**Part I:**

1. This dataset is a survey that aims to analyze how managers utilize decision-science methods within their decision-making processes. The variables are OrganizationType, Yrsofworking, Levelofmanagement, functionalrole, stem\_education, Impact of Big Data, accessibility, and DS usage. The primary response variable is DS usage. The first question to investigate is what the relationship between functional role and impact of Big Data, accessibility, and DS usage? The second question aims to determine what the relationship between level of management and DS usage acess is. The third question is what is the relationship between stem\_education and DS accessibility? The fourth question is what the relationship between Organization type and impact of Big Data is. Finally, what is the relationship between years of working and the impact of Big Data?
2. There are five variables that are used in the survey to assess whether someone will acquire data and use data-based tools to support day-to-day decision making. The response variable is an ordinal variable from a scale of 1-5 to assess how likely one is to use data. The five variables are organization type, years of working, level of management, functional role in the organization who makes decisions, and whether the post-secondary education background of the manager in in STEM.

The criteria for whether someone “uses” data-based decision making is whether they gave a 4 or a 5 on the Likert Response. A 3 or below is classified as does not routinely utilize data-driven decision making. If the sample of the population stays the same, the variables that have the greatest predictive power are those that have 15% or .15 probability or more of their respective contingency table totals (adding together 4 &5), or if the row totals for a row are over 50%. For the organization type table, there is .3142 probability that a private organization will give a rating of 4 or higher. From the years of working table, there is a .1450 probability that an employee > 5 years & < 10 will give a ranking of 4 or greater, and there is a .1754 probability that an employee >10 & < 15 will give a ranking of 4. There are two variables with predictive power within the levels of management table. There is a .1623 probability that a manager that is entry level will give a rating that is 4 or higher, and a .1859 probability that a mid-level manager will give a rating of 4 or higher. From the functional role table, there are no rows that score above .15 of probability. However, 68.75% of the operational/engineering row gives a 4 or a 5, despite the probability of the total table being 11.52. Finally, in the stem education table, there is a .3089 probability that a manager with a stem education will give a rating of 4 or 5 on the Likert scale. The following variables have been put in descending order in data table.

|  |  |
| --- | --- |
| **Variable** | **Probability** |
| Private Organization | 0.3142 |
| STEM education | 0.3089 |
| Manager Mid- Level | 0.1859 |
| Employee between 10 & 15 | 0.1754 |
| Manager Entry Level | 0.1623 |
| Employee between 5 &10 | 0.145 |
| Operational/Engineering | 0.1152 |

**Part II:**

1. **3.**  The first set of five partitioning chosen was the same as the contingency analysis tables with the five variables: organization type, years of working, level of management, functional role in the organization who makes decisions, and whether the post-secondary education background of the manager in in STEM. The max number of splits for organization type is two, with an Rsquare value of 0.018. The decision tree was split into Organization type (1,2), organization type (3), then split again into organization type (1) and organization type(2). The node or the leaf with the highest probability of a “use” of data science is Organization type 2, or private organization, with a probability of .4606 within that leaf. The second partition was the variable years of working. There were 3 splits, with an R-square value of .03, which only covers 3% of the variation in the dataset, or very low. Of the splits, the node with the highest probability was yrsofworking(3), or between 10 to 15 years of experience, with a combined probability of .5389. The third partition was level of management. The RSquare level was 0.022, which only covers 2.2% of the variation in the dataset. The node or branch with the highest probability was level of management (2), which is considered mid-level, with a combined probability of .5688 within that specific node. The next node was functional role with 4 splits and an RSquare value of .073, which is a low amount only covering 7% of the variation within the dataset. The highest probability for functional role nodes was functional role(5), or operational and engineering, with a probability of .7381. The final partition to find DS usage was educational background of manager. The node with the highest probability was Stem\_education(1), or that they did have a Stem education, with a probability of .5687.
2. i. A manager is more likely to use a DS tool if they have easier accessibility. For the node accessibility (4,5), the probability that a manager will use DS tools is .6296. Separated further into a further node, that probability increases to .7889 if a manager rates accessibility at a 5 level.

ii. In general, the higher the accessibility, the higher a manager will rank the importance of Big Data on their decision making. The nodes with the highest probabilities are Accessibility(4,5) with a probability of .5114 and Accessibility(4) with a probability of .5198.

**4.** Using the stratified validation data, the answers are like the original dataset. For the probability of Accessbility(4,5), the node produced a probability of .6211, compared to the original probability of .6296. For accessibility (5), the validation set produced a probability of .701, .089 probability from the original dataset.

For the impact of Big Data accessibility, the answers were similar. For accessibility (4,5), the node had a probability of .4926, compared to the original probability of .5114 and a probability of .5078 for the node accessibility (4) compared to the original of .5198. I would say that this decision tree does not overfit to the original data.

**Summary**

The purpose of this survey was to investigate which factors can contribute and have a relationship with decision-science/analytics usages for data-driven decision making within organizations. The variables that were investigated were organization type, years of working, level of management, functional role of the decision-making manager within the organization, whether the post-secondary education of the manager is in the math and sciences or non-STEM areas, and the perceived ease-of-access of data when a decision is made, or accessibility. To build a data-based culture, there are multiple decisions that an organization can take to accomplish these goals.

According to a survey done by PWC, one of the Big 4 accounting firms, executives that are highly data-driven are three times more likely to report significant improvements in decision-making compared to those that use less data (Stobierski, 2019). According to the data retrieved from the decision trees from the survey data, there are two separate ways organizations can accomplish this. The first is to make strategic decisions on who to hire for managers who are in decision-making roles and the second is to have more accessible data when making decisions. Years of working and levels of management are two variables that have similar probabilities for DS usage. Those with between 10 to 15 years of work experience have a probability of DS usage of .54 and Mid-level managers have a probability of .57. When organizations choose their managers, there is a higher probability that those who are mid-level or with 10-15 years of work experience will use decision science methods. When choosing a manager based on function, there is one specific function that has a significantly higher probability of using DS methods. Managers that make decisions who are operational/engineering have a .74 probability of using DS methods. Finally, for those managers that have a STEM education background, the probability that they will use DS methods is .57. If an organization wants to choose a manager that will make data-driven decisions, the ones that they should choose from in order of highest probability are operational/engineering managers, managers with a post-secondary education in STEM, Mid-level-managers, and finally, those with between 10 to 15 years of work-experience. Another important variable for whether a company uses DS methods for data-driven decision making is accessibility.

Accessibility of data is defined as the perceived ease-of-access of both data and data-based tools when a decision is being made. In the survey, those who gave a 4 or 5 in their accessibility ranking had a probability of using DS of .6296. For those that gave a 5, that probability rose to .7889. It is therefore imperative that organizations have the best accessibility of data for every function of business. The most reliable way to have accessible data is to have a data management strategy. There are five critical components to a data management strategy: identify the business needs and how data can address those needs, find the right technologies, avoid data silos, have data governance, and training and education (Lin). It is imperative to first identify your organizations objectives for the data, then have a process to manage, prepare, analyze, and store data (Lin). The right tools are imperative for data management. Dashboarding technologies such as Microsoft PowerBi and Tableau offer non-data professionals a way to have acesss and understand important data. Data silos can occur when different functions within the business have different methods for collecting, storing, and analyzing data. The most effective way to break up a data silo is to store the data in a data lake so that data can be shared easier across departments (Lin). Data governance addresses security, quality, transparency, and data privacy. Some of the most popular tools are SAS data management and IBM Data Governance. Finally, it is imperative that non data professional employees can understand, interpret, and communicate the data within the organization. Data literacy trainings across the organization is the best way to address this strategy.

# References

Lin, S. (n.d.). *5 ways to make data more accessible to every department.* Retrieved from RingCentralBlog: https://www.ringcentral.com/us/en/blog/5-ways-to-make-data-more-accessible-to-every-department/

Stobierski, T. (2019, August 26). *The advantages of Data Driven Decision Making.* Retrieved from Harvard Business School Online: https://online.hbs.edu/blog/post/data-driven-decision-making