysia

2.Barkin, E. (2011). CRM + Predictive Analytics: Why It All Adds Up. NewYork:DestinationCRM.Retrieved2014,from http://www.destinationcrm.com/Articles/Editorial/MagazineFeatures/CRM---Predictive-Analytics-Why-It-All-Adds-Up- 74700.aspx

3.Hayward,R. (2004). Clinical decision support tools: Do they support clinicians?FUTURE Practice, 66-68. Nigrini, M. (2011). Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations. Hoboken, NJ: John Wiley & Sons Inc. Parry, G. (1996, November–December). The characterization of uncertainty in Probabilistic Risk Assessments of complex systems. Reliability Engineering & System.2014 http://www.sciencedirect.com/science/article/pii/S095183209600 0695

4.Schiff M. (2012, March 6). BI Experts: Why Predictive Analytics Will ContinuetoGrow.Renton:TheDataWarehouseInstitute.Retrieved2014,from http://tdwi.org/Articles/2012/03/06/PredictiveAnalytics-Growth.aspx?Page=1

5.Strickland, J. (2013). Introduction toe Crime Analysis and Mapping. Lulu.com. Retrieved from http://www.lulu.com/shop/jeffreystrickland/introduction-to-crime-analysis-andmapping/paperback/product-21628219.html

6.Strickland, J. (2013). Introduction toe Crime Analysis and Mapping. Lulu.com.Retrievedfrom. http://www.lulu.com/shop/jeffreystrickland/introduction-to-crime-analysis-andmapping/paperback/product-21628219.html

7. Brett Lantz Machine Learning with R

8.Robert I. Kabacoff R IN ACTION(Data analysis and graphics with R)

9. Ioannis E[1]. Livieris[2],Emmanuel Pintelas[3] and Panagiotis Pintelas[4] "Gender Recognition by Voice Using an Improved Self-Labeled Algorithm"Department of Computer & Informatics Engineering, Technological Educational Institute of Western Greece, GR 263-34 Antirion, Greece Department of Mathematics, University of Patras, Grace.published: 5 March 2019